COPPE/UFRJ

UMA METODOLOGIA DE MINERAÇÃO DE DADOS PARA A PREVISÃO DE INSOLVÊNCIA DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO

Rui Américo Mathiasi Horta

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia Civil.

Orientadora: Beatriz de Souza Leite Pires de Lima

RIO DE JANEIRO, RJ - BRASIL
AGOSTO DE 2010

UMA METODOLOGIA DE MINERAÇÃO DE DADOS PARA A PREVISÃO DE INSOLVÊNCIA DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO

Rui Américo Mathiasi Horta

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

_____
Profª. Beatriz de Souza Leite Pires de Lima, D.Sc.

_____
Prof. Nelson Francisco Favilla Ebecken, D.Sc.

_____
Prof. Alexandre Gonçalves Evsukoff, Dr.

_____
Profª. Heloisa Márcia Pires, D.Sc.

_____
Prof. Carlos Cristiano Hasenclever Borges, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
AGOSTO DE 2010

Horta, Rui Américo Mathiasi

Uma metodologia de mineração de dados para a previsão de insolvência de empresas brasileiras de capital aberto/ Rui Américo Mathiasi Horta. – Rio de Janeiro: UFRJ/COPPE, 2010.

XIV, 151 p.: il.; 29,7 cm.

Orientadora: Beatriz de Souza Leite Pires de Lima

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2010.

Referências Bibliográficas: p. 137-151.

1. Previsão de insolvência. 2. Mineração de dados. 3. Balanceamento de bancos de dados 4. Comitê de classificadores. 5. Setores econômicos. I. LIMA, Beatriz de Souza Leite Pires de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

À minha querida mãe que mesmo na ausência sempre me conduz.

Ao meu pai pelos incentivos e motivações.

À minha esposa Tânia pelas compreensões e opiniões.

À minha filha Rafaela pelo carinho e a permanente alegria.

AGRADECIMENTOS

Agradeço:

A minha orientadora, Professora Beatriz de Souza Leite Pires Lima, pela confiança, estímulo e paciência a mim dedicados, mas, sobretudo pelo exemplo de profissionalismo e dedicação sempre presente neste e nos trabalhos que colaborou;

A Universidade Federal de Juiz de Fora pelo entendimento das dificuldades de se realizar este projeto e facilitar o mais possível este trabalho sempre acreditando no seu resultado;

Ao colega Custódio e também aos colegas da Faculdade de Economia e Administração da UFJF e a tantos outros que torceram pelo meu sucesso e me apoiaram sempre que precisei.

Dedico um agradecimento especial ao colega Carlos Cristiano, que participou ativamente na elaboração desta tese com sugestões, observações, críticas e muita motivação nas horas mais difíceis desse projeto.

A todos os funcionários da COPPE-PEC pelo eficiente apoio administrativo.

e, acima de tudo, a Deus.

Uma descoberta não consiste em ver o que todo mundo não viu, mas em pensar o que ninguém ainda pensou. (Goethe)

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

UMA METODOLOGIA DE MINERAÇÃO DE DADOS PARA A PREVISÃO DE INSOLVÊNCIA DE EMPRESAS BRASILEIRAS DE CAPITAL ABERTO.

Rui Américo Mathiasi Horta

Agosto/2010

Orientadora: Beatriz de Souza Leite Pires de Lima

Programa: Engenharia Civil

As empresas brasileiras vêm passando por muitas mudanças no seu ambiente de negócios, como por exemplo: o aumento permanente da concorrência; diminuição da vida útil dos produtos e serviços; aumento lento, mas constante de custos financeiros e sociais; aumento nas despesas para adequações às legislações fiscais, tributárias e sociais, dentre outras. Devido, sobretudo a esses motivos, tem havido um aumento no número de empresas que vem sofrendo com a sua descontinuidade causada pela insolvência. Conhecer antecipadamente as empresas que possam se tornar insolventes é preponderante para evitar futuros prejuízos financeiros e sociais. Por outro lado tem sido verificado que a descoberta de conhecimento em bases de dados originados nas demonstrações contábeis das empresas vem ganhando importância e interesse, sobretudo em áreas estratégicas em organizações de negócios. Estes dados são normalmente reais, legais e, sobretudo confiáveis.

Esta tese apresenta uma metodologia utilizando ferramentas de mineração de dados para previsão de insolvência adequada às mudanças que ocorrem com constância nos ambientes dos negócios. São utilizados dados contábeis e os resultados mostraram que os principais objetivos foram alcançados.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

A METHODOLOGY FOR DATA MINING FOR BANKRUPTCY PREDICTION OF BRAZILIAN COMPANIES CAPITAL OPEN.

Rui Américo Mathiasi Horta

August/2010

Advisor: Beatriz de Souza Leite Pires de Lima

Department: Civil Engineering

The Brazilian companies are passing for many changes in his environment of businesses, among those changes can be mentioned: the permanent increase of the competition; decrease of the useful life of the products and services; the increase slow, but constant of their financial and social costs; increase in the expenses for adaptations to fiscal, tax and social legislations that are constantly modified accompanying the changes in the business environment, among others. Due to those reasons, there is an increase in the number of companies that is suffering with its discontinuity caused by the insolvency. It is preponderant to identify in advance the companies that can become insolvent in order to avoid financial and social damages in the future. On the other hand, it has been verified that the knowledge discovery in databases originated of the accounting demonstrations of the companies is receiving importance and interest, mainly in strategic areas of businesses. In those areas, the data originated from accounting demonstrations are real, legal and, essencially, reliable and confidential. Then, the creation of a methodology using the data mining tools is fundamental to guarantee the efficiency in the application of modeling techniques for forecast of insolvency adequated to the changes that happen with constancy in business environment. Accounting data are used and the results showed that the main objectives were achieved.

SUMÁRIO

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1. Introdução

1.1 Considerações Iniciais

Neste capítulo é apresentada uma descrição geral da tese, onde se pretende fornecer ao leitor uma visão dos problemas tratados, dos objetivos principais e das contribuições, bem como uma descrição da organização deste trabalho.

1.2 Relevância do estudo

A importância do desenvolvimento de estudos na área de previsão de insolvência é muito abrangente. Ela permite que seja possível prever uma situação financeira difícil com certa antecedência, de forma que haja tempo hábil para serem adotadas medidas que reverta essa situação, como pode também influenciar em várias etapas importantes do negócio, como por exemplo: na tomada de decisão de investimento, na avaliação de desempenhos ou de concessão de crédito, na facilitação da aplicação dos procedimentos de auditoria, etc.

Nas últimas décadas, a previsão de insolvência em empresas é um assunto que vêm apresentando muitos estudos no domínio da contabilidade, finanças e nas técnicas de modelagem. Há um grande número de pesquisas acadêmicas no mundo, muitas dessas pesquisas são baseadas em técnicas mais contemporâneas de modelagem e dentre elas está a de Mineração de Dados ou Data Mining - DM - que nos recentes anos vem sendo muito pesquisada e se tornando cada vez mais desenvolvida e acessível. Numerosos modelos sobre previsão de insolvência foram estudados e desenvolvidos especialmente aplicando as técnicas de DM.

Nos anos recentes, uma revolução tem fervilhado à maneira pela qual a previsão de insolvência é medida e gerida (Saunders *et al.*, 2004). Contradizendo a história relativamente entediante e rotineira de previsão de insolvência, novas tecnologias e idéias têm emergido em meio a uma nova geração de estudiosos em engenharia financeira que está aplicando suas habilidades em construção de modelos e nas análises nessa área.

Algumas justificativas podem ser encontradas para este súbito impulso no aumento de interesse sobre o assunto:

1. Embora a recessão mais recente tenha atingido vários países em momentos diversos e consequências diferentes, a maioria das estatísticas de falências mostrou um significativo aumento de sua ocorrência em comparação à recessão anterior, (Saunders e Cornett, 2008). Considerando que tem havido um aumento permanente ou estrutural de falências em todo o mundo – possivelmente devido ao aumento da competição global, desenvolvimento da tecnologia, rápidas mudanças dos perfis dos consumidores, perturbações na estrutura financeira das empresas, dentre outros motivos – a pesquisa sobre previsão de insolvência, torna-se ainda mais importante hoje do que foi no passado.
2. Nas últimas décadas o ambiente econômico geral das empresas, na grande maioria dos países, tem mudado com uma enorme velocidade e experimentado tendências para baixo (dificuldades financeiras). Os ambientes financeiros e operacionais em que as companhias atuam têm mudado muito, elas tem passado por perturbações permanentes.
3. Como os mercados de capitais vêm se expandindo e tornando mais acessíveis a um maior número de empresas, a desintermediação está ocorrendo rapidamente, ou seja, aquelas empresas que vão obter recursos em bancos ou instituições financeiras tradicionais são cada vez mais prováveis de serem menores e possuírem classificações de créditos mais duvidosas.
4. Quase que paradoxalmente, apesar de um declínio na qualidade média dos empréstimos (devido ao motivo anterior, desintermediação), as margens de juros ou *spreads*, especialmente em mercados de empréstimos por atacado, têm-se tornado muito estreitas – ou seja, a compensação de risco-retorno advinda de empréstimos vem piorando. Várias razões podem ser citadas, mas um fator importante tem sido a competição por tomadores de empréstimos com menor qualidade, intensificada por parte de empresas financeiras, grande parte da atividade de empréstimos das quais se tem concentrado na ponta de risco mais elevado e de menor qualidade do mercado (Saunders e Cornett, 2008).
5. Concomitante com a recente crise financeira mundial, crises bancárias e de seguradoras de mercados de países altamente desenvolvidos como EUA, Japão e países europeus como Alemanha e Inglaterra, tem mostrado que valores de imóveis e de ativos físicos são muito difíceis de prever e de realizar através de

liquidação. Quanto mais fracos e incertos forem os valores das garantias reais, mais arriscada se torna a avaliação da capacidade financeira de uma empresa.

6. Com os avanços das tecnologias computacionais, tais como os sistemas de computadores relacionados à tecnologia da informação junto com a evolução da disponibilidade e desenvolvimento de base de dados, há um aumento de oportunidades de testar técnicas cada vez mais sofisticadas de modelagem à previsão de insolvência (Chye e Chin, 2004).
7. O crescimento da exposição do crédito, ou risco de contrapartida, devido à fenomenal expansão de mercados de derivativos, estendeu a necessidade de análises mais acuradas sobre previsão de insolvência utilizando registros contábeis e extra contábeis de empréstimos. Em muitos dos maiores bancos dos EUA, o valor teórico (não de mercado) de sua exposição extra balanço a instrumentos como *swaps* (troca de dívidas que podem ser, principalmente, de taxas de juros ou de moedas) e de mercado futuros de balcão (OTC) é mais do que dez vêzes o montante em seus registros contábeis de empréstimos (Saunders, 2000).
8. Outro grande incentivo para instituições financeiras desenvolverem novos modelos de previsão de insolvência é a imposição pós-1992 de exigências de reservas de capital para empréstimos, pelo BIS[1] e por bancos centrais (Cornett, 2007).
9. Empresas insolventes acabam gerando envolvimentos em vários setores econômicos com grandes custos econômicos e sociais. Pesquisas sobre o assunto são estimuladas por vários agentes econômicos públicos e privados, que induzem sempre às melhorias na acurácia dos modelos de previsão de insolvência. Os agentes privados visam à obtenção de melhores habilidades nas prevenções e nas ações corretivas das empresas que podem vir a ter problemas futuros de solvência. Já os agentes públicos podem detectar performances ruins de companhias e agir através de ações corretivas para evitar o insucesso da firma ou de um grupo de firmas e uma conseqüente perda de tributos acompanhado de aumentos dos custos sociais.

[1] BIS – Banco para Compensações Internacionais.

10. Na linha de extensão das pesquisas acadêmicas dos impactos dos *mercados imperfeitos e informações assimétricas*[2], trabalhos de previsão de insolvência têm aumentado. Mercados financeiros não são perfeitos por isso as avaliações de fluxo de fundos são insuficientes para determinar lucros ou bons projetos (isto é, projetos com valor presente líquido positivo) e, conseqüentemente, alguns valores criados nos projetos podem ser deixados sem financiamento.

11. Evitar possíveis custos diretos e indiretos de dificuldades financeiras. Custos diretos se referem a despesas judiciais e administrativas de insolvência. Esses custos foram estimados por White, Altman e Weiss (Ross *et al*., 2002, p. 346) e correspondem a aproximadamente a 3% do valor de mercado da empresa. Lawrence Weis (Brealey e Myers, 1996, p. 488) encontrou custos médios de cerca de 3% do valor contábil do ativo total e 20% do valor de mercado dos capitais próprios no ano anterior ao da falência. Warner (Ross, *et al.*, 2002, p. 346) constatou que os custos líquidos eram, em média, 1% do valor de mercado da empresa sete anos antes da falência, atingindo porcentagens maiores à medida que a falência se aproximava.

    Já os custos indiretos se referem à redução da capacidade de operação devido a uma possível falência da empresa levando a uma piora no relacionamento com clientes e fornecedores, havendo, frequentemente perda de vendas por temor de interrupção de serviço e perda de confiança. Altman (Ross, 2002, p. 347) estimou que os custos diretos e indiretos freqüentemente superem 20% do valor da empresa. Já Olsen (1996) mostra que os custos da concordata são mais de 20% dos resultados financeiros, embora, se os custos forem medidos sobre o valor contábil pré-concordata, o índice caíra para 6 %.

12. Setores econômicos apresentam situações contábeis, financeiras e operacionais distintas muito em função de exigências externas e internas da empresa. Podem ser citadas algumas dessas exigências referentes à segmentação das empresas: (a) caracterização do segmento, (b) produtos e mercados, incluindo aí perfil dos consumidores, (c) processo de produção, (d)

[2] Assimetria de informação ocorre quando um dos agentes numa dada transação dispõe de uma informação (crucial) que o outro não tem, ou quando um dos agentes não consegue descortinar as ações do outro. Este cenário é o mais comum no mercado que não é, de modo algum, o reino da "informação perfeita", pelo contrário (Martins e Lopes, 2007, p.31).

desempenho do setor, compreendendo os diversos mercados (locais, regionais, estaduais, nacionais, internacional), (e) posicionamento do setor econômico (da empresa) na cadeia de valor.

Para Kanitz (1978) vários estudos mostram que empresas insolventes começam a acusar sinais de dificuldades bem antes de chegar ao ponto crítico de uma falência ou concordata. É intuitivamente compreensível que a insolvência, sendo um processo que tem começo, meio e fim, se inicia muito antes de se concretizar. Portanto, devem existir nos demonstrativos contábeis publicados, antes da tragédia final, alguns indícios do que está para acontecer.

A insolvência, na verdade, começa quando uma empresa é incapaz de efetuar os pagamentos programados de sua dívida ou quando as projeções do fluxo de caixa da empresa indicam que ela logo será incapaz de efetuar esses pagamentos.

Prever com 100% de certeza que uma empresa vai se tornar insolvente é talvez impossível, visto o próprio ambiente de incerteza do mercado, e outras situações adversas a que todas as empresas estão sujeitas. Mas é possível identificar com antecedência as empresas com maiores probabilidades de falir em um futuro próximo. Isso não seria um prenúncio desagradável do fim da empresa por concordatas e falências, pois a grande maioria das empresas passa eventualmente por períodos difíceis e se recuperam, mas a idéia é conseguir identificar e apontar aquelas que têm maiores chances de não conseguir sobreviver à situação de crise em um futuro próximo, e assim adotar medidas preventivas. Para descobrir com antecedência e com um grau de segurança, qual a situação financeira de uma empresa é preciso, primeiro, determinar o que se denomina de fator de insolvência.

O fator de insolvência é um indicador daquilo que pode acontecer em futuro próximo, caso a empresa não corrija os rumos que está seguindo. É importante, ainda salientar que o fator de insolvência não é apenas um desagradável prenúncio de concordatas e falências. Uma aplicação menos pessimista surge quando nos concentramos do outro lado, isto é, do lado da solvência. Por exemplo: como uma instituição financeira deve decidir sobre pedidos de empréstimos ou financiamento de duas empresas, quando os recursos disponíveis, no momento, permitem atender a um pedido apenas? Uma das fórmulas para sair do impasse seria escolher a empresa que tivesse o melhor índice de solvência e, conseqüentemente, melhor situação financeira.

Além disso, o número de partes que são afetadas pelo fracasso corporativo de uma entidade é considerável. Bancos, investidores, governos, auditores, gerentes, fornecedores, empregados e muitos outros têm significativos interesses na acurácia da previsão de insolvência de uma companhia.

Apesar dos avanços em tecnologia da informação e das técnicas de modelagem terem se desenvolvido e se tornado cada vez mais acessível vindo a incrementar os estudos deste assunto, existem vários problemas relatados na literatura específica sobre a elaboração de modelos de previsão de insolvência (Balcaen e Ooghe, 2006; Kumar e Ravi, 2007; Nanni e Lumini, 2009; Tsai, 2009) e continuam também presentes quando da aplicação de DM.

Conforme será visto no capitulo 2, que discute trabalhos relacionados ao tema, foi encontrado poucos estudos com foco em alguns dos problemas aqui relatados referentes à elaboração de modelos de previsão de insolvência. Além disso, a maioria dos estudos relacionados pouco se concentra em se discutir separadamente a importância destes problemas, e não se conhece soluções em que sejam considerados os problemas abordados aqui conjuntamente. Também, para a extração de regras de classificação, em conjunto com os problemas citados anteriormente, há uma quantidade pouco significativa de trabalhos desenvolvidos na literatura específica.

A constatação anterior fica bem mais evidente quando se trata de estudos com dados contábeis de empresas brasileiras de capital aberto.

1.3 Objetivos

Como visto na seção anterior a importância de se obter um modelo de previsão cada vez mais eficiente vem incrementando estudos neste tema. Motivos não faltam para se buscar este objetivo, mas há questões ainda pouco exploradas nesta área. O presente trabalho tem como objetivo principal fazer um estudo no contexto de empresas brasileiras de capital aberto visando identificar com mais precisão, a previsão de insolvência de empresas do mercado brasileiro. Esta previsão é feita baseando-se nas relações contábeis anuais das empresas e focando uma melhor caracterização das empresas que tenham maior potencial de se tornarem insolventes.

Uma das questões ainda pouco exploradas nesta área é a origem dos dados a serem estudados. Nos estudos de modelagem de previsão de insolvência, as variáveis podem ter origem nos dados contábeis, no mercado e em ambos. Nesta tese também

será apresentada justificativas do uso somente dados contábeis para se obter modelos de previsão de insolvência mais confiáveis e precisos nesse tipo de modelo.

Outra questão que impacta os resultados da modelagem de previsão, mas ainda é pouco explorada na literatura específica, é a questão do balanceamento da base de dados. Em mercados com ambientes econômicos em condições normais (é o caso referente a este estudo), a quantidade de empresas existentes que podem compor a amostra das empresas insolventes é sempre bem menor do que o número de empresas que podem compor a amostra das solventes. Diante disso, os sistemas de aprendizado (classificadores) normalmente encontram dificuldades em induzir o conceito relacionado à classe minoritária. Nessas condições, modelos de classificação que são otimizados em relação à precisão têm tendência de criar modelos triviais, que quase sempre predizem bem a classe majoritária, gerando uma supremacia na classificação das empresas solventes sobre as insolventes, distorcendo o objetivo principal desta modelagem que é o de melhor caracterizar as empresas insolventes. Neste estudo será proposto um procedimento no qual o balanceamento da base de dados a ser estudada tenha melhor capacidade de caracterizar as empresas que podem vir a se tornarem insolventes. Também serão comparados os resultados com aqueles gerados por um procedimento bem referenciado na literatura com a finalidade de balancear bancos de dados.

Os demonstrativos contábeis evidenciam as estruturas patrimoniais e financeiras das empresas (Iudícibus, *et al.*, 2007, p.488), os conteúdos desses demonstrativos refletem similaridades patrimoniais e financeiras em função do setor econômico no qual elas pertencem (as empresas) refletindo características dos ambientes interno e externo que envolvem a empresa. Devido a isso, os atributos econômico-financeiros que comporão os modelos deverão ter origem nos demonstrativos contábeis das empresas pertencentes ao mesmo setor econômico. A setorização econômica das empresas também deve ser levada em conta nos modelos de previsão de insolvência. Nos sistemas de aprendizado de máquina na literatura específica, não há distinção entre as mesmas variáveis de setores econômicos distintos, podendo induzir os classificadores a caracterizações distorcidas no aspecto contábil financeiro. Consequentemente, os modelos elaborados são pouco confiáveis induzindo a erros de previsões. Essa é também outra questão impactante nos resultados do modelo desenvolvido e explorada nesta tese.

Outra questão a ser discutida nesta tese é a seleção de atributos para modelagem na previsão de insolvência. Selecionar atributos é preponderante na elaboração de modelos de previsão de insolvência quando do uso de dados originados, sobretudo, em demonstrativos contábeis. Apesar disso, a grande maioria dos estudos sobre previsão de inadimplência parte de um conjunto inicial de variáveis, escolhidas na maioria das vêzes de forma arbitrária, com base na popularidade da literatura específica e de seu sucesso preditivo nas pesquisas precedentes. A seleção de atributos fica ainda mais preponderante na medida em que os dados são confeccionados através de demonstrativos contábeis de empresas de mesmo setor econômico. Os atributos característicos para previsão de insolvência das empresas de certo setor econômico, provavelmente, serão distintos daqueles que caracterizarão a insolvência para as empresas de outro setor econômico. Além disso, a seleção de atributos desempenha uma tarefa essencial dentro do processo de modelagem, pois representa um problema de fundamental importância em DM, sendo freqüentemente realizada como uma etapa de pré-processamento, ela consiste em encontrar um subconjunto de atributos no qual o classificador utilizado em DM irá se concentrar.

Na modelagem de previsão de insolvência são utilizados um ou mais algoritmos de aprendizagem com o propósito de identificar um classificador que seja mais apropriado para o relacionamento entre o conjunto de atributos (dados contábeis) e o rótulo da classe dos dados de entrada (empresas solventes e insolventes). Esse modelo gerado pelo algoritmo deve se adaptar bem aos dados de entrada e previr corretamente os rótulos de classes de registros que ele nunca viu antes. Portanto, um objetivo chave do algoritmo de aprendizagem é construir modelos com boa capacidade de generalização; isto é, modelos que prevejam com precisão os rótulos de classes de registros não conhecidos previamente (Han e Kamber, 2006). E essa é a outra questão a ser estudada nesta tese, a escolha de um algoritmo de aprendizagem para identificar um modelo, também de acordo com o setor econômico, que seja mais apropriado para o relacionamento entre o conjunto de atributos e o rótulo da classe dos dados de entrada.

Outro problema a ser estudado nesta tese é a obtenção de um conjunto de regras de classificação para empresas insolventes de acordo com o setor econômico no qual as empresas pertençam. Em face das características operacionais das empresas atuantes em setores diferentes (consumo cíclico, bens industriais, construção e transporte, por exemplo), estes resultados facilitam um melhor entendimento do processo de insolvência para cada setor econômico estudado.

Na elaboração dos modelos de previsão de insolvência dois são os tipos de dados que podem ser utilizados na elaboração de modelos de previsão de insolvência: dados sequenciais ou dados em painel. Nesta tese será estudada a influência desses dois tipos de dados na composição do modelo de previsão de insolvência, as conveniências e as inconveniências do uso desses tipos de dados em modelos de previsão de insolvência.

Na imensa maioria dos estudos de previsão de insolvência os modelos desenvolvidos apresentam bem mais capacidade de classificar melhor as empresas solventes do que as empresas insolventes, distanciando do objetivo principal que é a de classificar com mais eficácia aquelas empresas que podem vir a se tornar insolventes. Neste estudo o foco preponderante é melhorar a capacidade do modelo de caracterizar empresas com alto potencial de se tornar insolvente.

Nesta tese, também é estudado e proposto um comitê de classificadores (*ensemble*) visando atender o objetivo do parágrafo anterior de forma a tornar mais robustos os modelos para prever empresas potencialmente insolventes. Esta técnica, também muito pouco estudada em previsão de insolvência, tende a fortalecer e creditar ainda mais o modelo de previsão elaborado.

1.4 Principais contribuições

As principais contribuições deste trabalho são: (i) justificativa do uso de dados contábeis na amostra para elaboração do modelo de previsão de insolvência; (ii) reconhecimento do problema do balanceamento da base de dados e a proposta de um procedimento para equacionar este problema; (iii) reconhecer a necessidade das empresas insolventes requererem mais atenções discriminando as variáveis contábeis de acordo com o setor econômico no qual pertença a empresa; (iv) reconhecer a necessidade de sistematizar a seleção de atributos; (v) pesquisar e encontrar os melhores classificadores de acordo com a amostra de dados de empresas de um setor econômico específico; (vi) obtenção de regras de classificação para empresas de acordo com o setor econômico; (vii) reconhecimento da necessidade de diferenciar o tipo de dado, seqüencial ou temporal versus dados de painel por setor econômico para apurar a classificação; (viii) reconhecer a necessidade de focar o estudo visando melhorar a capacidade de caracterizar as empresas insolventes e, (ix) melhorar a apuração dos resultados utilizando técnicas de comitê de classificadores. Algumas perguntas poderão ser respondidas no final desta tese:

1) Por que dados contábeis são mais utilizados do que os dados colhidos no mercado para elaborar modelos de previsão de insolvência?
2) Como os conjuntos de dados contábeis de empresas com classes desbalanceadas devem ser tratados e por que devem se tratados?
3) Na modelagem de previsão de insolvência os dados contábeis devem ser de acordo com o setor econômico da empresa para apurar os resultados.
4) Qual é o conjunto de indicadores contábeis que podem revelar os sintomas de uma insolvência iminente? Das variáveis selecionadas para serem estudas quais serão aquelas em que sofrerão efeitos do tratamento de dados em classes desbalanceadas e se esses efeitos melhorarão a capacidade da previsão.
5) Quais são os classificadores mais adequados para melhor caracterizar as variáveis de acordo com o setor econômico da empresa?
6) Quais são as regras de classificação geradas para se entender o processo de insolvência de empresas brasileiras de capital aberto de acordo com setor econômico no qual a empresa pertence? Nas regras geradas, estas são distintas nos setores econômicos? E sendo, quais são essas diferenças para cada setor econômico?
7) Qual tipo de dado é mais eficiente para elaborar modelos de previsão de insolvência, dados temporais ou dados de painel? Quais são as vantagens e as desvantagens na elaboração de modelos com os respectivos tipos de dados.
8) Quais são os efeitos nos resultados para uma melhor acurácia de caracterizar as empresas insolventes aplicando técnicas de comitê de classificadores?
9) Por que da necessidade de focar o estudo nos dados das empresas insolventes?

1.5 Organização da tese

Esta tese está organizada da seguinte forma:

O capítulo 2 trata dos conceitos de previsão de insolvência. Nesse capítulo são apresentados os conceitos de insolvência, as diferenças existentes nas suas conceituações para diversos autores e as várias abordagens desses conceitos. É feito uma revisão bibliográfica de vários estudos desenvolvidos sobre o tema, citando

modelos elaborados através de técnicas tradicionais, modelos elaborados através de técnicas de inteligência computacional/aprendizagem de máquina e os modelos elaborados através de técnicas de *soft computing*.

É também feita neste capítulo uma discussão sobre os modelos elaborados com dados contábeis versus modelos elaborados com dados de mercado, as características dos dados que comporão os modelos com os bancos desbalanceados, as variáveis sequenciais ou variáveis em painel e segmentação das empresas por setores econômicos.

O capítulo 3 descreve a metodologia utilizada nesta tese. Nesse capítulo são apresentadas as técnicas desenvolvidas para construir os bancos de dados para as empresas solventes e insolventes; as técnicas de balanceamento da base de dados com um algoritmo proposto para o balanceamento; os algoritmos de seleção de atributos utilizados; o comitê de classificadores; a técnica da votação majoritária (*majority voting)*; a segmentação dos demonstrativos contábeis das empresas por setores econômicos; as técnicas de previsão; as regras de classificação e os métodos de validação utilizados.

O capítulo 4 apresenta a descrição da montagem das bases de dados. Estas bases são elaboradas de acordo com as características das variáveis utilizadas, sequenciais ou de painel. São descritas também as variáveis econômico-financeiras consideradas no estudo separando-as em três grandes grupos: variáveis de liquidez, de endividamento e de rentabilidade.

No capitulo 5 são avaliadas as bases de dados com as variáveis do tipo sequencial e de painel com os índices econômicos-financeiras obtidos através dos demosntrativos contábeis de empresas brasilerias de capital aberto. Inicialmente, aplicam-se algumas técnicas de classificação consideradas eficientes para avaliar o desempenho destas bases sem um tratamento especifico. Em seguida, utiliza-se o algoritmo desenvolvido para balancear a base de dados (SEID). Na etapa seguinte é feita a seleção de características através de várias técnicas de balanceamento (SEIDwS). Os resultados obtidos são comparados com os resultados da literatura especifica.

No capítulo 6 são apresentadas as análises estatísticas descritivas das variáveis das empresas por setor econômico. São estudados seis setores econômicos nesta tese e estão de acordo com a classificação do Bovespa no ano de 2007: (i) materiais básicos; (ii) consumo cíclico; (iii) consumo não cíclico; (iv) bens industriais; (v) construção e transportes e; (vi) tecnologia da informação e telecomunicações. Estes setores foram escolhidos para estudo devido aos seguintes motivos: facilidade no acesso das

informações contábeis de empresas solventes e insolventes; capacidade de representar bem o ambiente econômico das empresas brasileiras de capital aberto durante o período de tempo estudado e devido à qualidade e a quantidade de dados contábeis dessas empresas.

No capítulo 7 são apresentadas as regras de classificação geradas de acordo com cada grupo de empresas pertencentes por setor econômico. Estas regras podem facilitar o entendimento de uma provável dificuldade financeira que a empresa venha a enfrentar no futuro.

No capítulo 8 são apresentadas as conclusões, limitações do estudo e alguns possíveis estudos futuros.

CAPÍTULO 2: Previsão de Insolvência

Neste capítulo são apresentadas algumas considerações sobre previsão de insolvência e definições sobre o que vem a ser insolvência de empresas no aspecto contábil financeiro, assim como os principais e mais recentes estudos no mundo e no Brasil sobre modelos de previsão de insolvência. Também são discutidos trabalhos que compararam os resultados dos modelos financeiros baseados em dados contábeis e modelos financeiros baseados em dados do mercado. A importância da fase de pré-processamento para a modelagem de previsão de insolvência é analisada, assim como as características dos dados que comporão os modelos de previsão de insolvência sendo eles desbalanceados, temporais ou em painel e segmentados em setores econômicos.

2.1 A Insolvência de Empresas

São vários os especialistas em contabilidade e finanças que discutem o que vem a ser efetivamente insolvência de empresas, havendo definições bem distintas para entender insolvência. Entretanto, em um aspecto a grande maioria concorda, a insolvência está relacionada à existência de um potencial risco na continuidade da empresa. Nesta seção serão apresentadas algumas definições de alguns estudiosos em contabilidade e finanças.

Para Brealey e Myers, 1996 (p. 485) a insolvência ocorre quando os acionistas exercem a seu direito de não cumprimento. Este direito é valioso; quando uma empresa está em dificuldades à responsabilidade limitada permite aos acionistas o simples abandono da empresa, deixando todos os seus problemas aos credores. Os antigos credores passam a ser os novos acionistas, e os antigos acionistas ficam sem nada. A definição desses autores foca o interesse preponderante dos investidores, que aplicam seus recursos em investimentos visando obter retornos adequados aos seus objetivos não se importando, inclusive, com prováveis perdas ocorridas em um passado bem próximo. Eles preferem estas perdas de momento em contraste com futuros e possíveis acúmulos de perdas maiores. Para esses autores os acionistas não têm como meta o controle administrativo e gerencial da entidade, diante disso são os resultados, que viabilizam os retornos financeiros, que preponderam para os acionistas e a insolvência torna-se consequência da inviabilidade desses investimentos.

Para Ross *et al.*, 2002 (p. 325) insolvência é a dificuldade financeira máxima que uma empresa pode ter, nesta situação a propriedade dos ativos da empresa é legalmente transferida dos acionistas aos credores. Essas obrigações decorrentes de dívidas são fundamentalmente distintas das obrigações para com os acionistas. Embora os acionistas esperem receber dividendos, nas dificuldades financeiras eles não têm direitos garantidos como os credores têm.

Para Damodaran, 2002 (p. 244) insolvência ocorre quando os fluxos de caixa de uma empresa venham a se tornar insuficientes para atender aos seus compromissos de dívida (juros e principal). Embora tal situação não queira dizer automaticamente falência, ela leva a inadimplência, com todas as suas consequências negativas.

Para Weston e Brigham, 2000 (p. 811) uma empresa se torna insolvente, quando ela não tem dinheiro suficiente para atender aos pagamentos programados de juros e do principal. Portanto, deve-se optar entre dissolver a empresa por meio de liquidação ou permitir-lhe que se reorganize e, assim, permaneça viva.

Já Lopes e Martins, 2007 (p. 30) insolvência está mais relacionada com a incapacidade da empresa de se endividar do que com o seu desempenho operacional. Aqui a relevância para os autores é a capacidade financeira em detrimento ao operacional, a perda da capacidade de endividamento representa a perda da capacidade de continuidade, entende-se endividamento como obtenção de recursos com terceiros (empréstimos, financiamentos, etc). A causa desse estado é o atraso de obrigações e a consequente perda da confiança junto a terceiros na continuidade da empresa. Os autores entendem que é o nível de alavancagem financeira que irá preponderar na continuidade da empresa, levando ela à insolvência ou não.

Para Matarazzo, 2003 (p. 255) a insolvência de uma empresa ocorre pela incapacidade de solver suas obrigações, ou seja, pela falta de dinheiro no momento de vencimento de uma dívida. A falta crônica de dinheiro e a perspectiva de longo prazo no agravamento da insuficiência de entradas em caixa em face das saídas comprometidas caracterizam a insolvência, cujas soluções só podem ser a concordata ou a falência. Para o autor a falta de dinheiro de curto prazo deve-se, principalmente a cinco motivos: (i) desempenho de vendas aquém do esperado; (ii) falta de controle das despesas; (iii) prejuízos acumulados permanentes; (iv) má administração dos ativos e passivos circulantes; e (v) excesso de investimento no ativo permanente.

Kumar e Ravi (2007) adotam como insolvência a definição da *The U.S. Bankruptcy Code* (11 USC 101, Clause 32, http://www.solvency.com/solvency.htm)

que são as seguintes: 1) um excesso do passivo sobre o ativo; 2) uma inabilidade para o pagamento das dívidas de acordo com a maturidade normal do negócio e; 3) a cessação da corporação como resultado de interesse administrativo usual do negócio, manifestado por algum ato positivo de insolvência.

Platt e Platt (2002) definem como empresas insolventes aquelas em que apresentam os seguintes indicadores: (1) vários anos com lucro operacional negativo, (2) suspensão do pagamento de dividendos ou (3) várias reestruturações ou demissões.

Para McLeay e Omar (2000) a insolvência ocorre em empresas em que as vendas das ações da empresa geram prejuízos aos investidores, reestruturação de capital e reorganização da empresa durante vários períodos e, também durante vários períodos de perdas acumuladas para os acionistas.

Keasey e Watson (1991) mencionam que o critério financeiro de insolvência é incompleto e arbitrário e concluem que "pode haver a necessidade de desenvolver modelos específicos para diferentes tipos de dificuldades financeiras".

Além de falência e dificuldades financeiras, várias outras definições econômicas de insolvência são utilizadas em estudos de previsão de fracasso empresarial. Um primeiro exemplo é "liquidez de insolvência", que significa que a empresa é incapaz de pagar suas obrigações financeiras quando os pagamentos devem ser pagos (Laitinen, 1994). Esta definição de falência, escassez na liquidez, está intimamente relacionada com o processo jurídico de falência no mercado (Laitinen e Kankaanpää, 1999). Outra definição aplica o "padrão de empréstimo". Ward e Foster (1997) argumentam que o não cumprimento do empréstimo é a melhor maneira de definir o fracasso, porque é um evento economicamente definido, em oposição à falência, que é um caso juridicamente definido. Eles sugerem que a definição padrão de empréstimo é mais coerente com a realidade econômica. No entanto, é evidente que esta definição não apresenta limites ao estudo de previsão, faltando um contexto mais esclarecido de problemas de crédito ou empréstimo. Taffler e Agarwal (2003) apresentam outras definições que são baseadas em eventos como as reconstruções de capital, o encerramento de atividades operacionais de grandes proporções ou a necessidade forçada da alienação de grande parte da empresa, o apoio informal (ou formal) do governo e renegociações de empréstimos pactuadas por razões de solvência com banqueiros. Hayden (2003) no seu estudo apresenta que à luz do novo acordo da Basiléia II, alguns eventos padrões são bem definidos, tais como perda de crédito associado a qualquer atraso no pagamento de mais de 90 dias ou uma reestruturação urgente envolvendo o perdão ou o adiamento de

montantes de capital de acordo com interesses de instituições financeiras. É claro que esses eventos predefinidos ocorrem mais freqüentemente que a falência. No entanto, estas definições de falha são puramente de crédito e, portanto, não são adequados quando a falha de análise em um contexto de negócios (ou seja, falha das empresas), em vez de em um contexto de crédito.

Cabe abrir um espaço e apresentar o aspecto jurídico da insolvência. No Brasil a Lei nº 11.101, de 9 de fevereiro de 2005, trouxe significativas mudanças no direito falimentar brasileiro. O objetivo principal dessa lei foi garantir que as empresas tenham a possibilidade de se reorganizar, caso passem por crises financeiras que a ponham em situações difíceis. Entre as mudanças, vale serem citadas a criação das figuras da Recuperação Judicial e da Recuperação Extrajudicial, como tentativas de manutenção da unidade produtiva e de seus benefícios sociais como geração de riqueza, empregos e impostos, por exemplo. Desapareceu, por outro lado, o instrumento jurídico da Concordata regido pela legislação precedente. No caso da Falência, ocorreram significativas mudanças na classificação dos credores e no caso de alienação de ativos.

Empresas classificadas na CVM (Comissão de Valores Mobiliários) em um dos três estados (Recuperação Judicial, Recuperação Extrajudicial e Falência) mais a classificação de concordatárias, para empresas anteriores a Lei nº 11.101, foram nesta tese incluídas na amostra das insolventes.

2.2 Técnicas para previsão de insolvência

Nesta seção serão apresentados diversos estudos desenvolvidos sobre o tema previsão de insolvência com o uso de modelos estatísticos, técnicas de inteligência computacional e aprendizado de máquina. Desses estudos, há trabalhos que são somente baseados em dados contábeis, que são a maioria, e trabalhos que utilizam não somente dados contábeis, mas também dados extra contábeis.

2.2.1 Modelos estatísticos

A previsão de insolvência tornou-se um assunto mais pesquisado a partir de estudos publicados na década de 60, sobretudo após o artigo no qual apresentava o modelo Escore-Z de Altman (1968). Este modelo tentou superar as deficiências das análises que eram baseadas em um único índice utilizando análise discriminante. O uso

da análise discriminante leva a um grupo de índices com capacidade de separar empresas boas de empresas ruins e ao mesmo tempo determina o peso relativo de cada índice, sem que para determinação desses pesos prevaleçam critérios arbitrários.

O mesmo Altman *et al.*, (1977) desenvolveu um novo modelo de classificação de insolvência chamado Zeta, este é uma atualização e aprimoramento do modelo Escore-Z original. O objetivo era o de construir uma medida que refletisse explicitamente desenvolvimentos recentes envolvendo quebra de empresas. A precisão da classificação deste modelo ficou com mais de 90% para um período antes de ocorrer à insolvência. E mais de 70% de precisão para cinco períodos antes da insolvência.

Martin (1977) elaborou um modelo de previsão de insolvência em que utilizou regressão logística. Ohlson (1980) empregou modelo *logit* para previsão de falência de firmas. Dietrich e Kaplan (1982) desenvolveram um modelo linear com três variáveis para classificação de empresas em nível de risco. Eles comparam o seu modelo com (i) o do Altman (1977) e (ii) de Wilcox (1973) e obtiveram melhores resultados do que os dois autores. O modelo obteve uma precisão melhor do que outros modelos de predição de risco de empréstimos.

Zmijewski (1984) elaborou dois modelos de previsão de insolvência com técnicas diferentes e depois os testou. O primeiro modelo foi elaborado replicando os dados das empresas insolventes para compor a amostra. O segundo modelo utilizou uma amostra em que não houve reposição dos dados das empresas insolventes. O autor usou como classificadores na etapa de estimação *probit* e máxima verossimilhança, na etapa de validação foram empregados *probit* simples e *probit* bivariado.

West (1985) utilizou análise fatorial para compor as variáveis. Ele demonstrou que a combinação de análise fatorial com *logit* obteve resultados bem animadores. Karels e Prakash (1987) conduziram os estudos através de três etapas: (i) investigou a condição de normalidade dos dados contábeis requerido pela técnica de análise discriminante múltipla (MDA); (ii) quando os dados não apresentavam normalidade, eles construíam variáveis que fossem normais, ou seja, eles normalizam as variáveis; (iii) usaram essas variáveis para comparar os resultados de predição da técnica de análise discriminante (LDA).

Haslem *et al.*, (1992) determinaram as implicações das estratégias externas e internas da empresa adotados e que refletem nos demonstrativos contábeis. Eles também determinaram os impactos das estratégias de lucros. Eles utilizaram correlação canônica para analisar as relações existentes entre as variáveis. Kolari *et al.*, (2002)

desenvolveram sistemas denominados de alarmes (EWS) baseados em análise *logit*. Jones e Hensher (2004) apresentaram um modelo de previsão de falência baseado em *logit* misto e compararam com os modelos *logit* multinomial (MNL). Canbas *et al*., (2005) propuseram a integração do sistema de alarmes (IEWS) combinando MDA, regressão logística, *probit* e análise de componentes principais (PCA) que ajudou a examinar e detectar empresas com sérios problemas. PCA permite explorar três componentes financeiros, com esses três componentes ele explorou MDA, regressão logística e modelos *probit*. Ele combinou esses três classificadores e construiu o IEWS. Os resultados são muito elaborados para aqui serem apresentados. Eles concluíram que IEWS tem mais habilidade preditiva do que outros modelos.

Já no Brasil uma das principais dificuldades ainda é a escassez de pesquisas desenvolvidas com o propósito de encontrar parâmetros para previsão de insolvência, além da escassez de dados adequados e confiáveis para a realização deste estudo. Felizmente, essa situação começa a ser mudada, mas ainda se está bem longe de poder fazer esse tipo de trabalho com a facilidade de obtenção de dados como ocorre com outros países. A seguir serão apresentados alguns trabalhos que acabaram conquistando destaque no estudo sobre o tema no Brasil.

Elizabetsky (1976) empregou análise discriminante para desenvolver três modelos: o primeiro com 5 variáveis, o segundo com 10 variáveis e o terceiro com 15 variáveis. A partir de uma amostra de 99 firmas (do ramo de confecção) com problemas financeiros e outra de 274 firmas similares e sem problemas, analisou 60 índices, dos quais, através de análise de correlação, separou 38 para utilizar a análise discriminante. Kanitz (1978) usou análise discriminante e construiu o chamado termômetro da insolvência. Seguindo linha semelhante à dos trabalhos de Altman, procurou, através do cálculo do Fator de Insolvência (FI), que é o nome da função discriminante em seu modelo, identificar empresas em condições de insolvência. Enquanto o modelo Kanitz se baseia na liquidez, o de Altman tem como base o ativo total. Mesmo sendo seu modelo composto por três índices de liquidez, Kanitz considera que os mesmos têm capacidade de previsão bem melhor que os outros modelos.

Por meio da técnica estatística de análise discriminante, Matias (1978) trabalhou com 100 empresas de diversos ramos de atividade, sendo 50 solventes e 50 insolventes e, partindo da análise de 33 variáveis, desenvolveu uma função discriminante com 6 índices.

A metodologia do trabalho seguinte de Altman, Baidya e Dias (1979) foi análise

discriminante, tendo sido utilizadas 23 empresas com problemas financeiros (PS) e 35 empresas de mesmo porte e mesmo ramo sem problemas financeiros (NP). Foram utilizadas cinco variáveis, da mesma forma que no modelo original de Altman.

Em seu estudo realizado em 1982, Silva (2006, p.266) utilizou a também ferramenta estatística de análise discriminante, introduzindo novos índices financeiros. Foram desenvolvidos vários modelos, que consistem de um conjunto de índices financeiros, cujo processo de escolha foi fundamentado em métodos estatísticos para selecionar, entre índices existentes, aqueles cujo conjunto tem maior representatividade para classificar empresas com probabilidade de serem boas ou insolventes.

Bragança e Bragança (1984) combinaram métodos estatísticos (análise discriminante) com a análise de indicadores econômico-financeiros na previsão de concordata e na atribuição de uma nota que reflita a saúde financeira das empresas. Os autores inovaram na composição das amostras com dados obtidos no DOAR (Demonstrativo de Origens e Aplicações de Recursos) das empresas, principal contribuição do estudo.

Kasznar (1986) aplicou análise discriminante para desenvolver um modelo linear a partir de uma amostra de 55 empresas industriais insolventes com problemas entre 1978 e 1982 e 69 empresas industriais solventes equivalentes. A partir daí, foi criado um modelo com 5 variáveis.

Carmo (1987) utilizou modelos de funções lineares obtidas a partir de modelos fatoriais e de análise de componentes principais e também a partir da análise discriminante (utilização da análise discriminante apenas para efeito de comparação com os outros modelos de análise fatorial). A vantagem dos outros modelos testados sobre o modelo de análise discriminante é que eles não exigem separação prévia entre o grupo de empresas solventes de empresas insolventes. Matias e Siqueira (1995) utilizaram análise de regressão Logística para desenvolver um modelo que recebeu a denominação de “Modelo de Previsão de Insolvência Bancária”. Santos (1996) construíram um modelo fundamentado em análise discriminante capaz de fornecer, com grau de precisão estatisticamente aceitável, indicações sobre a saúde financeira de empresas industriais.

2.2.2 Inteligência computacional

Nesta seção serão revistos diversos artigos aplicando algumas das mais importantes técnicas de inteligência computacional com a finalidade de prever dificuldades financeiras em empresas.

2.2.2.1 Redes neurais

Odom e Sharda (1990) compararam a habilidade de predição de Redes Neurais Artificiais (ANNs) e análise de discriminante (LDA) no risco de falência. ANNs apresentaram um desempenho melhor e provaram ser mais robustas que o método de LDA em amostras de tamanho reduzido. Roy e Cosset (1990) também compararam ANNs com regressão logística para avaliar risco de insolvência em países usando dados econômicos e políticos. ANNs obteve menores valores nos erros absolutos nas avaliações de risco e foram mais sensíveis do que regressão logística quando das mudanças de variáveis. Duliba (1991) comparou ANNs com quatro tipos de regressão em dados de empresas de transporte para previsão de insolvência.

Várias pesquisas em previsão de insolvência incluindo Lacher *et al.*, (1995), Sharda e Wilson (1996), Tam e Kiang (1992) e Wilson e Sharda (1994) relataram que ANNs produzem significativas melhoras na acurácia dos modelos de predição comparados com aqueles elaborados com técnicas estatísticas.

Para Zhang *et.al.*, (1999) estudos sobre ANNs aplicadas em previsão de insolvência apresentam que ANNs são técnicas poderosas para reconhecimento de padrões e de classificações em que são inseridas variáveis explicativas não lineares e não paramétricas adaptando as propriedades de treinamento. Modelos com ANNs têm apresentado sucessos para vários problemas contábeis-financeiros incluindo previsão de falência (Trippi, 1993 e Zahedi, 1996).

Altman *et al.*, (1994) também compararam as técnicas de análise discriminante (LDA) e redes neurais para previsão de insolvência. Os resultados indicam um grau equilibrado de precisão e outras características benéficas entre LDA e ANNs. Concluem da necessidade de haver estudos adicionais e testes que usem as duas técnicas e sugerem uma combinação para reforçar o preditor. Boritz e Kennedy (1995) fizeram um estudo para examinar a efetividade de diferentes redes neurais de predizer falência. Lee, Han e Know (1996) desenvolveram modelos de ANNs híbridas de previsão de falência. Eles elaboraram três modelos: (1) ANNs com MDA; (2) redes com ID3; e (3) ANNs com

SOFM (*self organizing feature map*). As redes utilizaram MDA e ID3, que é algoritmo de árvore de decisão desenvolvido por Quinlan (Han e Kamber, 2006 p. 292), como selecionadores e classificadores de variáveis. O modelo redes com SOFM combina o modelo *back-propagation* para o treinamento supervisionado e com SOFM para treinamento não supervisionado. Ahn, Cho, Kim (2000) propuseram um sistema inteligente híbrido que prediz o fracasso de empresas baseado nos dados de desempenho financeiros passados, combinando *rough set* e rede neural. Com essa combinação o que se pretende é reduzir as variáveis qualitativas sem perda de informação.

Lee e Cheng (2005) fizeram um estudo com o propósito de avaliar um previsor de crédito modelando com redes neurais artificiais e *multivariate adaptive regression splines* (MARS). A razão desta análise é primeiramente usar MARS para construir o modelo de avaliação de crédito. As variáveis significantes obtidas compuseram o modelo utilizando rede neural.

Pendharkar (2005) desenvolveu uma TV-ANN (*threshold-varying*) de aproximação para um problema de classificação binária. A TV-ANN proposta de aproximação foi comparada, para sua aprendizagem e desempenho preditivo, usando dados simulados e dados do mundo real. Os resultados das experiências indicam um melhor desempenho da TV-ANN de aproximação em relação à BP-ANN tradicional e análise discriminante linear no treinamento e no teste de desempenho de predição.

Fioramanti (2008) mostrou que, graças ao teorema de aproximação universal, uma ANN com duas camadas podem ultrapassar no desempenho um *Early Warning System* tradicional na predição de uma crise de dívida se o usuário souber escolher o número certo de unidades “escondidas”, treinamento da rede, e um eficiente algoritmo de treinamento. Isto é possível porque a relação que une o indicador de crise de dívida e as variáveis explicativas é altamente não-linear, portanto, a flexibilidade da ANNs deve render resultados que são pelo menos tão bons quanto os resultados dos métodos tradicionais paramétricos.

Chen e Du (2009) utilizaram redes neurais e técnicas de mineração de dados para elaborar modelos de previsão de insolvência. Os autores argumentam que os executivos, de acordo com seus interesses de curto prazo, podem manipular demonstrativos e distorcer resultados de empresas que estão em vias de se tornarem problemáticas. Diante disso eles utilizam variáveis financeiras, não financeiras e análise fatorial para extrair variáveis adaptáveis na amostra inicial. Youn, Hyewon *et al.*, (2009)

desenvolveram modelos de previsão de insolvência de empresas de hotelaria coreanas baseados em redes neurais artificiais e regressão logística com variáveis financeiras.

2.2.2.2 Algoritmos evolutivos

Nesta seção serão revistos artigos aplicando algoritmos evolutivos com a finalidade de prever dificuldades financeiras em empresas.

Varetto (1998) em seu estudo comparou, para classificação e predição de previsão de insolvência, técnicas tradicionais de estatística como análise discriminante (LDA) e uma técnica de inteligência computacional, Algoritmo Genético (GA). Shin e Lee (2002) propuseram um GA no estudo de modelos para previsão de insolvência de empresa. A vantagem do GA é que ele é capaz de extrair regras de fácil entendimento para um usuário especialista. Os resultados preliminares mostraram que as regras extraídas para previsão de insolvência com modelos de GA foram promissores.

Kim e Han (2003) propuseram um método de *data mining* baseado em algoritmo genético para obtenção de regras de classificação para previsão de insolvência voltada para especialistas do tema. Os resultados apresentados obtiveram melhores acurácia e cobertura do que os resultados obtidos com as técnicas de redes neurais e regras de indução.

Salcedo-Sanz *et al.,* (2005) propuseram uma abordagem para prever falência de empresas seguro de vida com base na aplicação de programação genética (GP). GP é uma classe de algoritmos evolutivos, que opera codificando a solução do problema como uma população de árvores LISP (uma família de linguagens de programação concebida por McCarthy em 1959). Este tipo de algoritmo fornece uma saída de diagnóstico na forma de uma árvore de decisão com funções e dados.

Lensberg *et al.,* (2006) desenvolveram um modelo de programação genética com seis variáveis de uma amostra de 1136 empresas norueguesas falidas e não falidas. O modelo fornece uma visão para a complexa interação de fatores relacionados com a falência, especialmente o efeito do tamanho da empresa. Os resultados sugerem que a informação contábil, incluindo a avaliação do auditor é mais importante para grandes empresas do que para as pequenas empresas.

Etemadi *et al.,* (2009) investigaram a aplicação de Programação Genética (GP) para a modelagem de previsão de falências. GP foi aplicado para classificar 144 empresas iranianas falidas e não falidas na bolsa de Teerã (TSE). Em seguida, análise discriminante múltipla (MDA) foi utilizada para aferição do modelo GP. Modelo GP

alcançou 94% de taxa de treinamento de 90% nas amostras de validação (*holdout*), respectivamente, enquanto modelo MDA atingiu apenas 77% no treinamento e 73% nas taxas de validação.

O objetivo principal do estudo de Abdou (2009) foi investigar a capacidade da GP na análise de modelos de *credit scoring* em bancos egípcios públicos. O segundo objetivo foi comparar o GP com a análise *probit* (PA), uma alternativa viável à regressão logística e com a medida WOE (*weight of evidence*).

2.2.2.3 Máquinas de vetor suporte

Huang *et.al.,* (2004) apresentaram máquinas de vetor suporte (SVM) para o problema de modelagem em previsão de insolvência na tentativa de fornecer um modelo com melhor poder explicativo. Foi usadas redes *backpropagation* (BPN) como referência e exatidão, tendo sido obtido de predição cerca de 80% para ambos os métodos BPN e SVM com dados de empresas dos Estados Unidos e dos mercados de Taiwan. No entanto, foi observada apenas uma ligeira melhoria com SVM.

Sinh *et al.,* (2005) investigaram a eficácia da aplicação de máquinas de vetor suporte para o problema de previsão de falências.

Shin, Lee e Kim (2005) mostraram que o classificador SVM supera BPN para problemas de previsão de falências de empresas. Os resultados demonstraram que a precisão e o desempenho de generalização do SVM foram melhores do que o BPN, sendo que o tamanho do conjunto de treinamento pode ser menor. Também foi analisado o efeito da variabilidade do desempenho em relação a diversos valores de parâmetros do SVM.

Min e Lee (2005) aplicaram máquinas de vetor suporte para o problema de previsão de falência sugerindo um novo modelo com melhor poder explicativo e estabilidade. Com essa finalidade, foi feito uma pesquisa usando 5 subamostras de validação cruzada para descobrir os valores do parâmetro ótimo de funcionamento do kernel do SVM. Além disso, para avaliar a acurácia de predição do SVM, foi comparado seu desempenho com os da MDA, análise de regressão logística (*logit*), e BPN. Os resultados mostram que SVM superou os outros métodos.

Min, Lee e Han (2005) propuseram métodos para melhorar o desempenho da SVM em dois aspectos: a seleção de atributos e otimização de parâmetros. GA é usado

para otimizar tanto um subconjunto de atributos e parâmetros do SVM simultaneamente para previsão de falências.

Chen e Shih (2006) propuseram um modelo de classificação automática para as classificações de crédito, um ranqueamento de crédito, através da aplicação de máquina de vetor de suporte. Foram utilizadas três novos grupos de variáveis: informações sobre o mercado de ações, o apoio financeiro do governo, e dos grandes acionistas, visando aumentar a eficácia da classificação. Pesquisas anteriores raramente consideravam essas variáveis.

Para Hua *et al.,* (2007) máquina de vetor de suporte foi aplicado no problema de previsão de falências, e provou ser superior aos métodos concorrentes, como a rede neural, as múltiplas abordagens discriminante linear e regressão logística.

Lee (2007) aplica máquinas de vetor suporte para o problema de ranqueamento de crédito corporativo em uma tentativa de sugerir um novo modelo com melhor poder explicativo e estabilidade. Para esta finalidade, o pesquisador utiliza uma base de dados com 5 desdobramentos na validação cruzada para descobrir os valores do parâmetro da função de kernel RBF do SVM. Os resultados mostram que o experimento com SVM supera o de outros métodos.

Ding, Song e Zen (2008) desenvolveram um modelo de previsão de insolvência em máquina de vetor suporte para um exemplo de empresas chinesas de alta tecnologia. Em geral, o SVM proporciona um modelo robusto com alta precisão para a previsão de socorro financeiro às empresas chinesas.

Kim e Sohn (2009) com máquina de vetor suporte elaboraram um modelo para prever insolvência das PME (pequenas e médias empresas no setor de tecnologia), considerando variáveis de entrada, tais como índices financeiros, indicadores econômicos e fatores de avaliação da tecnologia. Os resultados mostram que o desempenho de precisão do modelo SVM é melhor do que o BPN e regressão logística.

2.2.2.4 Outras técnicas

Alguns artigos abordam outras técnicas de estatística e aprendizagem de máquina aplicados nos estudos de previsão de insolvência como regras de indução, árvore de decisão, análise envoltória de dados, raciocínio baseado em casos (CBR) e *logit* quadrático.

Dimitras, Slowinski, Susmaga, Zopounidis (1999) em seu estudo, usam regras de indução para fornecer um conjunto de regras capazes de discriminar entre firmas saudáveis ou não a fim de prever insucesso empresarial. Os resultados são muito encorajadores, comparados com os da análise discriminante e *logit*, e provam a utilidade do método proposto para a previsão do fracasso empresarial.

Tay e Shen (2002) demonstraram que os modelos com regras de indução são aplicáveis a uma ampla gama de problemas práticos relacionados com a previsão econômica e financeira. Além disso, os resultados mostram que esses modelos são uma alternativa promissora aos métodos convencionais de previsão econômica e financeira.

Park e Han (2002) através do CBR, que é uma metodologia para resolução de problemas e tomada de decisões em ambientes complexos e de negócios, desenvolveram um modelo de previsão de insolvência. A modelagem com CBR não se mostrou suficiente e sendo necessário o emprego de outras técnicas mais elaboradas. Todas estas são áreas para futuras pesquisas.

Chen *et al.,* (2009) em seu artigo procuram oferecer uma alternativa para a modelagem de previsão de falência utilizando neuro *fuzzy*, uma abordagem híbrida que combina a funcionalidade de lógica *fuzzy* e da capacidade de aprendizagem das redes neurais. Os resultados empíricos mostraram que esta técnica demonstra uma melhor taxa de precisão, menor custo do erro de classificação e maior poder de detecção do que regressão *logit*, concluindo que a metodologia neuro *fuzzy* pode ser uma grande ajuda no alerta de falência iminente.

Nwogugu (2006) em seu artigo apresenta vários modelos dinâmicos de falência, recuperação e de tomada de decisão. Desenvolve um quadro e subsídios para a investigação utilizando sistemas dinâmicos e inteligência computacional na modelagem de falência, tomada de decisões e o raciocínio jurídico. Os resultados se apresentaram bem competitivos com as técnicas mais utilizadas.

Premachandra, Bhabra, Sueyoshi (2009) propuseram análise envoltória de dados (DEA) como uma rápida e fácil ferramenta para avaliar a falência da empresa. DEA é uma estimativa de variáveis de uma função de classificação para a separação das empresas solventes das empresas insolventes. Para os autores, DEA é um método eficaz para a avaliação de falência.

Tseng e Li (2005) propuseram um modelo *logit* quadrático (ou análise de regressão logística de intervalo quadrático) com base em uma abordagem de programação quadrática para processar variáveis de resposta binária. Os resultados

mostram que este modelo pode apoiar o modelo *logit* para discriminar os grupos, e fornece mais informações para os investigadores.

2.3 Dados de mercado versus dados contábeis

Nos estudos de modelagem em previsão de insolvência, os dados podem ter a origem nos registros contábeis (*accounting-ratio-based models*), nos dados de mercado (*market-based models*) ou utilizando os dois conjuntamente.

Tradicionalmente estudos sobre este tema são realizados com dados baseados em registros contábeis como os de Altman (1968) Escore-Z e no Brasil com Kanitz (1978). O uso de tais opções de dados sofre algumas criticas de lado a lado visando obter uma melhor capacidade de caracterizar as informações que eles geram. Dados estritamente do mercado são amplamente contestados devido à presunção da eficiência do mesmo.

A eficiência do mercado tem sido objeto de estudo da contabilidade desde que o trabalho de Fama (1970) foi publicado. Para Fama, o mercado seria mais ou menos eficiente em razão da reação a informação. As EMH (Hipóteses de Eficiência de Mercado) têm sido objeto de estudo pelos pesquisadores. Fama classificava suas hipóteses em três grupos: fraca, semi-forte e forte. Um grupo extenso de pesquisas surgiu no sentido de classificar o mercado num destes grupos. Ou seja, verificar em situações específicas se o mercado agia de forma mais ou menos eficiente. Dentre os estudos em que questionam a eficiência do mercado pode ser citado Haugen e Baker (2008) onde há evidências de que os mercados de ações dos EUA são altamente ineficientes. Timmermann e Granger (2004) comentam a contradição: os modelos do mercado pressupõem a sua eficiência, onde a existência do chamado “lucro anormal” não seria possível; mas os “práticos” usam para tentar obter este lucro.

Já em 1984 Mensah sugere que modelos elaborados com dados contábeis devem ser utilizados, acrescentou, todavia, da necessidade do seu aperfeiçoamento periódico.

Hillegelst *et al.*, (2004) compararam modelos de previsão de insolvência de Altman (1968), Black e Scholes (1973) e Ohlson (1980). Os resultados mostraram que o modelo Black e Scholes, que é elaborado com dados de mercado, foi mais eficiente do que os modelos tradicionais de Altman e Ohlson. Entretanto, para obter tais resultados foi necessário adotar várias modificações em relação aos modelos tradicionais como a

atualização dos coeficientes, ajustes dos dados para a indústria, separação dos dados de acordo com o defasamento temporal e mudanças de componentes.

Altman (2005) comparou os efeitos, para avaliação de crédito em mercados emergentes, do uso de dados com a origem nos livros contábeis e com os de origem no mercado. O autor utilizou dados de empresas do mercado mexicano. Para o autor, o impacto da utilização do valor de mercado contra o valor contábil pode ser considerável na classificação final de crédito para uma empresa. Tais efeitos podem, contudo, refletir a frequente ineficiência do mercado para empresas mexicanas de capital aberto. A volatilidade do peso mexicano e seus impactos sobre o mercado de capital mexicana pode mascarar os valores intrínsecos de títulos. Além disso, a inflação pode distorcer o valor contábil dos demonstrativos contábeis das empresas mexicanas por causa de "valores de renda não declarados" e as suas consequências nos lucros acumulados e no patrimônio líquido. Apesar destas deficiências, o mercado acionário mexicano foi suficientemente eficaz e abrangente o suficiente para agregar valor ao modelo no período pós crise do ano de 1994.

Reisz e Perlich (2007) estudaram a eficiência dos modelos de previsão de insolvência utilizando dados contábeis e dados de mercado. No modelo com dados de mercado, os autores adaptaram os modelos estruturais de Black e Scholes (1973) e Merton (1974) para empresas industriais americanas que apresentavam valores consideráveis em seus ativos imobilizados. A escolha destas empresas se deve às barreiras de entrada de novas empresas e que são também significativamente positivas na alavancagem da empresa e na volatilidade da empresa. Este modelo permitiu um melhor poder discriminatório do que os inferidos pelos modelos padrão Black e Scholes/Merton. No entanto, modelos com dados contábeis como o de Escore-Z de Altman superam os modelos estruturais em um ano posterior de previsões de falência, mas perdem relevância quando o horizonte de previsão é estendido.

Agarwal e Taffler (2008) compararam o desempenho do modelo Escore-Z (dados originados de registros contábeis) de Taffler (1984) com modelos de Hillegeist *et al.,* (2004) e Bharath e Shumway (2004) baseados no Black e Scholes (1973), elaborados com dados de mercado. O período estudado foi de 17 anos, de 1985 até 2001, e a área sob a curva ROC (*Receive Operating Characteristic*) foi usada como técnica de avaliação. Os autores tomaram como referência os estudos desenvolvidos por Stein (2005) e Blochlinger e Leippold (2006). Estes autores, em seus estudos,

demonstraram que “pequenas diferenças nos resultados obtidos no modelo de previsão de insolvência podem significar grandes impactos econômicos na sua aplicação”.

Os resultados encontrados por Agarwall e Taffler foram que em termos de acurácia de predição, ocorrem pequenas diferenças entre os modelos elaborados com dados de mercado e dados originados nos registros contábeis. Eles concluíram que a despeito das críticas existentes no uso de dados contábeis na elaboração de modelos de previsão de insolvência, estes modelos se mostraram mais eficientes do que aqueles elaborados com dados de mercado. De fato, modelos baseados em dados contábeis produzem significantes benefícios econômicos em relação a modelos elaborados com dados de mercado.

Chien-Chiang Lee, Jun-De Lee e Chi-Chuan Lee (2009) investigaram a hipótese de mercado eficiente em diferentes países desenvolvidos e em desenvolvimento econômico, inclusive o Brasil, durante o período de janeiro de 1999 a maio de 2007. Eles empregaram dados de painel aplicando testes de estacionariedade (a série de observações é invariante com respeito ao tempo). Os resultados encontrados indicam um nítido contraste com os encontrados na literatura existente, evidenciando que os preços reais das ações são processos estacionários sendo, por isso, inconsistentes com a hipótese de mercado eficiente.

Chen e Cheng (2009) em seu estudo visando avaliar o desempenho de algumas empresas para incluí-las ou não em seu portfólio de investimentos relatam duas análises para previsão de sucesso financeiro: análise técnica e análise fundamentalista. A análise técnica examina os dados históricos do mercado para identificar correlações entre preços de ativos e obter a previsão de preços futuros. A análise fundamentalista usa dados reais financeiros (registros contábeis). Para os autores a análise técnica apresenta resultados que melhor refletem sobre a situação real das empresas em particulares. A análise fundamental gera resultados que mostram as relações de causas e efeitos das variações ocorridas nos valores dos ativos. Para os autores a análise técnica é preferível para previsões de curto prazo. Análise fundamentalista é extensamente usada para previsões de períodos intermediários e longos. Os autores utilizam a análise fundamentalista para extrair decisões significativas para classificação do incremento das receitas para investidores de períodos intermediários e longos.

Nesta tese serão utilizados dados originados em registros contábeis (*accounting-ratio-based models*) de empresas brasileiras de capital aberto durante o período de 1996 até 2006 publicados na Comissão de Valores Mobiliários – CVM.

2.4 Pré-processamento dos dados

Sabe-se que a fase de pré-processamento de dados é bem extensa e envolve a identificação de diversos tipos de problemas e fases que podem se manifestar na etapa de descoberta de conhecimento (KDD). Para Pyle (1999) ela consome, em média, 90% do tempo necessário para o completo processo de KDD. Já para Han e Kamber (2006) esse tempo precedente consumido na etapa de pré-processamento é de fundamental importância para a etapa seguinte a de descoberta de conhecimento. É mais do que uma necessidade tediosa, as técnicas utilizadas na etapa de pré-processamento podem influenciar profundamente os resultados da etapa seguinte, a aplicação efetiva de um algoritmo de mineração de dados (DM). Por isso, o papel do impacto do pré-processamento com DM vem ganhando cada vez mais interesse ao longo dos anos. A etapa de pré-processamento está se tornando cada vez mais poderosa, mais rápida e mais transparente a cada dia, pode-se dizer que, em um futuro bem próximo haverá uma maior iteratividade controlada pelo usuário, por isso há muito a estudar na transição da etapa de pré para DM (Kriegel H-P *et al.*, 2007).

Ao contrário dos conjuntos de dados presentes em repositórios de dados, estes extraídos diretamente de sistemas de gerenciamento de dados freqüentemente apresentam diversos problemas, tais como: grande quantidade de ruído e inconsistências (atributos com valores incorretos); atributos de baixo valor preditivo; excesso de valores desconhecidos; classes desbalanceadas, ou seja, uma grande desproporção entre as distribuições das classes, base de dados relativamente pequenas, entre outros. Na maioria dos casos os dados do mundo real apresentam diversas imperfeições antes da extração dos padrões.

Embora muitos dos algoritmos utilizados na fase de DM tenham sido projetados para manipular dados em tais situações, pode-se esperar que esses algoritmos gerem resultados mais precisos caso a maioria dos problemas presentes nos dados tenham sido, removido ou corrigido gerando com isso conhecimentos mais representativos e mais preditivos.

Tarefas de pré-processamento de dados podem ser tipicamente solucionadas por métodos que extraem do próprio conjunto de dados as informações necessárias para tratar o problema.

2.5 Características dos dados

Na elaboração de modelos de previsão de insolvência, as principais características normalmente encontradas nos dados, na etapa de pré-processamento, que irão compor os modelos são: dados desbalanceados, dados sequenciais, dados de painel e dados setoriais. Esta seção será subdividida em subseções em que serão apresentados a conceituação e trabalhos realizados sobre cada assunto.

2.5.1 Bancos desbalanceados

Base de dados com distribuições de classes desequilibradas ou desbalanceadas são bastante comuns em muitas aplicações reais dentre elas está previsão de insolvência, onde o número de exemplo da classe de empresas insolventes é bem inferior ao número de exemplos de empresas classificadas como solventes.

Muitos sistemas de aprendizado de máquina assumem que as classes estão balanceadas e, dessa forma, esses sistemas falham em induzir um classificador que seja capaz de predizer a classe minoritária com precisão na presença de dados com classes desbalanceadas. Freqüentemente, o classificador possui uma boa precisão para a classe majoritária, mas uma precisão inaceitável para a classe minoritária.

O problema agrava-se ainda mais quando o custo de classificação incorreta da classe minoritária é muito maior do que o custo de classificação incorreta da classe majoritária.

Para Estabrooks e Japkowicz (2004) a solução de problemas com classes desbalanceadas podem ser categorizados em dois grupos: o interno, no qual cria novos algoritmos ou modifica os existentes para o problema de desbalanceamento considerado e o externo, que não modifica o algoritmo, mas faz uma reamostragem dos dados presentes com esses algoritmos para diminuir os efeitos causados pela classe desbalanceada.

Portanto, o tratamento de conjuntos de dados com classes desbalanceadas, por ser um problema mais discutido recentemente e pouco encontrado em estudos de previsão de insolvência, precisa ser superado para se obter uma melhor acurácia na classificação e na previsão na classe de empresas insolventes.

2.5.1.1 Estudos referentes a treinamento de bancos de dados desbalanceados

Neste subitem serão apresentados estudos relevantes e mais contemporâneos sobre o tratamento de bancos de dados desbalanceados.

Kubat e Matwin (1997) discutiram critérios para avaliar a utilidade dos classificadores induzidos propondo um método de *under- sampling* em que os exemplos são divididos em quatro categorias: ruídos, redundantes, próximos da borda e seguros.

Pelos resultados encontrados pode se concluir que o desempenho dos classificadores não são afetados quando os exemplos redundantes são removidos. Removendo os exemplos negativos redundantes há uma significante redução no número de sua acurácia, mas não afetando, demasiadamente, o seu resultado total.

Domingos (1999) propôs um método para fazer um classificador sensível ao custo através de um procedimento que o minimize. Este procedimento chama-se meta custo, trata o classificador como uma "caixa preta", desconhecendo o seu funcionamento. Diferente da estratificação – mudança da freqüência das classes de treinamento na proporção de seus custos – meta custo é aplicado para alguns números de classes. Em testes empíricos com grandes bases de dados o meta custo quase sempre tem apresentado uma grande redução dos custos comparado com o classificador C4.5 e duas formas de estratificação. O meta custo pode ter eficiência quando aplicado em grandes bases de dados.

Japkowicz (2000) discutiu basicamente duas questões: (1) que tipos de bancos desbalanceados lesam a acurácia do desempenho do classificador? (2) Como vários métodos categóricos podem resolver o problema?

Na conclusão do estudo *multilayerperceptron* – MLP - não foi sensitivo para problemas com classes desbalanceadas quando aplicados em domínios linearmente separados, a sensitividade aumenta com a complexidade do domínio. O tamanho da amostra de treinamento não parece ser um fator preponderante.

O estudo também constatou que os métodos *over-sampling* na classe minoritária e *under-sampling* na classe majoritária apresentam melhoras nos resultados sendo métodos muito eficientes. Para tratar com esse problema foi visto que usando *over-sampling* mais sofisticado ou *under-sampling* com distribuição uniforme randomicamente não geram melhores resultados, parecendo ser desnecessário aplicar estas técnicas.

Japkowicz e Stephen, (2002) estudaram a influência, em classes desbalanceadas, do desempenho de alguns classificadores como árvore de decisão (AD), MLP e SVM.

Para os autores: o problema de classes desbalanceadas depende de: i) o grau da classe desbalanceada; ii) a complexidade da concepção representada pelos dados; iii) o tamanho total do conjunto de treinamento; iv) o classificador envolvido.

Os classificadores não foram sensíveis para o problema do desequilíbrio de classes da mesma maneira, C5.0 (AD) foi mais sensível dos três, MPL foi o que apresentou os melhores resultados a seguir e manifesta diferentes padrões de sensibilidades. O SVM veio por último dos que foram estudados por não apresentar sensibilidade total ao problema.

Para os classificadores sensíveis ao problema do desequilíbrio de classes, o estudo mostrou que o método simples de reamostragem pode ajudar bastante visto que, em determinadas circunstâncias, os classificadores foram bem sensíveis aos desequilíbrios de classe.

Chawla *et al.,* (2002) aplicaram nas classes minoritárias técnicas de *over-sampling* – SMOTE – e mostraram que uma combinação do método de *over-sampling* da classe minoritária *under-sampling* da classe majoritária pode conseguir um melhor desempenho do classificador (área ROC) do que somente *under-sampling* da classe majoritária. Os autores propõem *over-sampling* na classe minoritária através da criação de exemplos "sintéticos" o bastante para ser equivalente ao *over-sampling* com reposição.

Esta abordagem foi inspirada em uma técnica que obteve êxito no reconhecimento de caracteres manuscritos (Ha e Bunke, 1997).

Os resultados apresentados pelo SMOTE melhoraram a acurácia dos classificadores na classe minoritária. SMOTE melhora novas aproximações para *over-sampling*. A combinação do SMOTE e *under-sampling* melhorou o desempenho em relação ao *under-sampling* simples.

Maloof (2003) investigou o treinamento de bancos de dados quando são desbalanceados e apresentam custos desiguais e desconhecidos. O autor sugere que a melhor métrica para avaliação do modelo é analisar a curva ROC.

Jo e Japkowicz (2004) sugeriram que a degradação de certas classificações são causadas não somente pelo desbalanceamento da base de dados, mas pode ser causada, também pelos pequenos disjuntos (subconjuntos com baixa correlação entre o conjunto

maior no qual está classificado). Devido a isso o foco principal da melhora do desempenho da classificação deve ser voltado para os pequenos disjuntos.

Weis (2004) estudou a relação que classes raras e casos raros representam na mineração de dados. Os problemas que podem resultar destas duas formas de raridade são descritos em detalhes, como são apresentados soluções destes problemas. O estudo também demonstra que classes raras e casos raros são fenômenos muito similares – ambas as formas de raridade apresentam problemas similares durante a mineração de dados.

Batista *et al.,* (2004) evidenciaram que classes desbalanceadas não são sistematicamente o motivo da influência na performance do sistema de treinamento. O problema ocorre no treinamento com pequenas minorias devido à presença de outros fatores complicadores, como sobreposição de classes. Dois métodos propostos pelos autores lidam diretamente com essas condições; juntos o método *over-sampling* com o método de "limpeza de dados" produz uma melhor definição das classes de clusters. Outra experiência apresentada, em geral, é que o método de *over-sampling* gera mais acurácia nos resultados do que o método de *under-sampling* considerando a área ROC. Este resultado parece contradizer resultados previamente publicados na literatura. Dois dos métodos propostos combina o SMOTE com Tomek e ENN apresentaram resultados muito bons para conjuntos de dados com um pequeno número de exemplos positivos.

Raskutti e Kowalczyk (2004) exploram os limites do treinamento supervisionado de duas classes discriminadas com dados de classes bem desbalanceadas. O foco do estudo do treinamento supervisionado é feito com SVM. É considerado o impacto de ambas as amostras, uma compensação nas classes desbalanceadas e também um balanceamento extremo quando uma das classes é ignorada completamente e o treinamento é acompanhado usando exemplos para classes únicas. Estudou-se também o impacto da seleção de atributos nos resultados.

Weiss *et al.*, (2007) compararam três métodos de tratamento de bancos de dados desbalanceados e que apresenta custos nos erros de classificação não uniforme. No primeiro método incorpora os custos do erro de classificação no algoritmo de treinamento enquanto os outros dois métodos empregam *over-sampling* ou *under-sapling* para fazer o treinamento dos bancos, mas balanceados. O estudo comparou empiricamente os efeitos destes métodos na ordem de determinar qual produz os melhores resultados de classificação e sob quais circunstâncias.

2.5.1.2 Estudos referentes a treinamento de bancos de dados desbalanceados para previsão de insolvência de firmas

Atiya (2001) desenvolveu um estudo sobre previsão de insolvência no qual apresenta duas contribuições. A primeira é a aplicação de redes neurais no modelo com bancos de dados desbalanceados. No segundo, no modelo a ser desenvolvido por redes neurais são propostos novos indicadores na elaboração do modelo. O autor apresenta uma melhora nos resultados com a adição desses novos indicadores aos tradicionais.

David *et al.*, (2005) investigaram três estratégias de comitê de classificadores em aplicações de decisões financeiras incluindo previsão de insolvência, visando obter modelos com maiores acurácia: validação cruzada (*cross validation*), *bagging* e *boosting*. Os autores empregaram MLP como classificador. Os resultados desta pesquisa confirmaram que comitê de classificadores de redes neurais é mais preciso e mais robusto do que o "único melhor" modelo de MLP.

Hung *et al.*, (2007) aplicaram probabilidade híbrida baseada em comitê de classificadores para previsão de insolvência utilizando votação majoritária (*majority voting*) e votação ponderada (*weighted voting*). Para os autores esta proposta do comitê de classificadores supera outras propostas utilizando a ponderação (*weighting*) ou a estratégia do voto (*voting strategy*).

Huang *et al.*, 2007 investigaram três estratégias para construção de modelos de híbridos baseado no SVM para *credit scoring*[3] e compararam suas performances com redes neurais (BPN), algoritmo genético (GA) e árvore de decisão (C4.5). Os autores obtiveram as seguintes conclusões:

(1) A abordagem baseada em SVM para modelos de *credit scoring* pode ser apropriada para classificações que são aplicações em que há duas classes, solventes ou insolventes, desse modo minimização de risco de crédito economiza recursos consideravelmente em transações futuras.

(2) É evidente que modelos baseados no SVM são muito competitivos com BPN e GA em termos de acurácia na classificação. Comparando com GA e BPN, modelos baseados em SVM para *credit scoring* podem atingir acurácia na classificação semelhantes.

[3] *Credit-scoring*: Sistemas de pontuação do risco de crédito. Termômetro de crédito. É uma fórmula que combina índices obtidos nos demonstrativos contábeis, ponderados de acordo com a técnica utilizada, estatística, inteligência artificial, por exemplo. Trata-se de uma ponderação relativamente complexa, mas que pode ser aplicada a qualquer empresa. Depois de realizadas as operações indicados na fórmula, obtém-se o *credit scoring*.

(3) Modelos baseados no SVM também têm acurácias similares na literatura. Ong *et al.*, (2005) reportaram que a acurácia do GA, BPN e C4.5 são 88,27%, 87, 93% e 87,06%, respectivamente, para um dos bancos de dados testados e são 77,34%, 75,51% e 73, 17%, respectivamente para outra base de dados testado.

Tsai e Wu (2008) estudaram a desempenho de um classificador simples de redes neurais com os (diversificado) múltiplos classificadores baseados em redes neurais. Teoricamente, os múltiplos classificadores deveriam ter desempenho melhor do que os classificadores singulares. Entretanto, considerando os resultados experimentais nas médias das acurácia de predição, o classificador múltiplo com redes neurais não apresenta um melhor desempenho do que os classificadores singulares em alguns casos. No particular, os resultados mostram que os classificadores singulares são mais adequados do que os classificadores múltiplos ou diversificados para previsão de insolvência ou *credit scoring*. Por outro lado, examinando os erros tipo I (representam as instâncias classificadas incorretamente como insolventes) e tipo II (representam as instâncias classificadas incorretamente como solventes) destes classificadores, não existe um resultado que se destaque em relação ao outro. Nestes casos, a decisão deverá ser feita considerando a combinação dos múltiplos classificadores para previsão de insolvência e *credit scoring*, além dos classificadores singulares.

Considerando os resultados experimentais, existem duas questões que devem ser discutidas em que os classificadores múltiplos não tiveram melhores performances do que os classificadores singulares. Primeiro, a divisão das amostras de treinamento podem ter sido pequenas para serem exploradas nos classificadores múltiplos e fazendo com que os classificadores múltiplos virem a ter um desempenho ruim. Segundo, no problema de domínio com classificação binária como previsão de insolvência e *credit scoring*, classificadores singulares talvez sejam mais estáveis. Em outras palavras, nos classificadores múltiplos e nos classificadores diversificados múltiplos podem não melhorar a desempenho da classificação binária.

Yu L. *et al.*, (2008) neste estudo, uma ANNs é proposto para avaliar o risco de crédito ao nível medição. O modelo proposto é composto de seis etapas. Na primeira fase, um *bagging* na amostragem inicial é aplicado para gerar diferentes dados devido à escassez de dados. Na segunda fase, a rede neural gera diferentes modelos com diferentes formações de subconjuntos obtidos a partir da etapa anterior. Na terceira fase, os modelos gerados pelas ANNs são treinados com diferentes bancos de treinamento e

conferindo a classificação da medição e a confiabilidade do valor classificado pode ser obtida. Na quarta fase, um algoritmo de correlação é usado para selecionar o conjunto adequado de membros. Na quinta etapa, os valores selecionados pela rede neural para o modelo (ou seja, conjunto membros) são dimensionados para um intervalo por uma transformação logística. Na fase final, o conjunto selecionado é agregado para obter classificação final por meio da medição.

Os autores tiraram as seguintes conclusões, a proposta do modelo do *comitê de classificadores* de rede neural, consistentemente, superam os outros modelos comparáveis. Incluindo três modelos únicos, dois modelos híbridos e de maioria de voto com base em conjunto modelo. Estes resultados obtidos revelam que a proposta do comitê de classificadores de rede neural pode fornecer uma promissora solução para análise de risco de crédito e isso implica que a proposta do comitê de classificadores de ANNs tem um grande potencial para outros problemas de classificação com classes binária.

Ravi V. *et al.*, (2008) elaboram e testam comitê de classificadores para previsão de insolvência. Tais comitês tratam-se de modelos que são constituídos de *multi-layered feed forward* neural network *trained with backpropagation* (MLFF-BP), a *probabilistic neural network* (PNN) e a *radial basis function neural network* (RBFN), *support vector machine* (SVM), *classification and regression trees* (CART) e a *fuzzy rule based classifier* Além disso, a análise componente principal (PCA) com base híbrida redes neurais, a saber. PCA - MLFF - BP, PCA - PNN e PCA - RBF também estão incluídas como componentes do conjunto. Além disso, GRNN e PNN foram treinados com um algoritmo genético para otimizar os parâmetros. Dois conjuntos: (i) a votação com base na maioria simples (*majority voting*) e (ii) *weightage* são implementadas.

2.5.2 Variáveis sequenciais versus variáveis em painel

Modelos de previsão de insolvência podem ser elaborados através de duas abordagens: empregando-se dados com variáveis sequenciais ou dados com variáveis de painel.

Dados sequenciais, também chamados de dados temporais, podem ser pensados como uma extensão de dados de registros, onde cada registro possui tempo associado a ele (Tan, Steinbach e Kumar, p. 35, 2006). Um conjunto de dados para previsão de insolvência registra o tempo no qual a variável representa. Modelos elaborados com este

tipo de variáveis têm a capacidade de definir as variáveis nos seus respectivos períodos de uma provável insolvência da empresa. Em geral, a deterioração financeira, que leva à falência aparece explicitamente no prazo de cerca de 3 anos antes da ocorrência (Ravi, *et al.*, 2008). A base de dados é elaborada com as variáveis em função do período (ano) da declaração de insolvência. Por isso modelos com variáveis pertencentes ao período de ocorrência da insolvência se tornam inviáveis de serem aplicados, apesar de, quando estão presentes nos modelos, são os que apresentam as melhores capacidades de influenciar nos resultados na classificação da amostra de origem nos testes (Horta, 2001 e Silva, 2006). As variáveis pertencentes aos 2º e 3º anos anteriores a decretação da insolvência são freqüentemente utilizados em pesquisas e estudos sobre o tema, como tem sido provado ser capaz de prever falência também com boa eficácia (Altman, 1968, Martin, 1977 e Canbas *et al.*, 2005).

Já os modelos que utilizam dados em painel oferecem uma serie de vantagens em relação aos modelos de séries temporais, pois estes modelos permitem controlar a heterogeneidade presente nos indivíduos, Hsiao (1986). Outra vantagem que os modelos com dados em painel permitem é o uso de mais observações, aumentando o número de graus de liberdade e diminuindo a colinearidade entre as variáveis explicativas. Além disso, estes modelos são capazes de identificar e mensurar efeitos que não são possíveis de serem detectados por meio de análise de dados de séries temporais isoladamente. Entretanto, os dados em painel possuem algumas limitações. Conforme Hsiao (1986), como as variáveis são analisadas no tempo, os dados em painel exigem um grande número de observações e, portanto, são mais difíceis de serem implementados.

Em modelos de previsão de insolvência elaborados com dados em painel cada empresa fornece dados contínuos durante períodos (anos) (Hung e Chen, 2009) podendo ser ou não igual para empresas classificadas como solventes ou insolventes. Nestes modelos pela facilidade de acesso a uma maior quantidade de dados obtidos nos demonstrativos contábeis, devido à criação de instância em quantidade bem maior do que o número de empresas, as aplicações e os estudos nestas bases dados acabam por apresentar melhores resultados. Daí a sua preferência e maior utilização do que os modelos elaborados com dados temporais (Balcaen e Ooghe, 2006, Lensberg, T., Eilifsen, A., e McKee, T. E. 2006).

2.5.3 Segmentação por setores econômicos

No estudo sobre previsão de insolvência a segmentação por setores econômicos tende a gerar resultados mais eficientes. Devido às semelhanças econômicas e financeiras desses segmentos, portanto, é necessário escolher setores que permitam a melhor comparação possível dos índices de uma empresa com os de outras, ou seja, setores devem compreender empresas possivelmente semelhantes do ponto de vista de sua composição patrimonial.

Quando da segmentação das empresas por setores econômicos, as classes de empresas são agrupadas de acordo com as suas características econômicas, contábeis, operacionais e patrimonial, o que pode auxiliar numa minoração da probabilidade de ocorrência de problemas na base de dados como, por exemplo, o dos pequenos disjuntos. Empresas de setores econômicos distintos e classificados como solventes para um setor e insolvente para outro setor podem apresentar indicadores contábeis financeiros bem semelhantes ou podem vir a gerar distorções na etapa de processamento de dados.

Para melhorar a avaliação de indicadores contábeis é sempre recomendável que os dados da empresa sejam, posteriormente, comparados ao setor econômico dentro do qual a empresa esteja inserida. É fundamental que a análise ou o estudo de previsão de insolvência seja feita através de amostras de empresas que pertençam ao mesmo setor econômico exatamente para guardar as mesmas características do respectivo setor econômico fazendo com que haja uma "facilitação" na etapa de reconhecer padrões.

Para Iudícibus (2008, p. 91) empresas de mesmo setor econômico apresentam em seus demonstrativos contábeis semelhanças devido a suas estruturas patrimoniais e econômicas. Indicadores como de liquidez, endividamento e rentabilidade, por exemplo, devem apresentar valores bem próximos na sua média setorial. Empresas nas quais apresentam índices bem distintos ao da média setorial no qual pertencem devem apresentar situações com certas anomalias econômicas ou financeiras principalmente, portanto, em condições normais os seus indicadores também devem apresentar comportamentos ajustados. Silva (2006, p. 190) afirma que é possível comparar o índice financeiro de uma empresa com o mesmo índice relativo a outras empresas de mesma atividade econômica, para sabermos como está a empresa em relação as suas principais concorrentes ou mesmo em relação aos padrões do seu segmento de atuação. Para Caouette *et al.*, (1999, p. 129) o analista de crédito compara diversos índices contábeis

do tomador em potencial com normas e tendências setoriais ou grupais pertinentes a estas variáveis.

A classificação setorial das empresas, estudada nesta tese, foi obtida na BOVESPA[4] e foi elaborada considerando-se, principalmente, os tipos e os usos dos produtos ou serviços desenvolvidos pelas empresas, com os seguintes propósitos:

- fornecer uma identificação mais objetiva dos setores de atuação das empresas, já a partir do primeiro nível da estrutura;
- permitir uma visão sobre empresas que, embora com atividades diferentes, atuem em estágios similares da cadeia produtiva ou com produtos/serviços relacionados e tendam a responder de forma semelhante às condições econômicas;
- facilitar a localização dos setores de atuação das empresas negociadas; e
- aproximar-se de critérios utilizados pelo mercado financeiro nacional e internacional.

Para a classificação das empresas, foram analisados os produtos ou serviços que mais contribuem para a formação das receitas das companhias, considerando-se, ainda, as receitas geradas no âmbito de empresas investidas de forma proporcional às participações acionárias detidas.

Os setores econômicos aqui estudados foram: bens industriais, construção e transporte, consumo cíclico, consumo não cíclico, materiais básicos, tecnologia da informação e telecomunicações.

O setor econômico de bens industriais é composto por empresas dos seguintes subsetores: comércio (máquinas e equipamentos, material de transporte), equipamentos elétricos, máquinas e equipamentos, material de transporte e serviços.

No setor de construção e transporte é composto por empresas dos seguintes subsetores: construção, engenharia e transporte.

Já no setor de consumo cíclico as empresas dos seguintes subsetores compõem o setor, automóveis e motocicletas, comércio (eletrodomésticos, livrarias e papelarias, produtos diversos, tecidos, vestuário e calçados), diversos, hotéis e restaurantes, lazer, mídia, tecidos, vestuário e calçados e utilidades domésticas.

[4]BOVESPA.Disponível em:http://www.bovespa.com.br/Home/Redirect.asp?end=/Empresas/ClassifSetorial/FormConsultaClaSetorialOqueE.asp. Acesso em: 19 mar. 2009.

As empresas do setor de consumo não cíclico pertencem aos seguintes subsetor: agropecuária, alimentos processados, bebidas, comércio e distribuição, diversos, fumo, produtos de uso pessoal e de limpeza e saúde.

No setor de materiais básicos é composto por empresas dos seguintes subsetores: embalagens, madeira e papel, materiais diversos, mineração, químicos e siderurgia e metalurgia.

No setor de tecnologia da informação e telecomunicações é composto por empresas dos seguintes subsetores: computadores e equipamentos e programas e serviços telefonia fixa e telefonia móvel.

CAPÍTULO 3. Desenvolvimento de uma estratégia para predição de insolvência

3.1 Considerações iniciais

O problema de determinar insolvência de empresas apresenta dificuldades e/ou propriedades inerentes as características das bases de dados que geralmente as compõe, a saber, i) o banco é geralmente desbalanceado, com um número bem maior de empresas solventes em relação às insolventes; ii) tem-se um maior interesse em obter uma classificação correta nas empresas insolventes (classe minoritária); iii) o entendimento da importância de cada atributo no processo de discriminação é crucial para um maior embasamento das decisões; iv) a base de dados é composta, geralmente, por empresas que pertencem a setores econômicos diferentes, o que pode traduzir em comportamento diferenciado dos atributos para cada setor representado na base de dados.

Além das dificuldades inerentes as características das bases de dados acima citadas, outras questões como a (v) origem dessas bases, se elas se originam dos demonstrativos contábeis ou do mercado e (vi) se as variáveis dessas bases são sequenciais ou de painel, também são propriedades intrínsecas às essas bases e podem influir nos resultados da previsão de insolvência.

Uma base de dados com tais propriedades, dificilmente apresenta um resultado adequado em termos de qualidade de predição quando se utiliza a aplicação direta de um classificador sobre a base, seja uma ANNs, um SVM ou árvore de decisão, mesmo que tal indutor esteja com a parametrização bem ajustada.

Apresenta-se, neste capítulo, uma estratégia para determinação da insolvência de empresas construída visando obter um funcionamento adequado em base de dados com as características descritas acima. O modelo é construído combinando uma técnica de balanceamento de dados por meio de um comitê de classificadores podendo, inclusive, ser utilizado em conjunto com um processo de seleção de características aplicado em uma etapa adequada da construção do comitê de classificadores.

Logicamente, tal estratégia pode ser implementada com diversos tipos de classificadores, métodos de seleção de características e medidas de desempenho. A seguir, descrevem-se alguns dos principais métodos que serão utilizados no modelo como classificadores, técnicas de seleção e medidas de desempenho. São métodos

diferenciados que serão avaliados quando utilizados em conjunto com a estratégia apresentada visando determinar seus desempenhos. Primeiro será apresentado os classificadores, comitê de classificadores, métricas de avaliação e validação e técnicas para bancos desbalanceados. Depois será apresentada uma estratégia para predição de empresas insolvente.

A seguir, mostram-se os principais modelos de seleção de atributos e incorpora a seleção na estratégia para predição de empresas insolventes.

Finalmente, exemplos numéricos são realizados visando validar os modelos desenvolvidos.

3.2 Etapas para construção da base de dados e predição de insolvência

A seguir é apresentada uma arquitetura dos procedimentos realizados na construção da base de dados com o objetivo de obter uma melhor eficácia na caracterização das empresas potencialmente insolventes.

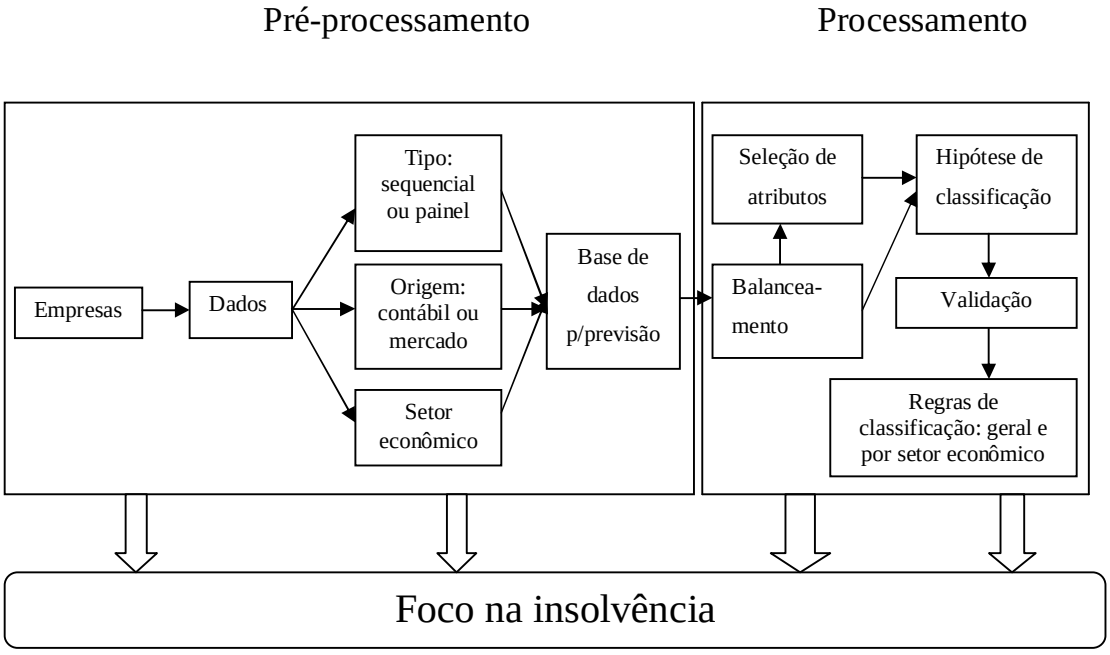

Figura 3. 1 Metodologia geral para previsão de insolvência.

O estudo se inicia na fase de pré-processamento, através de uma base de dados contendo os dados das empresas. Nesta etapa, são definidos o tipo, a origem e o setor econômico dos dados das empresas solventes e insolventes, forma-se assim a base de dados para a previsão que consolida, na maioria das vezes, a etapa de pré-processamento. Na etapa de processamento dos dados será realizada a seleção de atributos (*wrapper*), o balanceamento da base de dados, a hipótese de classificação, as regras para o setor geral e para os setores econômicos específicos e concluindo com a validação dos resultados.

As técnicas utilizadas na metodologia serão detalhadas nas seções posteriores.

3.3 Técnicas de Classificação

Nesta seção são apresentadas sucintamente as técnicas de classificação utilizadas neste estudo: Regressão Logística, Redes Neurais Artificiais, Máquina de Vetor Suporte e Árvore de Decisão. A escolha desses classificadores se deve a dois fatores: por serem metodologias distintas de processamento e também por serem os classificadores mais utilizados neste tipo de estudo.

3.3.1 Regressão logística

O modelo de Regressão Logística (RL) se baseia na função probabilística logística acumulada e é especificado como

$$P_i = F(Z_i) = F(\propto + \beta X_i) = \frac{1}{1+ e^{-Z_i}} = \frac{1}{1+e^{-(\propto+\beta X_i)}} \qquad (3.1)$$

Na equação 3.1, *e* representa a base de logaritmos naturais e P_i é a probabilidade de um indivíduo fazer uma certa escolha, dado X_i. Em relação a distribuição normal a logística tem caudas um pouco mais largas (Pindyck e Rubinfeld, 2004, p. 355).

Para estimar o modelo especificado na equação (3.1) primeiro multiplicamos ambos os lados da equação por $1+ e^{-Z_i}$ para obter

$$(1+ e^{-Z_i})P_i = 1$$

Ao dividir por P_i e subtrair 1, obtemos

$$e^{-Z_i} = \frac{1}{P_i} - 1 = \frac{1-P_i}{P_i}$$

Por definição, contudo, $e^{-Z_i} = 1/e^{Z_i}$, de modo que

$$e^{-Z_i} = \frac{P_i}{1-P_i}$$

Agora, tomando o logaritmo natural de ambos os lados,

$$Z_i = \log \frac{P_i}{1- P_i}$$

ou

$$log \frac{P_i}{1-P_i} = Z_i = \propto + \beta X_i \qquad (3.2)$$

A variável dependente nessa equação de regressão é o logaritmo das chances de que será feita uma escolha particular. Uma vantagem importante do modelo *logit* é que ele transforma o problema de prever a probabilidade dentro de um intervalo (0, 1) no problema de prever a chance de um evento ocorrer dentro do âmbito da linha real.

Se acontecer de P_i ser igual a 0 ou a 1, a probabilidade de P_i / (1-P_i) será igual a 0 ou infinito e o logaritmo da probabilidade será indefinido. Assim, a aplicação da estimação por mínimos quadrados ordinários à equação (3.2) é claramente inadequada. A estimação correta do modelo *logit* pode ser mais bem entendida com a distinção entre estudos em que as unidades básicas de análise são observadas individuais e estudos e que a análise envolve o uso de dados agrupados (Pindyck e Rubinfeld, 2004, p. 356).

3.3.2 Redes Neurais Artificiais

Para Haykin (2001, p. 28) uma Rede Neural Artificial (ANN) é um processador maciçamente paralelo, distribuído, constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.
2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Tipicamente, uma ANN consiste de um conjunto de unidades sensoriais (nós de fonte) que constituem a camada de entrada, uma ou mais camadas ocultas de nós computacionais e uma camada de saída de nós computacionais. O sinal de entrada se propaga para frente através da rede, camada por camada. Estas redes são normalmente chamadas de *perceptrons* de múltiplas camadas (MLP), as quais representam uma

generalização do *perceptron* de camada única (constituído em torno de um neurônio não-linear, isto é, o modelo de *McCulloch-Pitts* de um neurônio).

O MLP é aplicado através do seu treinamento de forma supervisionada com um algoritmo conhecido como *algoritmo de retropropagação de erro* (*error back-propagation*). Este algoritmo é baseado na regra de aprendizagem por correção de erro.

Basicamente, a aprendizagem por *retropropagação* de erro consiste de dois passos através das diferentes camadas da rede: um passo para frente, a propagação, e um passo para trás, a *retropropagação*. No passo para frente, um padrão de atividade (vetor de entrada) é aplicado aos nós sensoriais da rede e seu efeito se propaga através da rede, camada por camada. Finalmente, um conjunto de saídas é produzido como a resposta real da rede. Durante o passo de propagação, os pesos sinápticos da rede são todos fixos. Durante o passo para trás, por outro lado, os pesos sinápticos são todos ajustados de acordo com uma regra de correção de erro. Especialmente, a resposta real da rede é subtraída de uma resposta desejada (alvo) para produzir um sinal de erro. Este sinal de erro é então propagado para trás através da rede, contra a direção das conexões sinápticas – vindo daí o nome de "retropropagação de erro". Os pesos sinápticos são ajustados para fazer com que a resposta real da rede se mova para mais perto da resposta desejada, em um sentido estatístico. O algoritmo de retropropagação de erro é também referido na literatura como algoritmo de retropropagação (*back-propagation*) O processo de aprendizagem realizado com o algoritmo é chamado de *aprendizagem por retropropagação*, segundo Haykin (2001, p.184).

Para Kumar e Ravi (2007) a grande vantagem das redes neurais é que elas são boas funções de aproximação, previsão, classificação, clusterização e otimização de tarefas dependendo da arquitetura da rede neural. Os mesmos autores citam como desvantagem que a determinação de vários parâmetros associados com o algoritmo de treinamento não é tarefa simples. As arquiteturas das redes neurais precisam de um lote de dados para treinamento e ciclos de treinamento (iterações).

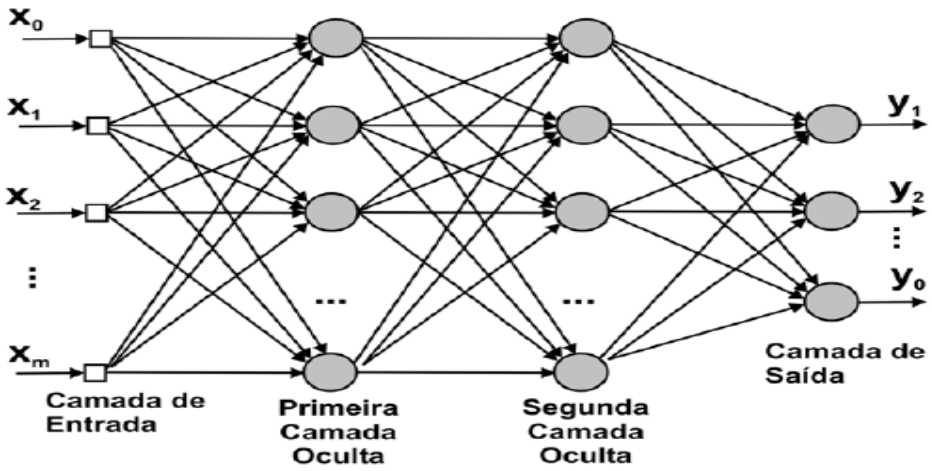

Figura 3. 2- Grafo arquitetural de um *perceptron* de múltiplas camadas com duas camadas ocultas (Haykin, 2001, p.186).

3.3.3 Máquina de Vetor Suporte

A Máquina de Vetor Suporte (SVM) é um algoritmo de aprendizagem poderoso baseado em recentes avanços na teoria de aprendizagem estatística (Vapnik, 1998). É um método de classificação para dados lineares e não lineares e podem ser usadas para classificação de padrões e regressão linear. No caso de padrões separáveis que podem surgir no contexto de classificação de padrões a idéia principal de uma máquina de vetor suporte é construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima. A máquina de vetor suporte é uma implementação do *método de minimização estrutural de risco*. Este principio indutivo é baseado no fato de que a taxa de erro de uma máquina de aprendizagem sobre dados de teste (isto é, a taxa de erro de generalização) é limitada pela soma da taxa de erro de treinamento e por um termo que depende da dimensão de Vapnik-Chervonenkis (V-C). No caso de padrões separáveis, uma máquina de vetor suporte produz um valor de zero para o primeiro termo e minimiza o segundo termo (Haykin, 2001, p. 349). Consequentemente, SVM tem se tornado uma das técnicas mais populares de ferramentas de aprendizado de máquina podendo ser aplicado tanto em problemas de predição quanto de regressão, apesar do fato de que ela não incorpora conhecimento do domínio do problema.

Uma noção que é central à construção do algoritmo e aprendizagem por vetor suporte é o núcleo do produto interno entre um "vetor suporte" x_i e o vetor x retirado do

espaço de entrada. Os vetores de suporte consistem de um pequeno subconjunto dos dados de treinamento extraído pelo algoritmo. Dependendo de como este núcleo de produto interno é gerado, pode-se construir diferentes máquina de aprendizagem, caracterizadas por superfícies de decisão não-lineares, próprias. Pode-se usar o algoritmo de aprendizagem por vetor suporte para construir os três seguintes tipos de máquinas de aprendizagem: (i) máquinas de aprendizagem polinomial; (ii) redes de função de base radial e (iii) *perceptrons* de duas camadas.

As principais características de SVM são (Tan *et al.*, 2006, p. 276):

1. O problema de aprendizagem do SVM pode ser formulado como um problema de otimização convexa, o qual algoritmos eficientes estão disponíveis para encontrar o mínimo global da função objetiva.
2. SVM executa controle de capacidade maximizando a margem do limite de decisão.
3. SVM pode ser aplicado a dados categóricos introduzindo variáveis simuladas para cada valor de atributo categorizado nos dados.

Na realidade, SVM podem servir como uma alternativa que combina as vantagens dos métodos estatísticos convencionais que são mais "teoricos-dirigidas" com a facilidade de serem analisador e métodos de aprendizado de máquina que são métodos mais orientados para os dados, livre de distribuição e robusto (Ravi *et al.*, 2008).

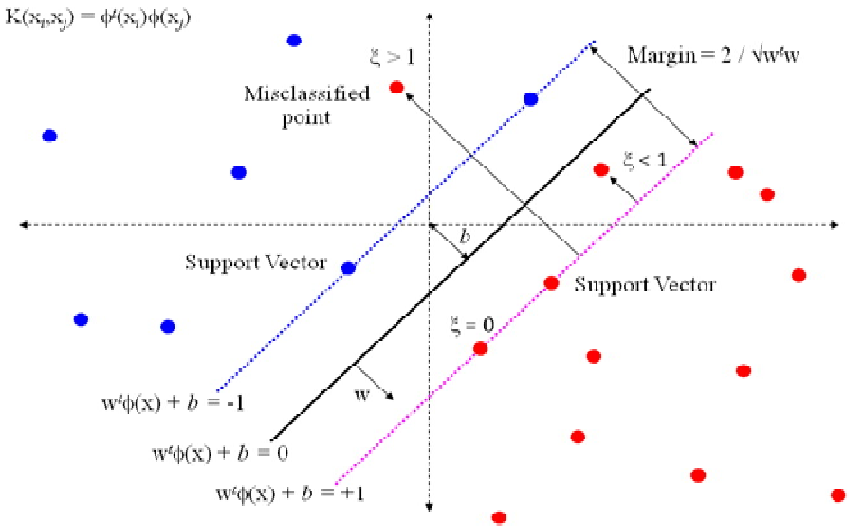

Figura 3. 3- Ilustração da idéia de um hiperplano ótimo para padrões linearmente separáveis

3.3.4 Árvore de decisão

A Árvore de Decisão (AD) é uma técnica de classificação que usa o recurso da técnica do particionamento e mede o grau de entropia para induzir na decisão em grupo de dados. Para Ravi e Kumar (2007) a vantagem dessa técnica é que ela resolve problemas tanto de classificação como de regressão. Utiliza a técnica de fácil compreensão humana, a regra binária "se então". As desvantagens citadas pelos autores são de que (i) *overfitting* pode gerar problemas na classificação e, (ii) como redes neurais, elas também exigem uma lote de amostras de dados, a fim de obter previsões confiáveis. Algumas das grandes vantagens na utilização de AD são a geração e a explicitação das regras de classificação facilitando um melhor entendimento do processo de discriminação diferente dos métodos de classificação como ANNs e SVM. Na subseção a seguir se discutirá a elaboração das regras de classificação.

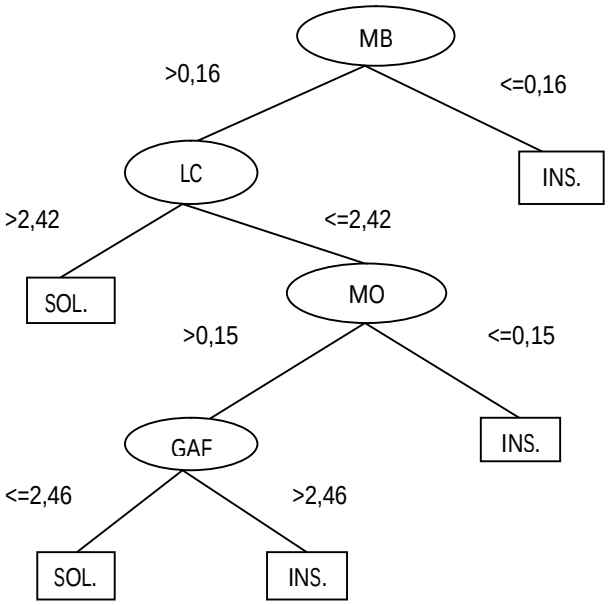

Figura 3. 4- Exemplo de uma árvore de decisão para definir se uma empresa é solvente ou insolvente utilizando variáveis contábeis.

3.3.4.1 Regras de classificação

As regras de classificação assumem a forma se *L* então *R* ou na forma simbólica $L \rightarrow R$ que também pode ser representada pelas formas equivalentes $R \leftarrow L$ ou *R*: - *L*. Normalmente, as partes esquerdas L e direita R da regra são complexas, sem atributos comuns entre eles, ou seja, atributo (L) ∩ atributo (R) = Ø. A parte esquerda L é denominada condição, premissa, cauda ou corpo da regra, e a parte direita R é denominada conclusão ou cabeça da regra. Uma regra de classificação assume a forma restrita de uma regra se L então classe = C_i, onde C_i pertence ao conjunto de K valores de classe $\{C_1, C_2,..., C_k\}$. Uma regra de classificação pode também ser definida na identificação de um conjunto de registros, onde os valores dos atributos – A = $\{a_1, a_2, a_3,..., a_n\}$ levam a uma classe B, ou seja, verifica-se a ocorrência de $A \rightarrow B$. Um modelo de aprendizagem baseado em regras é composto por expressões no seguinte formato:

SE [*condição*] ENTÃO [*conclusão*].

Regras de classificação visam descobrir um pequeno conjunto de regras na base de dados através de um classificador (U. Fayyad, et al., 1996).

O objetivo do usuário em utilizar um sistema de descoberta de conhecimentos (KDD) é encontrar regras que sejam compreensíveis [e.g., U. Fayyad, et al., 1996, B. Liu e W. Hsu 1996, B. Liu, W. Hsu e S. Chen, 1997, G. Piatesky-Shapiro e C. Matheus, 1994, G. Piatetsky-Shapiro, C. Matheus, P. Smyth, e R. Uthurusamy. 1994, A. Silberschatz e A. Tuzhilin. Regras são fáceis de entender quando elas estão em conformidade e em consistência com o conhecimento do especialista ou do seu usuário (M. Pazzani e D.Kibler. 1992). Regras são interessantes se elas são aplicáveis (ou facilitadoras de entendimentos) (G. Piatesky-Shapiro e C. Matheus, 1994) e/ou imprevistas (A. Silberschatz e A. Tuzhilin, 1996 B. Liu e W. Hsu, 1996, B. Liu, W. Hsu e S. Chen 1997). No entanto, para uma KDD saber quais as regras são interessantes e compreensíveis para um usuário não é uma tarefa fácil. Uma regra pode ser interessante para um usuário, mas não é interessante para o outro. Assim, se uma regra é interessante e compreensível ou não é subjetiva. Depende do conhecimento prévio do usuário sobre o domínio e seu atual interesse (Shu Chen e Bing Liu, 2001).

Reconhecer a aplicação de regras de classificação como ferramentas de aprendizagem em aplicações na vida real não é de forma alguma uma tarefa fácil (M. Pazzani e D.Kibler, 1992) porque essas ferramentas muitas vezes produzem regras que são completamente irrelevantes para o usuário dos conceitos existentes sobre o domínio e os interesses do usuário. Estas regras podem ser muito difíceis de compreender e/ou desinteressantes para o usuário. A raiz do problema é produzir um pequeno conjunto de regras precisas para formar um modelo de domínio (Quinlan, 1992). Regra de indução usa vários sistemas com vieses no processo de geração de regras. Estes vieses, porém, podem não estar de acordo com o conhecimento existente do usuário humano, resultando em regras incompreensíveis e desinteressantes em relação aos problemas. Ou seja, muitas das regras geradas podem não fazer sentido para o usuário e/ou não são interessantes para o usuário.

Árvore de decisão é uma das mais populares técnicas de extração de regras e apresenta boa capacidade de generalização. A fim de gerar as regras, cada caminho está traçado na decisão árvore, a partir do nó raiz ao nó folha, grava o resultado como os antecedentes e as folhas de nó classificação como as conseqüências. Árvores de decisão são fáceis de serem construídas automaticamente.

Uma das técnicas mais populares de extração de regras é baseada no algoritmo C4.5 (Quilan, 1993). Este algoritmo de árvores de decisão é baseado em conceitos teóricos de informações. A entropia é utilizada para medir a informação de um atributo em um conjunto de dados (Martens, *et al.*, 2007).

A entropia de uma amostra S é calculada da seguinte forma:

Entropia (S) = $-p_1 \log_2(p_1) - p_0 \log_2(p_0)$,

Onde p_1 (p_0) é a proporção de exemplos da classe 1 (0) na amostra S, e $p_0 = 1 - p_1$.

Basicamente, a entropia mede a ordem (ou desordem) nos dados no que diz respeito às classes. É igual 1 quando $p_1 = p_0 = 0{,}5$ (distúrbio máximo, mínima ordem) e 0 (ordem máxima, mínima desordem) quando $p_1 = 0$ ou $p_0 = 0$. Neste último caso, todas as observações pertencem à mesma classe. A medida de ganho conhecida como Gain (S, x_j) é definido como a expectativa de redução de entropia na separação do atributo x_j:

Gain (S, x_j) = Entropia (s) - $\sum_{v \in valor(x_j)} \frac{|S_v|}{|S|} \, Entropia\ (S_v)$,

Onde o valor (x_j) representa o grupo de todas as possibilidades dos atributos x_j, S_v a subamostra de S onde atributo x_j tem valor v e $|S_v|$ o número de observações em S_v. O Gain foi o critério usado pelo ID3, o precursor do C4.5, para decidir qual o atributo que deve dividir determinado nó (Quinlan, 1986). Entretanto, quando este critério é usado para decidir para dividir o nó, o algoritmo favorece uma divisão de atributos que há muitos valores distintos. Visando retificar isto, o C4.5 aplica uma normalização e usa o critério de razão (*Gainratio*) de ganhos definido como:

$$Gainration\ (S, x_j) = \frac{Gain(S,x_j)}{SplitInformation\ (S,x_j)} \text{ com}$$

$$SplitInformation\ (S, x_j) = -\sum_{k \in valor(x_j)} \frac{|S_k|}{|S|} \log_2 \frac{|S_k|}{|S|}.$$

3.4 Comitê de Classificadores (*Ensemble*)

São técnicas de melhora de classificação através da agregação de múltiplos classificadores. Estas técnicas são conhecidas como comitê de classificadores ou combinação de métodos de classificação. O método de comitê de classificadores constrói um grupo com base em classificadores para treinamento dos dados e faz a classificação votando nas predições feitas por cada classificador (Pang-Ning Tan *et al.*, 2006, p.276).

Para Tan *et al.,* (2006, p.277) os métodos de comitê de classificadores tendem a apresentar uma performance melhor do que os classificadores singulares. Duas são as condições necessárias para o comitê de classificadores apresentarem um desempenho melhor do que o classificador singular: (1) Os classificadores devem ser independentes um do outro, e (2) a base de classificadores deve ser melhor do que qualquer classificador singular. Na prática, é difícil de assegurar independência total entre os classificadores básicos, entretanto, foram observadas melhorias em precisões de classificação em métodos que usam comitês nos quais os classificadores básicos são ligeiramente correlacionados.

Pode ser dito também que um comitê de classificadores é um conjunto de classificadores cujas decisões individuais são combinadas de alguma forma para classificar um conjunto de dados cuja classe seja desconhecida. Uma condição necessária e suficiente para um conjunto de classificadores ser mais preciso do que algum de seus membros individuais é se os classificadores forem exatos e distintos

(Gary *et al.*, 2007). Dois classificadores são distintos quando cometem erros diferentes em novos conjuntos de registros.

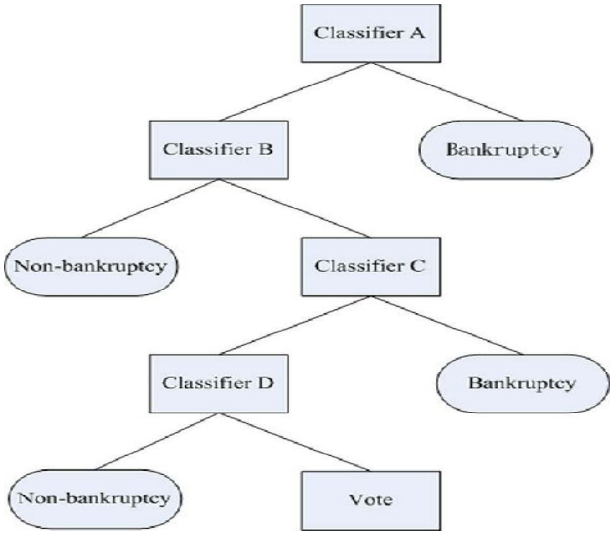

Figura 3. 5- Comitê de classificadores (Hung, *et al.*, 2009 p. 5301).

3.4.1 Métodos para construção dos comitês de classificadores

A idéia básica para construção de múltiplos classificadores é gerar vários bancos de dados do original e nestes novos bancos aplicarem classificadores para em seguida combinar os resultados destes classificadores.

Para Tan, Steinbach e Kumar (2006) os comitês de classificadores podem ser construídos de algumas maneiras:

1. Pela manipulação dos bancos de treinamento. Nesta técnica, múltiplas amostras de treinamento são criadas aplicando alguma distribuição. Esta distribuição determina como que estes novos bancos serão selecionados para o treinamento. A classificação é então feita por cada grupo de treinamento usando um classificador. *Bagging* e *boosting* (que serão apresentados em parágrafos seguintes) são dois exemplos de amostragem utilizados pelos métodos de comitê de classificadores.

2. Pela manipulação de acréscimos de atributos. Nesta técnica, o subgrupo de incrementos de atributos é feita para cada grupo de treinamento. O subgrupo pode ser feito aleatoriamente ou através da recomendação de um especialista.
3. Pela manipulação de classes específicas. Este método pode ser usado quando o número de classes é suficientemente grande. O treinamento de dados é transformado em problemas binários com particionamento aleatório de classes designadas em dois subgrupos disjuntos, A_0 e A_1.
4. Pela manipulação do algoritmo de treinamento. Alguns algoritmos podem ser ajustados no treinamento dos dados podendo resultar em diferentes modelos. Por exemplo, as ANN podem ter diferentes modelos trocando a sua topologia ou as entradas iniciais dos links entre os neurônios, similarmente, as árvores de decisão podem ser construídas variando os nós da árvore.

Dentre as diversas técnicas de construção dos *comitês de classificadores* que têm sido desenvolvidas, os modelos de *bagging e de boosting* se destacam.

A técnica de *Bagging* apresenta três etapas: a primeira etapa é a construção de L conjuntos de registros através de replicações *bootstrap* de um conjunto de registros inicial (Atiya Amir, 2001).

Bagging, é também conhecido como agregando *boostrap*, é uma técnica que repete amostras (com reposição) com um grupo de dados conforme uma distribuição probabilística uniforme. Cada amostra *boostrap* tem o mesmo tamanho do banco original. Devido à nova amostra ter sido elaborada com reposição, algumas instâncias podem assemelhar com a amostra inicial em alguns grupos de treinamento, como outros podem ser omitidos no novo grupo de treinamento. Na média, a amostra *boostrap* D_i contém aproximadamente 63% da base de dados original porque cada amostra tem a probabilidade $1 – (1 – 1/N)^N$ de cada amostra D_i selecionada. Se N é suficientemente grande, esta probabilidade converge para $1 – 1/e \approx 0{,}632$ (Pang-Ning Tan et al., 2006, p. 284)

O método *boosting* é um procedimento iterativo usado para alterar adaptativamente a distribuição de exemplos de treinamento de modo que os classificadores de base enfoquem exemplos que sejam difíceis de classificar. Este método atribui um peso a cada exemplo de treinamento e podem ser usados das seguintes maneiras:

1. Eles podem ser usados como uma distribuição da amostra para desenhar um conjunto de amostras de *bootstrap* a partir dos dados originais.
2. Eles podem ser usados pelo classificador de base para descobrir um modelo que tenha tendência na direção de exemplos de peso mais altos.

3.4.2 O método de combinação da votação majoritária

A combinação de vários classificadores ou de várias bases de dados de mesma origem é uma técnica de decisão que combina diferentes bases (ou classificadores) para gerar uma decisão conjunta. Todas as bases são preparadas para resolver o mesmo problema, que nesta tese é prever empresas insolventes. Existem várias combinações de bases (classificadores) individuais, entre as quais a maioria dos votos é o método mais utilizado (Lam e Suen, 1997; Oh, 2003; Xu, Krzyzak, e Suen, 1992). A técnica da maioria dos votos é um método simples e eficaz de combinação. Ele escolhe o rótulo de classe que é apoiada pela maioria dos múltiplos classificadores (Li e Sun, 2009). Existem três versões de maioria dos votos ou votação majoritária (Polikar, 2006), onde o comitê de classificadores escolhe a classe: (i) há unanimidade dos resultados entre os classificadores; (ii) prevê pelo menos mais da metade o número de classificadores (maioria simples); (iii) recebe a maior número de votos, se deve ou não a soma dos votos ultrapassarem 50% (pluralidade dos votos ou apenas uma votação por maioria).

Nesta tese a versão aplicada do método de combinação da maioria dos votos é o da maioria simples.

3.5 Métricas de avaliação

As métricas para avaliação desempenham um papel crucial para avaliar o desempenho do modelo e também para comparar o desempenho relativo de diferentes classificadores no mesmo domínio. Tradicionalmente, a precisão é a mais comumente utilizada para medir esses efeitos. No entanto, para classificação de classes desbalanceadas a precisão não é mais uma métrica interessante porque a classe minoritária tem muito pouco impacto sobre a precisão em comparação com a da classe predominante (M. V. Joshi *et al.,* 2001 e H. Käuck, 2004). Das métricas alternativas existentes para lidar com o problema do desequilíbrio de classes citadas por Sun (2007)

foram escolhidas: matriz de confusão, área sob a curva ROC e medida F. Já para a avaliação do classificador serão utilizadas validação cruzada e resubstituição.

3.5.1 Matriz de confusão

Os diferentes tipos de erros e acertos realizados por um classificador podem ser sintetizados em uma matriz de confusão. Na Tabela 3.1 é mostrada uma matriz de confusão para um problema que possui duas classes rotuladas como positiva e classe negativa.

	Predição Positiva A	**Predição Negativa B**
Classe Positiva	Verdadeiro Positivo -TP	Falso Negativo – FN
Classe Negativa	Falso Positivo – FP	Verdadeiro Negativo - TN

Tabela 3. 1 Matriz de confusão para duas classes de problemas

Para comparar os resultados dos classificadores encontrados nos estudos referentes ao tema presente neste estudo são utilizados os erros Tipo I e Tipo II representados na matriz de confusão da Tabela 3.2.

- Sensibilidade ou recall: mede a capacidade de medir os exemplos positivos

  $Sens = TP/(TP + FN)$

- Especificidade: mede a capacidade de medir os exemplos negativos

  $Spec = TN/(FP + TN)$

A matriz de confusão			
		Predição	
		Solventes	Insolventes
Real	Insolventes	Erro Tipo II	
	Solventes		Erro Tipo I

Tabela 3. 2 Erros na matriz de confusão para duas classes de problemas

- Erro Tipo I – representam as instâncias classificadas incorretamente como insolventes, ou seja, são as taxas de erros daquelas instâncias (empresas) que

apresentam uma saúde financeira boa e são classificadas como instâncias em que a sua saúde financeira é ruim.

- Erro Tipo II – representam as instâncias classificadas incorretamente como solventes. Contrapondo-se ao erro Tipo I, esta representa as taxas de erros daquelas instâncias que apresentam uma saúde financeira ruim e são classificadas como instâncias em que a sua saúde financeira é boa.

3.5.2 Área ROC

Por definição, uma curva ROC é um gráfico bidimensional em que o eixo horizontal representa a taxa de erro da classe negativa (1-*Spec*) e no eixo vertical os valores de sensitividade. O desempenho de um classificador é medido pela área sob a curva ROC (AUC).

Pontos na diagonal representam classificadores aleatórios, de acordo com a probabilidade a priori de cada classe. Acima (abaixo) da diagonal encontram-se classificadores com desempenho melhor (pior) que o classificador aleatório.

3.5.3 Medida F

Para se entender a métrica Medida F é importante conhecer a definição de sua composição. *Recall* e *Precision* ou precisão são duas métricas amplamente usadas em aplicações onde a detecção bem sucedida de uma classe é considerada mais significativa do que outras classes. Uma definição formal destas métricas é apresentada a seguir:

Precisão, $p = \frac{TP}{TP+FP}$

Recall, $r = \frac{TP}{TP+FN}$

A precisão determina a fração de registros que realmente acabam sendo positivos no grupo que o classificador declarou como classe positiva. Quanto maior a precisão, menor o número de erros positivos no grupo que o classificador declarou como classe positiva. O *recall* mede a fração de exemplos positivos previstos corretamente pelo classificador. Classificadores com valores altos de *recall* têm poucos exemplos positivos mal classificados como a classe negativa. Na verdade, o valor de *recall* é equivalente à taxa de positivos verdadeiros.

O desafio dos algoritmos de classificação é elaborar um modelo que maximize tanto a precisão como o *recall*.

É a métrica entre precisão e *recall* dada pela fórmula abaixo:

$$Medida\ F = \frac{2\ x\ recall\ x\ precisão}{recalll + precisão} = \frac{2xTP}{2xTP + FP + FN}$$

A princípio, F representa uma média harmônica entre *recall* e a precisão, i.e., $F = \frac{2}{\frac{1}{r}+\frac{1}{p}}$.

3.5.4 Validação cruzada

A validação cruzada é uma técnica de reamostragem que permite que todos os dados da base de dados sejam utilizados para treinamento e teste. Na validação cruzada os dados iniciais são particionados aleatoriamente em K subconjuntos D_1, D_2,...,D_k aproximadamente iguais em tamanho. Em seguida é estimado um modelo utilizando K-1 subconjuntos e é testado o subconjunto restante. O procedimento anterior é repetido K vezes, utilizando sempre um subconjunto diferente para teste (Han e Kamber, 2006).

3.5.5 Validação por resubstituição

Na validação por resubstituição ocorre a construção da hipótese de classificação com todos os dados para em seguida aplicar esta mesma hipótese de classificação por sua vez em cada uma das observações.

Braga- Neto *et al.,* (2004) afirma que para a maioria das classificações e regras de classificação, resubstituição subestima a taxa de erro de classificação, e em alguns casos esta tendência pode ser grave, o mais extremo caso pode ser da classificação de um único vizinho mais próximo, onde a estimativa da resubstituição é sempre zero. Por outro lado, validação cruzada está próxima do imparcial no seguinte sentido: se o procedimento é repetido para diferentes amostras extraídas em uma população, então a estimativa do erro médio aproxima o erro esperado dos classificadores projetados em todos os possíveis nas amostras do mesmo tamanho. Devido a considerações de viés, validação cruzada é geralmente preferida sobre resubstituição. No entanto, para

pequenas amostras, a variância do estimador precisa ser considerada, e aqui resubstituição é superior a validação cruzada (Devroye *et al.*, 1996).

3.6 Técnicas de tratamento de bancos desbalanceados

Uma maneira de solucionar o problema de classes desbalanceadas numa base de dados é balancear artificialmente a distribuição das classes no conjunto de exemplos. Dentre os diversos métodos propostos existentes para o tratamento de bancos desbalanceados, duas abordagens principais são utilizadas nesta tese, são elas:

- Remoção de exemplos da classe majoritária - u*nder-sampling*. Os métodos *under-sampling*, visam balancear o conjunto de dados por meio da eliminação de sub-amostras da classe majoritária;
- Inclusão de exemplos da classe minoritária - *over-sampling*. Os métodos de *over-sampling* visam balancear a distribuição das classes por meio da replicação de exemplos da classe minoritária;

Over-sampling aleatório é um método heurístico que replica aleatoriamente exemplos da classe minoritária. Em sua implementação, essa replicação é feita sem reposição.

Under-sampling também é um método não heurístico que elimina aleatoriamente exemplos da classe majoritária.

Ambos os métodos por serem aleatórios possuem limitações conhecidas. *Under-sampling* pode eliminar dados potencialmente úteis, e *over-sampling* pode aumentar a probabilidade de ocorrer superajustamento aos dados, uma vez que a maioria dos métodos de *over-sampling* faz cópias exatas dos exemplos pertencentes à classe minoritária.

Alguns trabalhos recentes têm tentado superar as limitações existentes tanto nos métodos de *under-sampling*, quanto aos métodos de *over-sampling*. Por exemplo, Chawla *et al.*, (SMOTE) (2002) combinam métodos de *under* e *over-sampling*. Nesse trabalho, o método de *over-sampling* não replica os exemplos da classe minoritária, mas cria novos exemplos dessa classe por meio da interpolação de diversos exemplos da classe minoritária que se encontram próximos. Dessa forma, é possível evitar o problema do superajustamento.

O algoritmo SMOTE, que é uma técnica bastante empregada, servirá como comparativo com os algoritmos aqui propostos.

3.7 Considerações

Em relação às técnicas apresentadas na literatura especificamente para a predição de insolvência observa-se que basicamente, não se tem a preocupação com o desbalanceamento entre as amostras das empresas insolventes e solventes. Geralmente os autores optam por trabalhar com um banco balanceado desrespeitando a relação natural entre empresas insolventes e solventes, o que torna a base de dados pouco representativa. Feita esta opção, pelo uso de um banco balanceado, torna-se comum e aceitável o uso do nível de acerto na predição com medida de desempenho do classificador. Esta é a estratégia mais utilizada em trabalhos que tratam do problema de insolvência de empresas.

Em relação ao setor econômico, as técnicas de predição de insolvência conhecidas, geralmente utilizam o setor econômico como um atributo adicional, na expectativa que este atributo seja relevante no processo de discriminação.

Quanto a procedimentos de seleção de características, são aplicados de forma usual, tanto para modelos em filtro como em capsula (*wrappper*), como se faz para qualquer base de dados.

Deve-se destacar que pelas propriedades especificas de uma base de dados de empresas solventes e insolventes, tem-se a expectativa que uma ferramenta que leve em conta estas propriedades obtenha um nível de predição mais eficiente. Na seção a seguir, descrevem-se os passos e as motivações utilizadas para o desenvolvimento de uma ferramenta para a predição de insolvência em empresas de capital aberto. Logicamente, os classificadores, as estratégias para seleção de características e as medidas de desempenho descritas anteriormente serão componentes da estratégia apresentada e devem ser avaliadas para identificar as que mais se adaptam ao modelo apresentado.

3.8 Uma estratégia para a predição de empresas insolventes

Descreve-se, nesta seção, um método construído especificamente para a predição de insolvência em uma base de dados desbalanceada composta por empresas de diversos setores econômicos. Considera-se como sendo uma ferramenta especifica por levar em conta, na sua construção, todas as propriedades descritas na seção 3.1, que geralmente caracterizam uma base de dados de empresas solventes/insolventes.

Deve-se destacar que se considera o maior diferencial da base de dados o desbalanceamento existente entre as empresas consideradas solventes e as classificadas como insolventes. Desta forma, o principal enfoque da metodologia esta em corrigir este desbalanceamento de maneira a proporcionar uma melhor qualidade na predição de novas empresas. Outras considerações como, por exemplo, um maior interesse na predição de empresas insolventes, utilização do setor econômico, entre outros são levados em conta visando complementar a construção do modelo.

Conforme descrito na seção 3.6, um dos principais modelos para trato de base de dados desbalanceados baseia-se em procedimentos randômicos de diminuição dos dados da classe majoritária (*under-sampling*), incremento dos dados da classe minoritária por meio da replicação randômica com reposição (*over- sampling*), e na combinação das duas estratégias. Neste caso, não se tem a geração de novas instâncias, simplesmente o balanceamento é feito com a manipulação da base de dados original.

Outro modelo de destaque para o balanceamento tem como estratégia a inserção de novas instâncias geradas artificialmente na classe minoritária (SMOTE). Este método tem apresentado bons resultados de uma forma geral. A maior dificuldade é a falta de garantia que se tem das instâncias sintéticas pertencerem realmente a classe a que foram associadas.

Deve-se destacar que estas estratégias baseiam-se em um processo totalmente estocástico para a obtenção de bases balanceadas. O modelo desenvolvido busca diminuir este componente estocástico visando: i) a utilização dos dados da classe minoritária de forma mais intensa ou redundante, pois, busca-se um maior nível de acerto nesta classe; ii) a decomposição da classe majoritária de forma a torná-la de dimensão mais próxima a classe minoritária e permitir a criação de bases com composições diferentes de setores econômicos representados na base de dados original.

É importante ressaltar que a obediência a estes dois objetivos traz como característica adicional a diminuição da aleatoriedade na obtenção do balanceamento.

Dai, denomina-se tal modelo de *Semi-Determinístic Ensemble Strategy for Imbalanced Data* (SEID).

A forma definida para se levar em conta estes dois objetivos conjuntamente foi por meio de um comitê de classificadores. Um procedimento de comitê apresenta, naturalmente, uma facilidade de implementação dos objetivos para cada classe descrito acima. No caso da necessidade de redundância das instâncias minoritárias tem-se a facilidade de utilização de todas suas instâncias em cada base do comitê. No caso das instâncias majoritárias, onde se pretende particionar ou decompor seus elementos, podem-se colocar parcelas de suas instâncias em bases diferentes para gerar os classificadores que formam o comitê. Desta forma, a partição não prejudica nem a representatividade dos dados da classe majoritária, que devem compor pelo menos uma base de dados do comitê, nem a dimensão do banco, pois uma estratégia de comitê lida bem com bancos menos completos por não basear a decisão em somente um dos classificadores gerados. Além disto, os parâmetros para determinar tamanhos mínimos da base dos classificadores do comitê servem para evitar a utilização de bases de dimensão consideradas inadequadas.

Ressalta-se que esta estratégia baseada em comitê para o balanceamento permite o uso de um procedimento de seleção de características de uma forma diferenciada, que será descrita mais adiante. A seguir, apresenta-se o método para predição de insolvência em empresas:

Considera-se inicialmente a composição do conjunto de treinamento $Str = Str_m \cup Str_M$, ou seja, formado pela união de instâncias da classe minoritária (Str_m) e da clase majoritária (Str_M) com #(Str_M) > #(Str_m), onde #(*) é a cardinalidade do conjunto. Os conjuntos de treinamento gerados para a obtenção dos classificadores base serão balanceados com n_{ic} instâncias de cada classe. Para que se obtenham conjuntos de treinamentos de acordo com o modelo proposto, adota-se para o valor mínimo de instância por classe n_{ic} :

$$n_{ic} \geq \max\left(\#(Str_m), \#(Str_M)/n_{bc}\right) \tag{3.3}$$

Com n_{cb} sendo o número de classificadores base usados no comitê de classificadores e o operador max (*) assume o maior valor entre os avaliados. Quanto maior o valor de n_{ic} mais próximo o algoritmo se torna do algoritmo de *bagging*. A seguir, apresenta-se o pseudo-código do SEID.

Pseudo-código: *comitê de classificadores para base de dados desbalanceadas* (SEID)

início

Defina o número de classificadores base n_{cb}

Defina o número de instâncias para cada classe n_{ic}

% construção dos n_{cb} *classificadores base*

para i=1, n_{cb}

% classe minoritária

$Str_i \leftarrow Str_m$

% completar, quando necessário, aplicando um processo de bootstrap na classe minoritária

para $j = \#(Str_m) + 1, n_{ic}$

$Str_i \leftarrow Str_i \cup$ *j-ésima instância obtida aplicando bootstrap na amostra* Str_m

fim

%classe majoritária

para $j = 1, \ \#(Str_M)/n_{cb}$

$Str_i \leftarrow Str_i \cup$ *j-ésima instância obtida de* Str_M *sem reposição*

fim

%completar, quando necessário, aplicando um processo de bootstrap na classe majoritária

para $j = \#(Str_M)/n_{cb} + 1, n_{ic}$

$Str_i \leftarrow Str_i \cup$ *j-ésima instância obtida aplicando bootstrap na amostra* Str_M

fim

fim

Treine os n_{cb} *classificadores base*

%classificação de novas instâncias

Aplique técnica de votação majoritária para classificar os dados de teste

fim.

3.9 Seleção de atributos

A seleção de atributos representa um problema de fundamental importância em mineração de dados (DM), sendo freqüentemente realizada como uma etapa de pré-processamento. Os objetivos principais da seleção de atributos em DM para previsão de insolvência são: (i) o desenvolvimento de modelos compactos, (ii) o uso e refinamento do modelo de classificação ou predição para avaliação e (iii) a identificação de índices financeiros relevantes (Piramuthu, 2006).

Os algoritmos usados para seleção de atributo podem ser separados em duas atividades principais: busca do subconjunto de atributos e avaliação dos subconjuntos de atributos encontrados, como pode ser visto na Figura 3.1 (Liu, e Motoda, 1998).

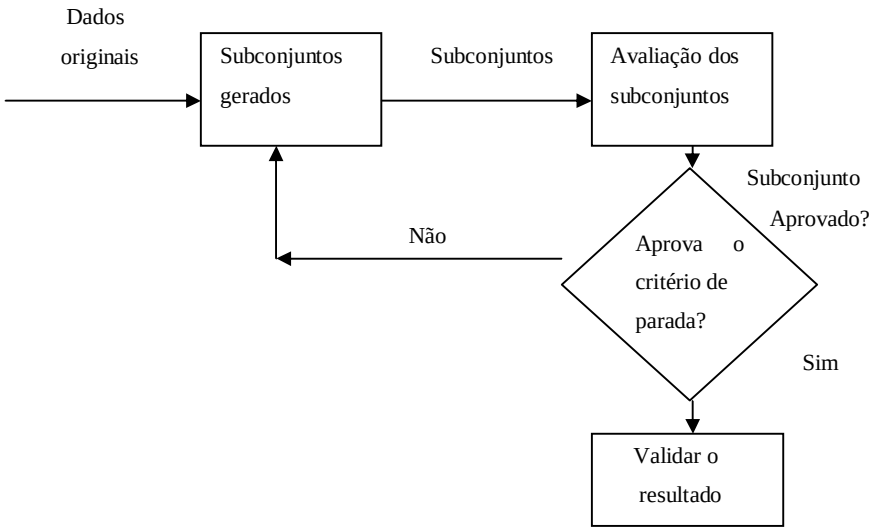

Figura 3. 6 Passos na Seleção de Atributos.

Na atividade de busca de subconjuntos de atributos seleciona-se um subconjunto de variáveis relevantes com o apoio de um algoritmo de busca. Neste trabalho foram utilizadas duas abordagens de direção de busca ou ponto de partida: (i) seleção *forward* e (ii) seleção aleatória. Na seleção *forward* vai-se adicionando um atributo por vez ao subconjunto até que não se consiga melhorar a qualidade do subconjunto de atributos. Esta abordagem foi utilizada no algoritmo de busca *GreedyStepwise*. Já abordagem usando seleção aleatória como ponto de partida, utilizou o algoritmo de busca *Genetic Selection* que seleciona aleatoriamente um subconjunto de atributos de um subconjunto de candidatos utilizando algoritmo genético. Estas abordagens foram escolhidas por terem sido as que apresentaram os melhores resultados nas bases de dados estudadas nesta tese.

Avaliar o subconjunto de atributos selecionados é medir quão bom um determinado atributo é segundo um critério de avaliação (informação, distância, dependência, consistência, precisão). Em outras palavras, é como ele interage com o algoritmo de aprendizado. Essa interação pode ser subdividida, basicamente, em duas abordagens principais: filtro e *wrapper* (Kohavi e John, 1997). Neste estudo serão

utilizadas as abordagens filtro e *wrapper* e também a abordagem estatística de análise de componentes principais.

3.9.1 Abordagem filtro

A abordagem filtro introduz um processo separado, o qual ocorre antes da aplicação do algoritmo de aprendizagem propriamente dito. A idéia é filtrar atributos irrelevantes segundo algum critério antes do aprendizado ocorrer (Figura 3.2). Essa etapa do pré-processamento considera características gerais do conjunto de dados para selecionar alguns atributos e excluir outros. Sendo assim métodos de filtro são independentes do algoritmo de aprendizado que, simplesmente, receberá as instâncias somente com os subconjuntos dos atributos selecionados pelo método filtro.

Na abordagem filtro, a meta é selecionar um subconjunto de atributos que preserva a informação pertinente no conjunto inteiro de atributos (Freitas, 1998, p. 66). Esta abordagem usa como uma das técnicas de avaliação de atributos medidas de consistência (*consistency*). Estas medidas são fortemente dependentes do conjunto de treinamento e preferem hipóteses consistentes que possam ser definidas a partir do menor número possível de atributos. Assim, essas medidas encontram o subconjunto mínimo de atributos que satisfaz a proposta de inconsistência aceita, geralmente definida pelo usuário. Porém, um problema associado às medidas de consistência é que elas não conseguem distinguir entre dois atributos igualmente bons e, consequentemente, não conseguem detectar atributos redundantes. A inconsistência é definida como dois exemplos possuindo os mesmos valores de atributos, mas em classes diferentes.

Para Dash e Liu, 2003; Liu e Setiono, 1996, um subconjunto de atributos importantes é definido por meio da taxa de inconsistência quando:

1. um exemplo é considerado inconsistente se existirem pelo menos dois exemplos exatamente iguais exceto pelo valor da classe;
2. a contagem de inconsistência para um exemplo é dada pelo número de vezes que esse exemplo aparece nos dados subtraído o maior número entre as diferentes classes e
3. a taxa de inconsistência de um subconjunto de atributos é a soma de todas as contagens de inconsistência de todos os exemplos do subconjunto nos dados dividido pelo número N de exemplos.

Por exemplo, se para um determinado subconjunto de atributos, um exemplo inconsistente aparece N_{Ei} vêzes dos quais Nc_1, pertencem à classe Nc_2 pertencem classe C2 e Nc_3 pertencem a classe C3, sendo $N_{Ei} = Nc_1 + Nc_2 + Nc_3$. Se Nc_3 é o maior valor entre os três, a contagem de inconsistência é dada ($N_{Ei} - Nc_3$). Desse modo, dados um subconjunto de atributos e um limiar mínimo de taxa de inconsistência, definida pelo usuário, caso a taxa de inconsistência desse subconjunto seja menor que o limiar, ele será dito consistente. Em geral, essa medida é combinada com alguma outra, por exemplo, o tamanho do subconjunto de atributos em questão.

A outra técnica de avaliação de atributos na seleção aqui também utilizada foi o CFS (Seleção de atributos baseado em correlação). O CFS (Hall e Holmes, 2003) utiliza a seguinte métrica para fazer a avaliação:

$$U(X,Y) = 2.0 * \left[\frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)} \right] \quad (3.3)$$

Depois de calcular uma matriz de correlação, CFS aplica uma estratégia heurística de busca para encontrar um bom subconjunto de atributos de acordo com (3.4).

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (3.4)$$

onde H(X) e H(Y) de (3.4) são os atributos e, $Mérit_s$ heurística é o "mérito" de um subconjunto S_k contendo atributos, r_{cf} é a média da correlação entre classes, e r_{ff} a média da correlação entre os atributos.

Os algoritmos de busca do melhor subconjunto de atributos utilizados na abordagem filtro foram: *Genetic Selection* (GS) e *GreedyStepwise* (GSP).

Genetic Selection é realizado através de algoritmos genéticos (AGs) que são algoritmos de otimização global, baseados nos mecanismos de seleção natural e da genética. Eles empregam uma estratégia de busca paralela e estruturada direcionada à busca de pontos de “alta aptidão”, ou seja, pontos nos quais a função a ser minimizada ou maximizada tem valores relativamente baixos ou altos. Apesar de aleatórios, AG não são buscas aleatórios não-direcionadas, pois exploram informações históricas para encontrar novos pontos de busca onde são esperados melhores desempenhos. Embora

possam parecer simplistas do ponto de vista biológico, esses algoritmos são suficientemente complexos para fornecer mecanismos poderosos e robustos de busca adaptativa (Rezende, 2005, p.227).

O método *GreedyStepwise* seleciona as variáveis utilizando a abordagem seleção *forward* como direção de busca. Esta abordagem inicia a busca comum num conjunto vazio de atributos. Iterativamente, vai adicionando um atributo por vez ao subconjunto até que não se consiga melhorar a qualidade do subconjunto de atributos e pára quando a adição/eliminação dos atributos restantes começa a apresentar uma diminuição nos resultados avaliados. Esse método usa as possíveis combinações entre as estratégias e direções de busca completa e heurística. O GSP admite que uma variável selecionada em uma etapa possa ser eliminada em outra posterior. Este método pesquisa avidamente através do espaço de subconjuntos de atributos (Witten e Frank, 2005). Outra vantagem deste método é que o processo de seleção pode começar a partir da equação construída com um subconjunto de variáveis independentes e, inclusive, com todas elas (Aranaz, 1996, p.213).

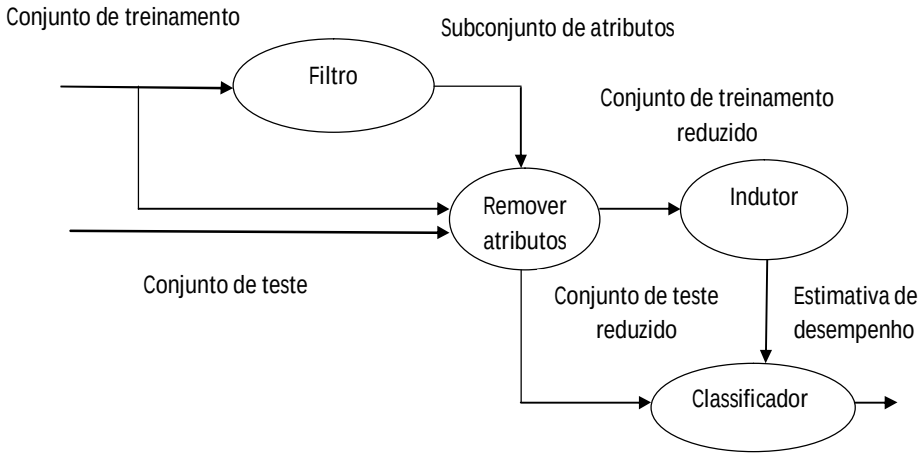

Figura 3. 7 Abordagem Filtro.

3.9.2 Abordagem *wrapper*

A abordagem *wrapper* ocorre em conjunto ao algoritmo básico de aprendizagem, porém, utilizando tal algoritmo como uma referência para analisar, a cada iteração, o subconjunto de atributos em questão – Figura 3.8. Em outras palavras, métodos *wrapper* geram um subconjunto candidato de atributos, executam o

algoritmo de aprendizado considerando apenas esse subconjunto de atributos selecionado do conjunto de treinamento, e utilizam a precisão resultante do classificador induzido para avaliar o subconjunto de atributos em questão. Esse processo é repetido para cada subconjunto de atributos até que o critério de parada determinado seja satisfeito.

Esta abordagem avalia os atributos usando estimativas de precisão providas por algoritmos de aprendizado pré-determinados (Freitas A. A., 1998, p. 66).

Um argumento utilizado com muita frequência para utilizar a abordagem *wrapper* é que o mesmo algoritmo de aprendizado que vai usar o subconjunto de atributos selecionado deve prover uma estimativa melhor de precisão que um outro algoritmo, o qual pode possuir um *bias*[5] de aprendizado totalmente diferente (Kohavi e John, 1997). Porém, a maior desvantagem dos métodos *wrapper* é o custo computacional, o qual resulta da execução do algoritmo de aprendizado para avaliar cada subconjunto de atributos a ser considerado (Pila, 2001; Lee *et al.*, 1999; Baranauskas *et al.*, 1999; Kohavi e John, 1997).

A abordagem utilizada para a direção de busca foi a seleção *forward* (definido no item anterior).

No estudo serão utilizados os mesmo indutores no modelo *wrapper* já definidos anteriormente, ou seja, RL, MLP, SVM e AD.

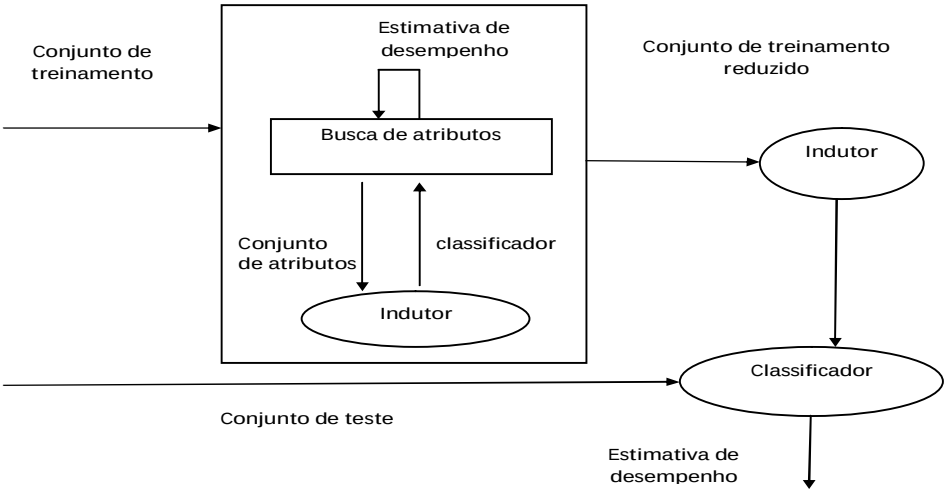

Figura 3. 8 Abordagem *Wrapper.*

[5] O *bias* de aprendizado é definido como qualquer preferência de uma hipótese sobre outra, além da simples consistência com os exemplos.

3.9.3 Análise de Componentes Principais (ACP)

Análise de componentes principais é um método estatístico. O objetivo da análise é tomar *p* variáveis X_1, X_2,..., X_p e encontrar combinações destas para produzir índices Z_1,Z_2,...Z_p que sejam não correlacionados na ordem de sua importância, e que descrevam a variação nos dados. A falta de correlação significa que os índices estão medindo diferentes "dimensões" dos dados, e a ordem é tal que Var (Z_1) ≥ Var (Z_2) ≥ ...≥ Var (Z_p), em que Var (Z_i) denota variância de Z_i. Os índices Z são então os componentes principais. Ao fazer uma análise de componentes principais, há sempre a esperança de que as variâncias da maioria dos índices serão tão baixas a ponto de serem desprezíveis. Neste caso, a maior parte da variação no conjunto de dados completos pode ser descrita adequadamente pelas poucas variáveis Z com variâncias que não são desprezíveis, e algum grau de economia é então alcançado (Manly, 2005, p.89). A idéia básica do método é de transformar *p* variáveis tipicamente correlacionadas em $k<p$ combinações lineares não correlacionadas.

3.10 Uma estratégia de predição de insolvência com seleção de atributos

Apresenta-se agora uma técnica para a seleção de características a ser acoplada no modelo de predição desenvolvido. A idéia é considerar a aplicação dos métodos de seleção de forma individualizada nas bases que compõem o comitê. Desta forma, setores econômicos representados nestas bases podem direcionar mais adequadamente o processo de seleção, priorizando atributos mais relevantes para os setores presentes na base em questão. Espera-se que este tratamento local dos atributos reforce o desempenho de cada uma das bases de dados que compõem o comitê tornando assim, o processo de votação majoritário mais robusto de uma forma geral. A seguir, complementa-se o pseudo código do modelo para predição de insolvência com a inclusão do procedimento de seleção de características.

Pseudo-código: comitê de classificadores para base de dados desbalanceadas com seleção de características (SEIDwS)

início

Defina o número de classificadores base n_{cb}

Defina o número de instâncias para cada classe n_{ic}

% construção dos n_{cb} *classificadores base*

Construa as instâncias da classe minoritária das n_{cb} *base de dados conforme algoritmo SEID*

Construa as instâncias da classe majoritária das n_{cb} *base de dados conforme algoritmo SEID*

% aplicação de uma estratégia de seleção de características

para i=1, n_{cb}

Aplique a seleção de características no classificador base i

fim

Treine os n_{cb} *classificadores base com os atributos selecionados para cada classificador base*

%classificação de novas instâncias

Aplique técnica de votação majoritária para classificar os dados de teste

fim.

3.11 Validações dos algoritmos propostos

A validação dos algoritmos propostos será realizada em três etapas visando atender três objetivos da validação, (i) testar os algoritmos propostos nesta tese em base de dados diferentes daquelas aqui estudadas; (ii) comparar os resultados gerados pelo SEID e SEIDwS com outras pesquisas realizadas nesse tema, e (iii) testar o mesmo algoritmo nas bases de dados aqui estudadas de acordo com cada amostra de empresas de diferentes setores econômicos. As etapas (i) e (ii) serão apresentadas aqui, já a etapa (iii) será apresentada nos capítulos seguintes.

O cumprimento da primeira etapa foi feito testando os algoritmos SEID e SEIDwS em oito bases de dados originadas do Repositório UCI para Aprendizado de Máquina[6]. Dessas bases de dados, três são normalmente utilizadas para testes de estudos sobre modelagem de previsão de insolvência (*Japanese Credit Screening, Australian Credit Approval, German Credit Data*) e outras cinco são utilizadas para testar estudos relacionados a bases de dados desbalanceados (*Breast Cancer Wisconsin Diagnostic, Haberman's Survival, Hepatitis, Ionosphere e Pima Indians Diabetes*).

[6] http://archive.ics.uci.edu/ml/

3.11.1 Validação do SEID e do SEIDwS nas bases do UCI Repositório para Aprendizado de Máquina

Nesta subseção são apresentados os resultados da validação do SEID através das oito bases do UCI, o procedimento é o mesmo apresentado na Figura 3.1. O classificador AD foi o utilizado.

A Tabela 3.3 apresenta os resultados dos testes do SEID e do algoritmo SMOTE, neste algoritmo o *K* usado foi igual a 5. As bases de dados utilizadas do UCI foram as que normalmente são utilizadas para testes na previsão de insolvência, nelas as classes são discriminadas em empresas insolventes (INS) e solventes (SOL). Já a Tabela 3.4 apresenta os resultados com bases de dados do UCI com fins de testar bases desbalanceadas.

Os softwares utilizados foram o WEKA 3.5.6 (Witten e Frank, 2005) e o Matlab 7.1. Em todas as análises apresentadas nas etapas de classificação e de seleção de atributos foram aplicadas 10 partições na validação cruzada.

Duas abordagens de direção de busca ou ponto de partida foram utilizadas: (i) seleção *forward* e (ii) seleção aleatória. A abordagem seleção aleatória utilizou o algoritmo de busca *Genetic Selection* (GS). Em GS foram usados os valores do tamanho da população e do número das gerações iguais a 20, e a probabilidade de *crossover* e mutação igual a 0.6 e 0.033, respectivamente.

Na abordagem filtro, as técnicas de avaliação de atributos foram medidas de consistência (c*onsistency*) e o CFS (Seleção de atributos baseado em correlação). Na abordagem *wrapper* os algoritmos indutores foram RL, ANNs, SVM e AD.

Os resultados apresentados nas tabelas seguintes, que obtiveram os melhores resultados, utilizaram GS, *wrapper* e AD.

Bases de dados do UCI	Nº de atributos	Classe	Instâncias	Base original		SEID		SMOTE	
				F	AUC	F	AUC	F	AUC
Japanese Credit Screening	15	INS	383	0,382	0,572	0,709	0,84	0,731	0,864
		SOL	307	0,77	0,572	0,866	0,84	0,878	0,864
Australian Credit Approval	14	INS	383	0	0,474	0,549	0,876	0,778	0,900
		SOL	307	0,96	0,474	0,954	0,876	0,967	0,900
German Credit Data	20	INS	300	0,547	0,73	0,836	0,900	0,798	0,881
		SOL	700	0,813	0,73	0,878	0,900	0,859	0,881

Tabela 3. 3 – Resultados dos testes do algoritmo SEID com as bases de dados sobre insolvência do UCI.

Já a Tabela 3.4 apresenta os resultados com bases de dados do UCI com fins de testar bases desbalanceadas.

Bases de dados do UCI	Nº de atributos	Classe	Instâncias	Base original		SEID		SMOTE	
				F	AUC	F	AUC	F	AUC
Breast Cancer Wisconsin Diagnostic	32	MIN	212	0,907	0,923	0,912	0,948	0,934	0,984
		MAJ	357	0,944	0,923	0,949	0,948	0,978	0,984
Haberman's Survival	3	MIN	81	0,358	0,6	0,5	0,703	0,521	0,731
		MAJ	225	0,82	0,6	0,821	0,703	0,82	0,731
Hepatitis	19	MIN	32	0,528	0,708	0,713	0,91	0,831	0,928
		MAJ	123	0,9	0,708	0,893	0,91	0,923	0,928
Ionosphere	34	MIN	126	0,874	0,892	0,94	0,987	0,976	0,989
		MAJ	225	0,935	0,892	0,949	0,987	0,987	0,989
Pima Indians Diabetes	8	MIN	268	0,802	0,751	0,792	0,887	0,798	0,896
		MAJ	500	0,614	0,751	0,838	0,887	0,891	0,896

Tabela 3. 4 – Resultados dos testes do algoritmo SEID com as bases de dados sobre bancos desbalanceados do UCI.

Os resultados dos testes feitos com o algoritmo SEID nas oito bases de dados do UCI mostraram que em sete bases os resultados foram bem convincentes - F com valores bem significativos nas classes minoritárias e AUC com valores bem acima de 0,8 - podendo ser concluído que essas técnicas podem ser utilizadas com um bom grau de eficácia. Tanto nas bases de previsão de crédito quanto nas bases desbalanceadas para testes os resultados, a exceção fica na base de dados Haberman´s. Na base Haberman´s os resultados fugiram a regra anterior e não podem ser considerados tão eficientes, entretanto a AUC obteve 0,7, valor aceitável. Nas outras bases de dados quando a comparação é feita com o algoritmo SMOTE os resultados são bem competitivos, chegando bem próximos aos resultados deste algoritmo.

Vale ressaltar, entretanto de que a técnica do algoritmo SEID trata somente com variáveis reais da base de dados diferente do SMOTE que acaba gerando dados artificiais em relação à base original caracterizando mais tendências e menos realizações. A comparação com a base original (sem balanceamento) evidencia a necessidade, a importância e a eficácia da aplicação de um algoritmo de balanceamento quando o custo do erro nas classes minoritárias é relevante, situação que ocorre em previsão de insolvência de empresas.

Os resultados encontrados pelo algoritmo SEID na base de dados *German*, base mais desbalanceada entre as testadas, foram melhores do que os encontrados pelo

SMOTE, podendo ser inferido de que em bases mais desbalanceadas os resultados do SEID são mais promissores o que os do SMOTE.

A tabela 3.5 apresenta os resultados dos testes do SEIDwS e do algoritmo SMOTE, neste algoritmo o K usado também foi igual a 5. Já a Tabela 3.6 apresenta os resultados com bases de dados do UCI com fins de testar bases desbalanceadas.

Bases de dados do UCI	Nº de	Classe	Instâncias	Base original		SEIDwS		SMOTE	
	atributos			F	AUC	F	AUC	F	AUC
Japanese Credit Screening	6	INS	383	0,38	0,57	0,73	0,88	0,77	0,91
		SOL	307	0,77	0,57	0,9	0,88	0,9	0,91
Australian Credit Approval	5	INS	383	0	0,47	0,57	0,88	0,79	0,93
		SOL	307	0,96	0,47	0,97	0,88	0,98	0,93
German Credit Data	7	INS	300	0,55	0,73	0,88	0,94	0,81	0,93
		SOL	700	0,81	0,73	0,93	0,94	0,93	0,93

Tabela 3.5 – Resultados dos testes do algoritmo SEIDwS com as bases de dados sobre insolvência do UCI.

Já a Tabela 3.6 apresenta os resultados com bases de dados do UCI com fins de testar bases desbalanceadas.

Bases de dados do UCI	Nº de	Classe	Instâncias	Base original		SEIDsS		SMOTE	
	atributos			F	AUC	F	AUC	F	AUC
Breast Cancer Wisconsin Diagnostic	9	MIN	212	0,907	0,923	0,93	0,983	0,988	0,992
		MAJ	357	0,944	0,923	0,96	0,983	0,993	0,992
Haberman's Survival	3	MIN	81	0,358	0,6	0,5	0,705	0,571	0,726
		MAJ	225	0,82	0,6	0,823	0,705	0,819	0,726
Hepatitis	7	MIN	32	0,528	0,708	0,753	0,94	0,848	0,948
		MAJ	123	0,9	0,708	0,918	0,94	0,959	0,948
Ionosphere	10	MIN	126	0,874	0,892	0,98	0,992	0,984	0,994
		MAJ	225	0,935	0,892	0,989	0,992	0,991	0,994
Pima Indians Diabetes	5	MIN	268	0,802	0,751	0,802	0,906	0,832	0,916
		MAJ	500	0,614	0,751	0,864	0,906	0,911	0,916

Tabela 3. 6 – Resultados dos testes do algoritmo SEIDwS com as bases de dados sobre bancos desbalanceados do UCI.

As Tabelas 3.3 e 3.4 apresentam os resultados dos testes feitos com algoritmos de balanceamentos (SEIDwS e SMOTE) sem a etapa da seleção de atributos, já nas Tabelas 3.5 e 3.5 são apresentados os resultados dos algoritmos de balanceamentos com a etapa da seleção de atributos. Pelos resultados apresentadas nas duas últimas Tabelas a realização da etapa de seleção de atributos exerce influência positiva na classificação de bancos desbalanceados, evidenciando, assim a necessidade de estar presente no pré-processamento de bancos desbalanceados a etapa da seleção de atributos. O SEIDwS apresentou resultados bem promissores e competitivos com o SMOTE como estratégia de balanceamento quando testado com dados do UCI.

A tabela 3.5 apresenta as comparações dos vários estudos publicados sobre o tema na literatura específica utilizando como parâmetros acurácia, Erro Tipo I e Erro Tipo II. As comparações foram feitas através dos melhores resultados encontrados pelos autores. Os estudos utilizados para comparação são de Tsai (2008), Tsai e Wu (2008) e Nanni e Lumini (2009). Estes autores utilizaram as bases de dados do UCI, as mesmas utilizadas pelo algoritmo proposto nesta tese, o SEIDwS.

	SEIDwS	Tsai e Wu	Tsai	Nanni e Lumini
Japanese Credit Screening	%	%	%	%
Acurácia	88,64	87,94	85,88	86,38
Erro Tipo I	13,02	14,42	90,05	18,8
Erro Tipo II	9,92	10,05	22,40	9,4
Australian Credit Approval	%	%	%	%
Acurácia	90,67	97,32	81,93	85,89
Erro Tipo I	14,23	12,16	21,89	17,4
Erro Tipo II	12,02	11,55	13,89	11,8
German Credit Data	%	%	%	%
Acurácia	83,52	78,97	74,28	73,93
Erro Tipo I	28	44,27	55,39	60
Erro Tipo II	7,54	8,46	9,63	18,2

Tabela 3. 5– Comparação dos resultados do algoritmo SEIDwS com as bases de dados do UCI com outros estudos publicados.

Na tabela 3.5, os resultados mostram a eficácia do algoritmo SEIDwS. A comparação mostra que SEIDwS obteve melhores resultados na acurácia, nos Erros Tipo I e II, e que em todos esses parâmetros há um ganho do SEIDwS sobre os outros estudos. No Erro Tipo II (classifica instância falidas no grupo das não falidas) o SEIDwS obteve melhores resultados sobre os outros testes em dois das três bases de dados. Somente na base *Japanese Credit Screening* do estudo de Nanni e Lumini os resultados ficaram um pouco inferiores, (9,92 X 9,4).

CAPÍTULO 4 Base de dados na previsão de insolvência de empresas

Empresas insolventes começam a acusar sinais de dificuldades bem antes de chegar ao ponto crítico de uma falência ou concordata, vários estudos mostram isso entre outros, Altman, 1968; Kanitz, 1978; Silva, 2006; Kumar e Ravi, 2007; Sun e Li, 2008; Nanni e Lumini, 2009. É compreensível que a insolvência, sendo um processo com começo, meio e fim, que existam nos demonstrativos contábeis publicados, antes da tragédia final, alguns indícios do que está para acontecer. Para encontrar estes indícios, foram aplicadas nesta tese, técnicas de DM em uma base de dados elaborada através de demonstrativos contábeis de empresas brasileiras de capital aberto. O objetivo principal desta tese é caracterizar melhor estes indícios, sobretudo através da discriminação daquelas empresas que apresentam grandes possibilidades de virem a se tornar insolventes. Nas seções a seguir são apresentados todos os procedimentos desenvolvidos neste estudo utilizando a base de dados de empresas da Bovespa. Tais dados tendem a ser bastante diferenciados tanto em termos de setores econômicos das empresas quanto em relação à quantidade de empresas destes setores econômicos.

4.1 Descrição da montagem da base de dados

Para desenvolver um estudo de previsão de insolvência (classificação) a base de dados é composta por dois grupos de instâncias. O primeiro grupo refere-se a um conjunto de empresas que apresentam problema de insolvência em determinado período. Tal problema é aqui entendido como sendo uma solicitação de concordata, decretação da falência ou recuperação judicial de acordo com a Lei nº 11.101 de 2005. O segundo grupo diz respeito a empresas saudáveis no sentido precisamente contrário ao do primeiro grupo, ou seja, empresas que não apresentaram problemas de insolvência no período de tempo estudado. Deve-se ressaltar que não é previsto nesta base de dados um caso intermediário de empresas que estão em processo de insolvência. Logicamente, o processo de insolvência demanda um determinado tempo. A expectativa é que, considerando-se dados temporais, os dados das instâncias insolventes capturem esta tendência no período considerado.

Em relação aos dois grupos, foram consideradas apenas empresas comerciais e industriais privadas de capital aberto negociadas na BOVESPA e classificadas por setor econômico. Tal restrição pretende estabelecer um corte qualitativo, pois no caso brasileiro, as demonstrações contábeis, fundamentais para a análise em questão, estão disponíveis de forma regulamentada apenas para tais empresas. As empresas do setor financeiro não foram consideradas porque os seus demonstrativos contábeis são elaborados de acordo com legislações específicas para o setor. A baixa confiabilidade das demonstrações contábeis das empresas de capital fechado levou a sua não inclusão no presente estudo.

Por outro lado, o fato de também não serem consideradas neste estudo as empresas estatais, refere-se à possibilidade de tais empresas terem uma fonte adicional de recursos não disponível como as demais, isto é, podem ter seus problemas resolvidos por outros mecanismos que não os estabelecidos pelo mercado. Da mesma forma e pelas mesmas razões, não foram incluídas empresas de controle estrangeiro neste estudo. Também não se incluíram dados de balanços consolidados (*holding*), com vistas a um estudo de tipo singular das empresas. Foi respeitada a proporção do ativo das empresas para evitar distorções patrimoniais e também considerada as suas localizações geográficas.

Para a realização desta tese foi construído primeiro o grupo de empresas que haviam requerido concordata ou falência, o que representa o primeiro passo para estabelecer grupos específicos de classificação. A partir deste grupo, atendendo a segmentação econômica das empresas, criou-se a base de dados principal do trabalho, formada por empresas insolventes e empresas saudáveis financeiramente, ou seja, solventes.

A partir do grupo das empresas insolventes foram analisados os demonstrativos contábeis, Balanço Patrimonial e Demonstrativo de Resultado do Exercício do ano ao pedido de concordata ou falência, e de quatro anos precedentes ao pedido. Assim, foram obtidos 22 indicadores contábeis anuais das empresas, com base na literatura específica do assunto, que foram classificadas como concordatária ou falida na BOVESPA durante o período de 1996 a 2006. Para cada empresa classificada como insolvente, foi selecionada uma quantidade superior de empresas de capital aberto com controle privado nacional, financeiramente saudáveis (no sentido de que não há solicitação de concordata por parte da empresa no período considerado), com tamanho do ativo, sempre que possível compatível, e pertencente ao mesmo setor de atividade, buscando

respeitar, localização geográfica e idade. O estabelecimento de uma quantidade superior de empresas sadias para cada inadimplente, por outro lado, baseia-se na hipótese de que quanto maior a quantidade de dados existente, menor a probabilidade de erro e objetivando, também ficar mais próximo da realidade econômica.

A escolha do período para o estudo de caso deve-se a implantação do Plano Real em julho de 1994, fato que gerou muitas mudanças nos ambientes externos e internos das empresas, sendo que o fato mais significativo referente aos demonstrativos contábeis foi a considerável queda dos índices inflacionários levando a uma maior confiabilidade dos mesmos, o que propicia obter mais informações gerenciais e econômicas confiáveis sobre as empresas. O outro motivo que também determinou a escolha do período estudado foi o estabelecimento da nova lei de recuperação e falências no Brasil. Portanto, decidiu-se pelo conjunto dos anos de observações após o Plano Real e que estivesse, supostamente, sob pouca influência da nova lei de falências que entrou em vigor em junho de 2005.

A base de dados empregada nesta tese, o grupo de empresas insolventes é muito reduzido devido às seguintes razões:

a) Um grande número de empresas que vão à falência, tais como as limitadas, não tem obrigação de publicar seus balanços;
b) Muitas empresas de Sociedades Anônimas que faliram, publicaram balanços muito sumários para que pudessem ser analisados;
c) Foram excluídas as empresas de serviços, as imobiliárias, as instituições financeiras e de participações porque não pertencem aos setores econômicos que se pretende estudar.
d) Em alguns casos teve-se disponibilidade a dados contábeis confiáveis de empresas insolventes, mas não ocorreu o mesmo com os dados contábeis das empresas solventes do mesmo setor econômico.

4.2 Bases de dados consideradas

Para a realização deste trabalho foram construídas duas bases de dados que se diferenciam quanto ao tipo de dados que as compõem: variáveis sequenciais e variáveis de painel.

4.2.1 Base de dados com variáveis sequenciais

Nesta abordagem, foi elaborada a base de dados considerando-se variáveis sequenciais que foram caracterizadas no segundo capítulo. A base de dados seqüenciais totaliza 175 empresas, com 147 classificadas como solventes e 28 classificadas como insolventes durante o período em estudo.

Foi considerado na base de dados sequenciais um total de 22 variáveis para cada empresa selecionada. Para cada uma destas variáveis sequenciais são definidos os períodos nos quais elas pertencem, por exemplo, LC_1 representa o valor da variável Liquidez Corrente de dois períodos anteriores ao ano em que a empresa foi declarada insolvente; LC_2 representa o valor desta mesma Liquidez Corrente no ano anterior à insolvência da empresa e LC_3 o valor da liquidez corrente do ano em que a empresa foi considerada insolvente. Sendo assim, nesta abordagem, as 22 variáveis contábeis consideradas se transformam em 66 (22 x 3 anos). Uma apresentação mais detalhada destas variáveis será feita na seção a seguir.

O período de anos foi escolhido de acordo com o período que vem sendo usado na literatura especifica, como por exemplo, Altman *et al.,* 1994 e Silva, 2006. Para estes autores, modelos de previsão de insolvência com variáveis contábeis sequenciais se mostraram bem eficientes considerando-se até três anos anteriores a insolvência. A partir de três anos, as variáveis começam a perder significativa capacidade de caracterizar a insolvência (redução na sua entropia), e o seu nível informacional para esse fim diminui. Para os mesmos autores, as variáveis sequenciais mais capazes de caracterizar a insolvência são aquelas referentes ao período em que a empresa é declarada insolvente, entretanto, dizem os autores, variáveis deste período não devem ser utilizadas na prática por pertencerem exatamente ao período da ocorrência da insolvência. Tais variáveis refletem bem a deterioração financeira em que se encontrava a empresa neste ano. Portanto, os melhores modelos devem ser aqueles que utilizam variáveis sequenciais mais distantes do ano da declaração da insolvência, permitindo assim que a empresa tome ações em períodos mais longos, visando impedir a sua deterioração financeira. Por outro lado, deve-se tomar cuidado para que a capacidade de previsão do modelo também não fique comprometida, pelo motivo das variáveis consideradas estarem mais distantes do período de declaração da insolvência, deteriorando sua entropia para esta modelagem.

Nesta base de dados as variáveis são discriminadas pelo período no índice subscrito. Por exemplo, a variável LC (liquidez corrente), quando representa o período no qual a empresa foi declarada insolvente se apresenta como LC_3, quando, porém a liquidez corrente se refere a um período (um ano) antes da empresa se declarar insolvente ela é representada por LC_2. Portanto, os índices subscritos das variáveis representam o período no qual a empresa está situada no tempo em relação ao período em que ela foi declarada insolvente.

4.2.2 Base de dados com variáveis de painel

Já nesta abordagem, considerando-se dados de painel, foram criadas 1.610 instâncias sendo 1.470 referentes a empresas solventes e 140 empresas insolventes. Nesta abordagem, a base foi composta por dados referentes aos demonstrativos contábeis dos cinco anos anteriores ao ano em que a empresa foi declarada insolvente. De acordo com Altman *et al.*, 1994 e Hung e Chen, 2009 as empresas insolventes começam a apresentar características ou indícios de insolvência num período de cinco anos antes do ano que ocorre efetivamente a falência, sendo que esses indícios apresentam mais consistência (maior entropia) no período próximo de três anos antes.

Foram utilizados dados de empresas solventes num período de dez anos para avaliar o comportamento da predição em bancos desbalanceados já que neste período ocorre uma boa caracterização dessa empresas. Outros motivos foram: (i) uma adequação ao ano (2005) no qual ocorreu a mudança na lei de falência e também (ii) a considerável queda dos índices inflacionários ocorrido no ano de 1994 refletindo melhor nos demonstrativos no ano de 1996, o inicio da base de dados.

Diante disso o conjunto de dados das empresas insolventes, com 28 empresas, totalizou 140 instâncias (28x5). Já o conjunto de dados das solventes totalizou 1.470 (147x10) instâncias com dados financeiros. As variáveis consideradas totalizam 22.

A escolha dos referidos períodos nesta tese, tanto nas variáveis sequenciais como para as variáveis de painel, visa utilizar os dados contábeis como sistema de informação no qual é apto a fornecer informações sobre processos e eventos de negócio que afetam a organização e também apresentando desempenhos passados de uma empresa e, portanto pode ser um bom informativo para predizer o futuro (Horngren *et al.*, 1999, p. 248).

4.3 Variáveis consideradas

Os 22 indicadores financeiros (variáveis) considerados neste estudo não são a totalidade das informações disponíveis nos demonstrativos contábeis, mas sim os únicos geralmente possíveis de obtenção para o estudo em questão e os mais utilizados para análise das demonstrações contábeis. A partir destes dados disponíveis é que foram determinados os indicadores ou índices financeiros, cuja finalidade é elaborar os modelos de previsão de insolvência.

Os indicadores financeiros considerados podem ser classificados de acordo com Iudícibus (1998); Matarazzo (2003); Pereira (2006) em três grandes grupos de índices contábeis-financeiros: liquidez, endividamento e rentabilidade. Estes índices são descritos detalhadamente a seguir.

4.3.1 Índices de liquidez

Os índices de liquidez visam fornecer um indicador da capacidade da empresa de pagar suas dívidas a partir da comparação entre os direitos realizáveis e as suas exigibilidades. A seguir serão definidos os índices de liquidez aqui utilizados e a sua relação entre contas ou grupos de contas das demonstrações contábeis.

- Liquidez Imediata (LI) - Determina a capacidade da empresa de pagar suas obrigações de curto prazo, valendo-se de suas disponibilidades em caixa, bancos ou aplicadas no mercado financeiro de curtíssimo prazo.

  LI = Disponibilidades / Passivo Circulante.

- Liquidez Corrente (LC) – Determina a capacidade da empresa de pagar suas obrigações de curto prazo.

  LC = Ativo Circulante / Passivo Circulante.

- Liquidez Seca (LS) - Indica a capacidade da empresa de pagar suas obrigações de curtíssimo prazo, valendo-se de seus ativos mais líquidos.

  LS = (Ativo Circulante – Estoques) / Passivo Circulante.

- Liquidez Geral (LG) – Determina a capacidade da empresa de pagar todas as obrigações de curto e longo prazo.

LG = (Ativo Circulante + Realizável Longo Prazo) / (Passivo Circulante + Exigível Longo Prazo).

- Relação Saldo de Tesouraria sobre Ativo Total (RSTA) – Mostra, em percentuais, a proporção do Saldo de Tesouraria (ST) sobre o Ativo Total (AT). Onde:

  ST = CCL – NCG

  Onde,

  Necessidade de Capital de Giro (NCG) = (Clientes – Provisão para devedores Duvidosos + Adiantamento a fornecedores + Estoques + Imposto a Recuperar + Outros Bens e Valores) – (Fornecedores + Adiantamento de Cliente + Salários e Contribuições Sociais + Impostos, Taxas e Contribuições a recolher+ Outras Contas a Pagar).

  Capital Circulante Líquido (CCL) = Ativo circulante – Passivo circulante

  RSTA = ST / AT.

- Termômetro financeiro (TERFIN) – Esse índice evidencia uma reserva financeira da empresa para fazer frente às eventuais expansões da necessidade de investimento operacional em giro, principalmente aquelas de natureza sazonal. Assim, necessidades transitórias de investimento em giro podem ser cobertas até o limite do saldo disponível existente.

  TERFIN = (Saldo de Tesouraria / Necessidade de Capital de Giro)

4.3.2 Índices de endividamento

Os índices referentes a endividamento são os seguintes:

- Endividamento Total sobre Ativo Total (ETAT) – Revela a dependência da empresa de suas exigibilidades, isto é, qual a participação dos recursos de terceiros no montante investido em seus ativos.

  ETAT = (Passivo Circulante + Exigível Longo Prazo + Duplicatas Descontadas) / AT.

- Endividamento Oneroso sobre Ativo Total (EOAT) – Indica a participação das fontes externas de financiamento (Recursos Externos) em relação ao AT da empresa.

EOAT = (Empréstimos e Financiamentos Totais de Curto Prazo+Empréstimos e Financiamentos Totais de Longo Prazo + Debêntures de Curto Prazo+Debêntures de Longo+Empresas Coligadas e Controladas de Curto Prazo+Empresas Coligadas e Controladas de Longo Prazo+Duplicatas Descontadas) / AT.

- Endividamento Total sobre o Patrimônio Líquido (ETPL) – Mede a proporção dos recursos de terceiros em relação aos recursos próprios existentes na empresa. Indica, também, a capacidade que a entidade tem de levantar fundos no mercado com base no montante de seus recursos próprios.

  ETPL = (Passivo Circulante + Exigível Longo Prazo + Duplicatas Descontadas) / Patrimônio Líquido.

- Endividamento Oneroso sobre o Patrimônio Líquido (EOPL) – Indica a participação das fontes externas de financiamentos em relação ao capital próprio. Tornam insolventes apresentam alta variação em períodos contínuos.

  EOPL = (Empréstimos e Financiamentos Totais de Curto Prazo+Empréstimos e Financiamentos Totais de Longo Prazo + Debêntures de Curto Prazo+Debêntures de Longo+Empresas Coligadas e Controladas de Curto Prazo+Empresas Coligadas e Controladas de Longo Prazo+Duplicatas Descontadas)/ Patrimônio Líquido.

- Endividamento Oneroso de Curto Prazo sobre Oneroso Total (EOCpOT) – Indica a porcentagem do volume da dívida onerosa de curto prazo em relação à dívida onerosa total (curto e longo prazo).

  EOCpOT= (Empréstimos e Financiamentos Totais de Curto Prazo+ Debêntures de Curto Prazo+Empresas Coligadas e Controladas de Curto Prazo+ Duplicatas Descontadas) / (Empréstimos e Financiamentos Totais de Curto Prazo+Empréstimos e Financiamentos de Longo Prazo + Debêntures de Curto Prazo+Debêntures de Longo+Empresas Coligadas e Controladas de Curto Prazo+Empresas Coligadas e Controladas de Longo Prazo+Duplicatas Descontadas).

- Grau de Imobilização do Capital Próprio (IMCP) – Determina a participação do Patrimônio Líquido em relação ao volume total investido no Ativo Permanente.

  IMCP = Ativo Permanente / (Exigível Longo Prazo + Patrimônio Líquido).

- Grau de alavancagem financeira (GAF) – Determina quanto a empresa consegue alavancar, ou seja, aumentar o lucro líquido através da estrutura de financiamento.

  GAF = Retorno sobre o Patrimônio Líquido / Retorno sobre o Ativo Total.

4.3.3 Índices de rentabilidade

Os índices neste grupo visam medir o quão eficiente a empresa usa seus ativos e administra suas operações, e são os seguintes:

- Margem Bruta (MB) – Identifica o desempenho dos custos de produção, dado que o lucro bruto é obtido da diferença entre as vendas líquidas e o custo dos produtos ou das mercadorias vendidas.

  MB = Lucro Bruto / Receita Operacional Líquida.

- Margem Operacional (MO) – Identifica o desempenho operacional da empresa computado o resultado financeiro mais o resultado de equivalência patrimonial.

  MO = Lucro Operacional / Receita Operacional Líquida.

- Margem Líquida (ML) – Representa o percentual da receita operacional líquida que sobrou após serem deduzidas todas as despesas e computados os resultados não operacionais, a provisão de imposto de renda e as participações estatutárias.

  ML = Lucro Líquido / Receita Operacional Líquida.

- Rentabilidade sobre o Patrimônio Líquido (ROE) - Corresponde à rentabilidade que a empresa propiciou aos recursos investidos pelos seus acionistas.

  ROE = Lucro Líquido / Patrimônio Líquido.

- Giro dos Ativos (GA) – Relaciona as vendas líquidas da empresa com seu ativo total (ou investimento), de forma a demonstrar seu giro, verificando se o volume das vendas realizadas no período foi adequado em relação ao Capital Total investido na empresa.

  GA = Receita Operacional Líquida / AT.

- Rentabilidade Operacional do AT (ROA) – Indica a lucratividade operacional que a empresa propicia em relação ao total de investimentos.

  ROA = Lucro Operacional / AT.

- Rentabilidade Líquida do AT (ROI) – Mostra o retorno do total de investimentos efetuados pela empresa, ou seja, a capacidade que os ativos apresentam de gerar lucros.

  ROI = Lucro Líquido / AT.

- Lucro antes dos juros, impostos (sobre lucros) (EBIT) - Representa a eficiência financeira da empresa determinada pelas estratégias operacionais adotadas excluindo os valores de depreciação e amortização (não demandam contrapartida monetária imediata).

  EBIT = Lucro Operacional – Resultado Financeiro.

- Lucro antes dos juros, impostos (sobre lucros), depreciações/exaustão e amortização (EBITDA) – Revela, em essência, a genuína capacidade operacional de geração de caixa da empresa, ou seja, sua eficiência financeira determinada pelas estratégias operacionais adotadas.

  EBITDA = EBIT + Depreciação, Amortização e Exaustão – Equivalência Patrimonial.

- Retorno sobre o patrimônio líquido adaptado (RTA) (Modelo Du Pont adaptado) – Mostra como o retorno sobre o Patrimônio Líquido é afetado pelo Giro do Ativo, pela Margem Bruta e pela alavancagem.

  RTA = GA x MB.

Assim, os atributos da base de dados serão compostos por 5 variáveis relacionadas à liquidez das empresas, 7 relativas ao endividamento e 10 associadas à rentabilidade das empresas envolvidas. Acredita-se que estas variáveis ou, pelo menos algumas delas, tenham potencial para capturar indícios de insolvência em um processo de discriminação.

CAPÍTULO 5 Predição de insolvência de empresas

Neste capítulo serão avaliadas as bases com dados do tipo sequencial e de painel em relação a suas capacidades de predição de um processo de insolvência. Inicialmente, aplicam-se algumas técnicas de classificação consideradas eficientes para avaliar o desempenho destas bases sem um tratamento especifico. A seguir, utiliza-se o SEID, comparando-se com os resultados obtidos com o algoritmo de balanceamento SMOTE. Consideram-se também procedimentos de seleção de características do SEIDwS, bem como avaliação da insolvência de empresas por setor econômico.

5.1 Avaliação da base de dados sequencial

Inicialmente uma avaliação de alguns classificadores geralmente utilizados para previsão de insolvência será feita sobre a base sequencial, sem nenhum tipo de tratamento adicional.

5.1.1 Aplicação de classificadores na base sequencial

As técnicas empregadas para a classificação das empresas são: Regressão Logística (RL), Máquina de Vetor Suporte (SVM), *Multilayerperceptron* (MLP), e Árvore de Decisão (AD). Estes classificadores foram escolhidos por serem considerados eficientes bem como por serem largamente utilizados na determinação de insolvência de empresas. O principal objetivo é avaliar a qualidade da predição obtida com a aplicação direta do discriminante sobre uma base de dados com o perfil da base sequencial, ou seja, desbalanceada e com diferentes setores econômicos representados.

Foram feitos ajustes paramétricos prévios para cada classificador utilizado, visando obter uma parametrização adequada para esta base. Os resultados apresentados são obtidos por meio de validação cruzada com 10 partes. Para que haja um melhor entendimento do desempenho de cada classificador, os resultados para cada classificador são apresentados por meio da matriz de confusão, medida F e AUC.

Na Tabela 5.1 são apresentados os resultados para a base sequencial original, isto é, desbalanceada e com todos os setores representados. Os parâmetros utilizados pelos classificadores foram os seguintes:

(i) Na RL utilizou-se o valor *ridge* no log-verossimilhança igual a 1,0 E^{-8}

(ii) No SVM o kernel utilizado foi o RBF com δ igual a 0,01 e *c* igual a 1,0.
(iii) MLP empregou-se uma camada de entrada, uma oculta e uma de saída, com número de ciclos igual a 500.
(iv) Na AD utilizou-se como fator de confiança 0,25 e o número de poda igual a 3.

Inicialmente foi utilizada a base de dados original desbalanceada sem seleção de atributos e também sem distinção de setor econômico.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	17	11	0,63	0,757	8	20	0,39	0,626	13	15	0,553	0,696	20	8	0,769	0,888
S	9	138	0,932	0,757	5	142	0,919	0,626	6	141	0,93	0,696	4	143	0,959	0,888

I = Insolventes; S = Solventes; MC = Matriz de confusão; F = Medida F; AUC = Área sob a curva ROC.

Tabela 5. 1 Resultados da aplicação dos classificadores na base de dados do tipo sequencial

Este resultado corrobora com a afirmação de Japkowicz e Stephen, 2002 de que a maioria dos algoritmos de aprendizado de máquina apresenta dificuldades em criar um modelo numa base de dados muito desbalanceada que classifique com precisão os exemplos da classe minoritária.

De acordo com os resultados encontrados e apresentados na Tabela 5.1 o classificador que obteve os melhores resultados no processamento da base de dados original foi AD. Este classificador obteve uma melhor matriz de confusão, medida F e AUC. Em seguida o classificador que obteve os melhores resultados foi o de RL.

Como o classificador AD foi o melhor dos métodos de classificação empregados ele será usado nos outros testes desta seção da tese. A seguir será avaliado o impacto da aplicação de estratégias de balanceamento na base sequencial.

5.1.2 Balanceamento da base de dados sequencial

Objetiva-se, nesta seção uma avaliação empírica da influência do balanceamento em bases com o perfil de dados sequencial. Aplica-se uma distribuição das instâncias da base, de acordo com o SEID, gerando bancos com menos instâncias, porém balanceados visando avaliar a qualidade da predição. Deve-se destacar que não se trata da aplicação do SEID, mas somente da avaliação dos bancos que formarão o comitê de classificadores de uma forma individual. Um bom desempenho do classificador nestes bancos menores, mas balanceados, servem de indício de que o SEID poderá apresentar

resultados eficientes. Deve-se ressaltar que, pelo tamanho da classe minoritária na base seqüencial, optou-se pela geração de 3 sub-bases de dados balanceadas. Destaca-se, também, que a desigualdade dada pela inequação (3.3) não será obedecida de uma forma rígida. Optou-se por considerar este valor somente como referência para definir o número de instâncias por sub-bases. Desta forma, foi possível testar algumas variações no número de instâncias que definem as sub-bases.

Na figura 5.1 é apresentado o fluxo do procedimento feito para se chegar aos resultados encontrados.

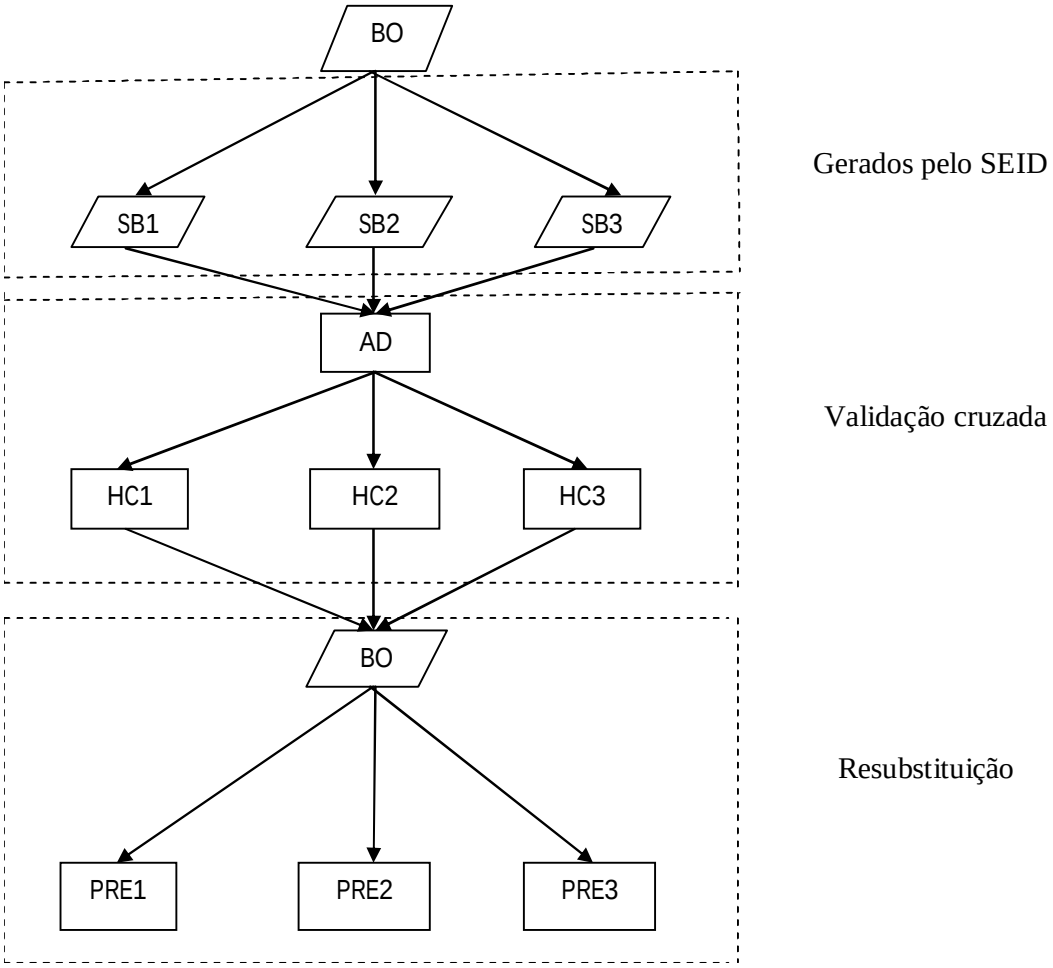

Figura 5. 1- Fluxo referente aos procedimentos para se chegar aos resultados após os balanceamentos da base original.

BO – Base de dados original
SEID – Algoritmo proposto nesta tese
SB1,2,3 – Sub-bases balanceadas através do SEID para treinamento
HC1,2,3- Hipótese de classificação 1,2 e 3
AD – Classificador árvore de decisão
PRE1,2,3 – Predição (MC, F, AUC) 1,2,3.

O fluxograma acima é composto de três etapas: a primeira etapa onde são geradas as sub-bases pelo SEID; a segunda etapa em que são elaboradas as hipótese de classificação de cada sub-base por meio de validação cruzada; e a terceira etapa onde é feita a validação da base original por resubstituição.

Inicialmente, serão geradas as 3 sub-bases, através do SEID, com um total de 60 empresas em cada sub-base. Neste caso, optou-se por aplicar um *under-sampling* na classe majoritária. Na Tabela 5.2 são apresentados os resultados referentes às classificações na base original geradas pelas hipóteses de classificação destas sub-bases.

	HC1_60				HC2_60				HC3_60			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	28	0	0,823	0,932	26	2	0,838	0,941	27	1	0,857	0,950
S	14	133	0,956	0,932	8	139	0,965	0,941	9	138	0,968	0,950

Tabela 5. 2 Resultados referentes à base de dados com 60 instâncias.

Na Tabela 5.2 pode-se observar, através dos resultados obtidos nos modelos (HCs) gerados por SEID, que a classe minoritária (empresas insolventes) apresentou uma acentuada melhora em relação aos resultados empregando a base de dados desbalanceada, apresentados na Tabela 5.1. O modelo HC3 obteve as melhores medidas de F e AUC.

Em seguida, foi feito um novo teste aumentando-se o tamanho das bases de dados entre as empresas solventes e insolventes. A composição adotada foi de 40 empresas solventes e 40 empresas insolventes, totalizando, portanto 80 empresas. Estas bases de dados foram geradas também pelo SEID.

Na Tabela 5.3 são apresentados os resultados dos 3 modelos (HC1_80, HC2_80 e HC3_80) referentes às classificações na base original após os procedimentos de balanceamento com o SEID que gerou as novas três bases balanceadas com um total de 80 empresas em cada uma.

	HC1_80				HC2_80				HC3_80			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	25	3	0,793	0,902	26	2	0,764	0,923	26	2	0,818	0,931
S	13	136	0,954	0,902	14	133	0,943	0,923	11	136	0,957	0,931

Tabela 5. 3 Resultados referentes à base de dados com 80 instâncias

Para esses novos modelos gerados pelas novas bases de dados o HC3 apresentou o melhor valor para a área ROC 0,931 a medida F também foi o melhor resultado até essa etapa. Esses modelos não foram melhores do que os gerados na etapa anterior.

5.1.3 Balanceamento e seleção de características para base sequencial

Estuda-se nesta seção o desempenho do balanceamento da base de dados aplicados em conjunto com uma estratégia de seleção de características.

Uma forma natural e razoável para viabilizar esta combinação se dá por meio da aplicação de um modelo de seleção (filtro) na base de dados original. Definidos os atributos mais relevantes estes são utilizados para a construção de cada sub-base balanceada que será utilizada no comitê. Estes modelos foram testados e os são apresentados na tabela 5.4 utilizando as abordagens filtro, ACP e *wrapper* respectivamente.

	Filtro				ACP				Wrapper			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	16	12	0,351	0,674	15	13	0,37	0,531	18	9	0,444	0,745
S	47	100	0,772	0,674	38	109	0,743	0,531	36	111	0,749	0,745

Tabela 5.4 – Resultados dos modelos com seleção de atributos feita anteriormente ao balanceamento.

Analisando as tabelas nota-se que os resultados não foram compatíveis para um nível aceitável para uma previsão de insolvência de empresas, com valores de F com média 0,38 e AUC com média próxima de 0,65 para as empresas classificadas em insolventes.

Uma possível justificativa para a obtenção de resultados tão inexpressivos com a seleção de atributos é a composição heterogênea das sub-bases de dados em relação aos setores econômicos representados.

A seguir, aplica-se então, a estratégia de seleção determinada no algoritmo SEID com seleção (SEIDwS) para a determinação das sub-bases.

Neste modelo, a seleção é feita de forma individual para cada sub-base, permitindo variação entre os atributos selecionados para cada sub-base. A Figura 5.2 se refere ao fluxo dos procedimentos para se chegar aos resultados após os balanceamentos e a seleção de atributos da base de dados original.

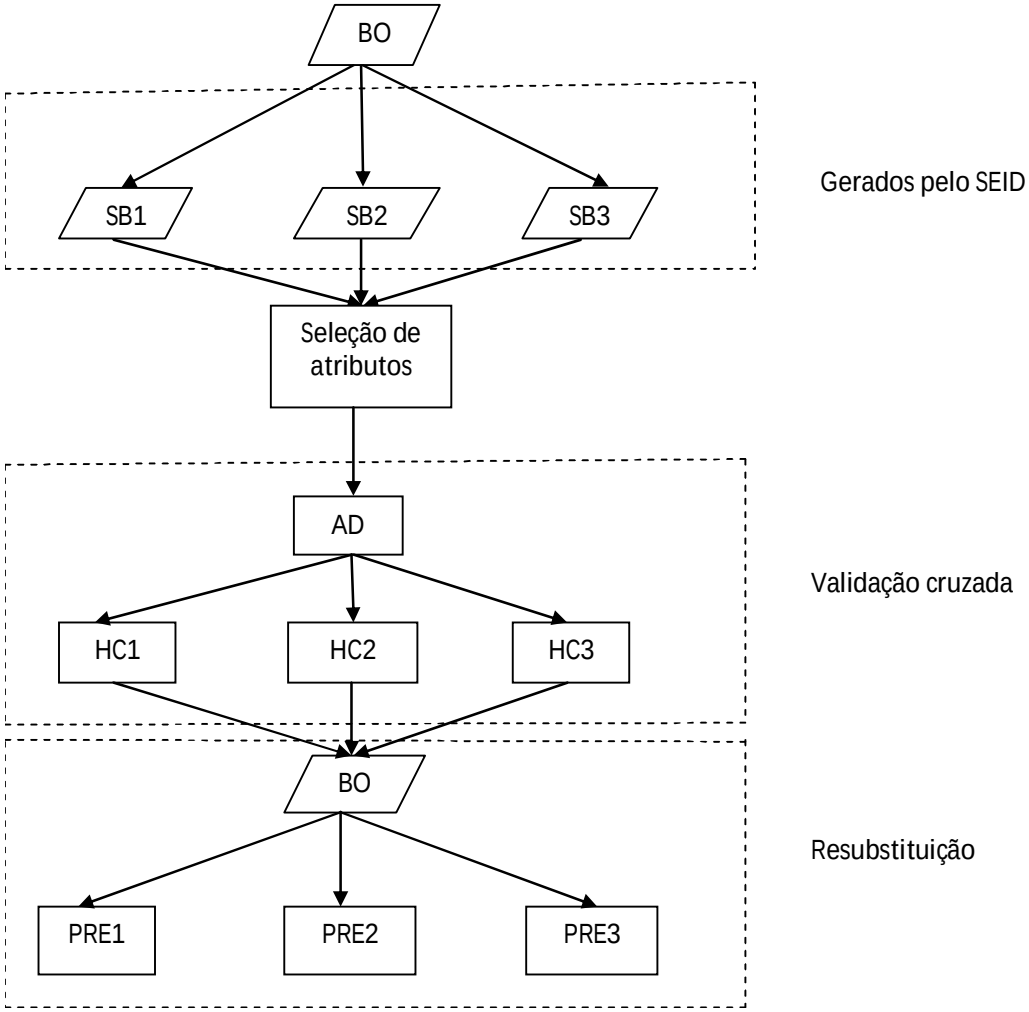

Figura 5. 2- Fluxograma referente aos procedimentos para se chegar aos resultados após os balanceamentos e a seleção de atributos da base de dados original.

O procedimento aqui é semelhante ao representado na figura 5.1, entretanto neste fluxograma ocorre a seleção de atributos após a geração das sub-bases, portanto nesta subseção a classificação é realizada após a seleção das variáveis independentes. A seleção de atributos utilizou as abordagens filtro, ACP e *wrapper*.

5.1.3.1 Resultados para a abordagem filtro de seleção de atributos

Nas aplicações das abordagens filtro foram testados na avaliação dos subconjuntos de atributos os algoritmos *CfsSubsetEval* (M. Hall, 2000; Mark A. Hall e Geoffrey Holme, 2003) e C*onsistency* (Witten e Frank, 2005). Já nos métodos de busca foram testados GS (*genetic search*) com GD (*greedystepwise*). Os melhores resultados foram encontrados aplicando *Consistency* e GS como algoritmos de avaliação e busca respectivamente e os procedimentos executados são os já explicitados na Figura 5.2. Os resultados encontrados estão na Tabela 5.4, onde os HCs representam os modelos gerados empregando-se a base de dados original.

	HC1_60 CONS+GS				HC2_60 CONS+GS				HC3_60 CONS+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	22	6	0,637	0,834	26	2	0,652	0,802	24	4	0,857	0,903
S	19	128	0,911	0,834	25	122	0,9	0,802	4	143	0,972	0,903

Tabela 5. 4– Resultado das classificações com modelos aplicando a seleção filtro.

Nesta abordagem, das 66 variáveis empregadas inicialmente, foram selecionadas dez: ROA_1, LC_1, LS_1, LC_2, LS_2, ROI_2, ROA_2, $ETAT_3$, $EOAT_3$, LC_3.

Das 10 variáveis selecionadas três ($ETAT_3$, $EOAT_3$, LC_3) são referentes ao ano em que as empresas foram declaradas insolventes, portanto não deveriam compor o modelo. Todas as variáveis selecionadas são originadas do Balanço Patrimonial, o que reflete a preponderância da situação patrimonial da entidade na caracterização de empresas com grandes possibilidades de se tornarem insolventes

5.1.3.2 Resultados para a seleção de atributos por ACP.

Nesta seção a técnica de seleção de atributos aplicada nas sub-bases foi ACP. Por este método as variáveis selecionadas com 75,36% de variância total (valor gerado pelo *weka*) foram: LC_1, ROI_1, LC_2, LS_2, MO_2, $TERFIN_2$, LI_3, LC_3, LG_3, MO_3. Os resultados das classificações após o balanceamento é apresentado na Tabela 5.5.

	HC4_60 ACP				HC5_60 ACP				HC6_60 ACP			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	25	3	0,476	0,850	24	4	0,857	0,936	19	9	0,404	0,870
S	52	95	0,775	0,850	4	143	0,646	0,936	47	100	0,742	0,870

Tabela 5. 5– Resultado das classificações com modelos aplicando ACP.

Das dez variáveis selecionadas pela técnica do ACP quatro pertencem ao período da declaração da insolvência (LI_3, LC_3, LG_3, MO_3). Nesta seleção há variáveis relacionadas ao desempenho operacional (MO_2 e MO_3) e, portanto tem origem no Demonstrativo de Resultado do Exercício. Entretanto, verifica-se que quando são comparados os resultados das classificações com a abordagem filtro, os resultados usando ACP são inferiores.

5.1.3.3 Resultados para a técnica de seleção de atributos *wrapper*.

Nesta seção a abordagem utilizada foi *wrapper, e* os métodos de busca foram GS e GD, sendo que o GA foi o que gerou melhores resultados. O algoritmo de aprendizado predeterminado nesta abordagem foi o AD. A Tabela 5.6 apresenta estes resultados.

	HC7_60 WRAPPER+GS				HC8_60 WRAPPER+GS				HC9_60 WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	26	2	0,963	0,978	24	4	0,857	0,936	26	2	0,963	0,978
S	0	95	0,993	0,978	4	143	0,646	0,936	0	95	0,993	0,978

Tabela 5. 6– Resultado das classificações dos modelos aplicando *wrapper*.

As variáveis selecionadas pelo método de seleção *wrapper* foram: $EOCpOT_1$, LG_1, $EOCpOT_2$, LC_2, $TERFIN_2$ e $TERFIN_3$.

Os resultados encontrados com a aplicação da abordagem de seleção de atributos *wrapper* ficaram bem melhores em relação aos das outras duas técnicas (filtro e ACP) utilizadas nas subseções anteriores. Há uma acentuada melhora tanto na classificação das empresas solventes como nas insolventes. Tais resultados corroboram as conclusões do estudo de Somol *et al.,* (2005). Vale salientar de que mesmo com uma amostra relativamente reduzida, a utilização da abordagem *wrapper* em variáveis contábeis de empresas brasileiras para previsão de insolvência foi bem mais eficaz do que as outras duas abordagens. O método de busca GS foi o que apresentou melhor resultado nas abordagens filtro e *wrapper*. A ACP realiza uma transformação linear, portanto, não utiliza um método de busca.

Três variáveis presentes nas abordagens filtro, ACP e *wrapper*, utilizando variáveis sequenciais, chamam atenção: TERFIN, LC e EOCpOT. Tais variáveis têm origem estritamente no Balanço Patrimonial, evidenciando a influência demasiada da situação patrimonial numa caracterização de insolvência, ficando para o segundo plano os aspectos operacionais, evidenciados no Demonstrativo de Resultado do Exercício.

5.1.4 Aplicação das estratégias SEID e SEIDwS

Nesta seção, a estratégia SEID desenvolvida para a predição de insolvências em empresas será aplicada na base sequencial. Na prática, a aplicação completa do SEID é obtida com o uso da votação majoritária em relação aos resultados dos modelos das sub-bases obtidas na definição da instância que está sendo avaliada. Desta forma, as sub-bases passam a representar um comitê de classificadores conforme descrito.

Deve-se ressaltar que para a geração dos classificadores utiliza-se a validação cruzada em 10 partes, tanto para as sub-bases do SEID quanto para o SMOTE. Agora, com a utilização do SEID completo a validação será feita pelo método da substituição tanto para o SEID como para o SMOTE.

Da mesma forma, serão apresentados os resultados utilizando-se da estratégia SEIDwS. Neste caso, os atributos selecionados continuam os mesmos para cada sub-base. Porém, a votação majoritária deve aumentar a robustez na predição obtida para as instâncias avaliadas. O procedimento é exposto na Figura 5.3.

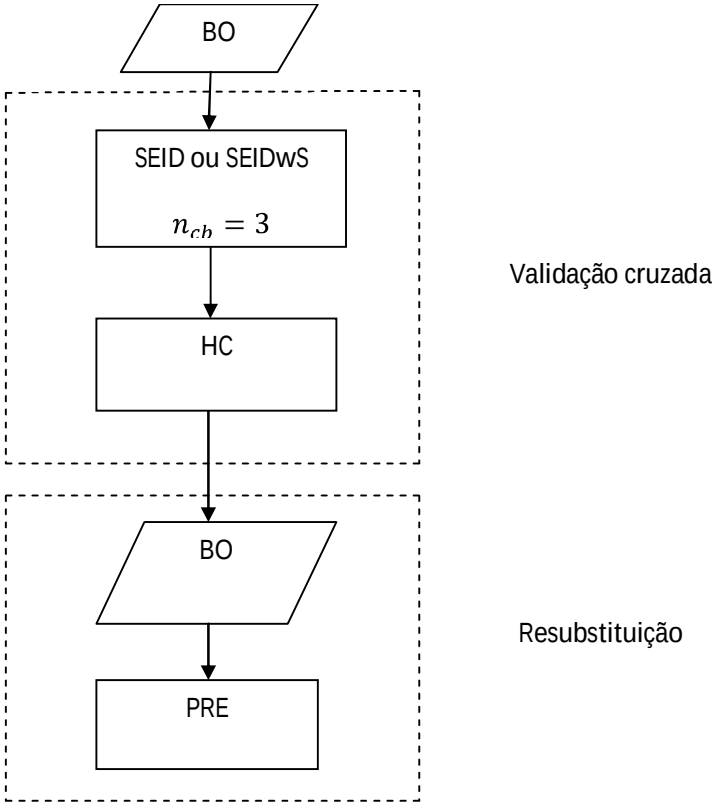

Figura 5. 3– Procedimento de classificação com o SEID ou SEIDwS.

No caso do SEIDwS, serão avaliados os melhores algoritmos de seleção determinados para as estratégias filtro e *wrapper* e também com o algoritmo ACP.

Os resultados obtidos são mostrados na tabela 5.7:

	SEIDwS CONS+GA				SEIDwS ACP				SEIDwS WRAPPER			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	26	2	0,732	0,890	26	2	0,7	0,876	26	2	0,912	0,941
S	17	130	0,932	0,890	21	126	0,916	0,876	0	147	0,982	0,941

Tabela 5.7 Resultados referentes à base de dados balanceadas aplicando SEIDwS.

A tabela 5.7 indica que a aplicação de um procedimento de seleção de características em conjunto com o SEIDwS tendeu a melhorar a previsão das empresas insolventes na base de dados utilizados.

Dos resultados acima apresentados a abordagem *wrapper* para seleção de atributos foi a que melhor resultado gerou para a classificação de empresas insolventes. Esta técnica acertou quase todas as empresas insolventes 26 de 28. Por outro lado apresentou também um excelente desempenho para as empresas solventes, bem melhor do que as outras técnicas de seleção. Nesta base de dados a técnica de seleção ACP

apresentou um desempenho inferior à base de dados em que não há aplicação de seleção de atributos para as empresas solventes acertando 126 empresas contra 147. Já a abordagem filtro empregando-se *consistency* como algoritmo de avaliação obteve resultados mais eficazes comparados com a base sem seleção de atributos e a técnica ACP, entretanto apresenta resultados inferiores a abordagem *wrapper*.

5.1.5 Comparação dos resultados encontrados

Na tabela 5.8 é feita uma comparação dos melhores resultados encontrados, através das métricas para avaliação, da base original desbalanceada com a base balanceada mais a seleção de atributos (SEIDwS).

	BASE ORIGINAL				SEIDwS WRAPPER			
Classe	MC		F	AUC	MC		F	AUC
I	20	8	0,769	0,890	26	2	0,912	0,941
S	4	143	0,959	0,890	0	147	0,982	0,941

Tabela 5. 8 Comparação dos resultados encontrados.

Pode ser visto na Tabela 5.8 a importância da aplicação de técnicas de balanceamento de bancos de dados visando melhorar a classificação da classe minoritária, neste caso a classe das empresas insolventes. Podemos evidenciar, também, a influência da técnica de seleção de atributos *wrapper* para melhorar o resultado. Os resultados encontrados aplicando esta técnica apresentaram significativas melhoras em relação às técnicas de filtro e estatística.

5.2 Avaliação da base de dados de painel

Inicialmente uma avaliação de classificadores será feita sobre a base de dados de painel, sem nenhum tipo de tratamento adicional. A base de dados de painel é composta por 140 instâncias classificadas como insolventes e 1470 instâncias classificadas como solventes. A dimensão das sub-bases foi determinada em 210 x 210, ou seja, 210 instâncias classificadas como insolventes e 210 instâncias classificadas como solventes. O número de variáveis permanece o mesmo, igual a 22.

5.2.1 Aplicação de classificadores na base de dados de painel

As técnicas empregadas para a classificação das empresas são: Regressão Logística (RL), Máquina de Vetor Suporte (SVM), *Multilayerperceptron* (MLP), e Árvore de Decisão (AD). Estes classificadores foram escolhidos por serem considerados eficientes bem como por serem largamente utilizados na determinação de insolvência de empresas. O principal objetivo é avaliar a qualidade da predição obtida com a aplicação direta do discriminante sobre uma base de dados com o perfil da base de dados de painel, ou seja, desbalanceada e com diferentes setores econômicos representados.

Foram feitos ajustes paramétricos iniciais para cada classificador utilizado, visando obter uma parametrização adequada para esta base. Os resultados apresentados são obtidos por meio de validação cruzada com 10 partes. Para que haja um melhor entendimento do desempenho de cada classificador apresentam-se os resultados de cada classificador da matriz de confusão, medida F e AUC.

Os parâmetros utilizados pelos classificadores foram os mesmos utilizados pelos dados sequenciais. Foram utilizadas as variáveis dos três grupos descritos na seção anterior.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	75	65	0,607	0,907	85	55	0,756	0,804	89	51	0,764	0,757	80	60	0,661	0,935
S	32	1438	0,932	0,757	0	1470	0,982	0,804	4	1466	0,982	0,757	22	1448	0,972	0,935

Tabela 5. 9- Resultados dos classificadores no treinamento da base de dados original

Dos classificadores testados AD e RL foram os que obtiveram os melhores resultados nos testes. O AD teve um desempenho bem melhor do que o RL. Em todos os classificadores, as empresas solventes obtiveram menores erros de classificação. Isto pode ser visto tanto nas matrizes de confusão como nas medidas F. Estes resultados contrariam o objetivo principal desta classificação que é a de descobrir melhores conhecimentos na classe de empresas insolventes.

Os resultados de todos os classificadores aqui testados, empregando dados de painel, foram melhores do que os resultados encontrados quando empregados dados sequenciais. Tal afirmativa pode ser evidenciada através de todas as métricas utilizadas com os dados de painel em relação à base de dados com dados sequenciais (Tabela 5.9 e Tabela 5.1).

5.2.2 Balanceamento e seleção de características para base de painel

Estuda-se nesta seção o desempenho do balanceamento da base de dados aplicados em conjunto com uma estratégia de seleção de características.

Assim como na base seqüencial foi feito primeiramente a seleção de atributos antes do balanceamento, os resultados encontrados não foram compatíveis para um nível mínimo aceitável em uma previsão de insolvência de empresas (valores de F e AUC próximos a 0,65). Diante disso, a etapa de seleção de atributos foi executada após a realização do balanceamento das bases de dados. Este modelo foi testado e os resultados são apresentados nas tabelas seguintes.

Possivelmente o resultado de pouca eficiência seja consequência da heterogeneidade das sub-bases em termos de setores econômicos representados.

Aplica-se então, um modelo de seleção baseado no algoritmo SEID com seleção (SEIDwS) descrito no capitulo 3. Neste modelo, a seleção é feita de forma individual para cada sub-base, permitindo variação entre os atributos selecionados para cada sub-base. A Figura 5.2 se refere ao fluxo dos procedimentos para se chegar aos resultados após os balanceamentos e a seleção de atributos da base de dados original.

Nesta subseção a classificação é realizada após a seleção das variáveis independentes. A seleção de atributos utilizou as abordagens filtro, ACP e *wrapper*.

5.2.2.1 Resultados para a abordagem de seleção de atributos filtro

Nas aplicações das abordagens filtro foram utilizados na avaliação dos subconjuntos de atributos os mesmos modelos para os dados seqüenciais, ou seja, algoritmos *CfsSubsetEval* e C*onsistency*. Para o método de busca foram utilizados GS (g*enetic search*) e GD (*greedystepwise*). O classificador utilizado foi o AD. Os melhores resultados foram encontrados aplicando *Consistency* e GS como algoritmos de avaliação e busca respectivamente, e os procedimentos executados são os já explicitados na Figura 5.2. Os resultados encontrados estão na Tabela 5.10, onde os HCs representam os modelos gerados empregando-se a base de dados original.

	HC1_210 CONS+GS				HC2_210 CONS+GS				HC3_210 CONS+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	111	29	0,485	0,896	110	30	0,427	0,801	111	29	0,437	0,809
S	207	1263	0,915	0,896	265	1265	0,891	0,801	257	1213	0,895	0,809

Tabela 5. 10- Resultado para as sub-bases utilizando seleção de atributos com abordagem filtro.

As variáveis selecionadas nesta abordagem foram: ETAT, ETPL, LS, MB e RTA, ou seja, cinco variáveis selecionadas num conjunto de vinte e duas. Destas variáveis, apenas uma, a MB, tem sua origem no Demonstrativo de Resultado do Exercício, evidenciando a importância da situação patrimonial na caracterização de empresas possivelmente insolventes.

Na abordagem filtro de seleção de atributos os resultados se mostraram inferiores aos obtidos sem a aplicação do balanceamento e da seleção de atributos. Porém, a predição das empresas insolventes, de maior interesse, obteve um considerável acréscimo no nível de predição para todas as sub-bases consideradas por serem menores do que a original, houve um decréscimo nos valores de medida F e AUC.

5.2.2.2 Resultados para a abordagem de seleção de atributos ACP

Nesta subseção é apresentada a Tabela 5.11 com os resultados após se fazer o balanceamento e a seleção de atributos com ACP.

	HC4_210 ACP				HC5_210 ACP				HC6_210 ACP			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	90	50	0,246	0,765	98	42	0,287	0,715	110	30	0,275	0,755
S	501	969	0,779	0,765	445	1025	0,808	0,715	550	920	0,76	0,755

Tabela 5. 11– Resultado para as sub-bases utilizando seleção de atributos abordagem ACP.

As variáveis selecionadas por ACP foram: ROE, ROA, LC, LS, EOAT. Todas as variáveis selecionadas pela técnica ACP tem sua origem no Balanço Patrimonial. A variância total encontrada foi de 80%.

5.2.2.3 Resultados para a abordagem de seleção de atributos *wrapper*

Nesta subseção, ocorre o balanceamento para depois ser realizada a seleção de atributos com *wrapper*. AD foi também o classificador utilizado nesta sub-seção. Os resultados encontrados estão na Tabela 5.12.

	HC7_210 WRAPPER+GS				HC8_210 WRAPPER+GS				HC9_210 WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	121	19	0,595	0,916	120	20	0,71	0,902	124	16	0,933	0,943
S	146	1324	0,941	0,916	78	1392	0,966	0,902	3	1467	0,944	0,943

Tabela 5. 12– Resultado para as sub-bases utilizando seleção de atributos abordagem *wrapper*.

Os resultados evidenciam que o balanceamento seguido da seleção de atributos na abordagem *wrapper* gera um ganho de desempenho em relação a classificação após o balanceamento e sem a aplicação de técnicas de seleção de atributos (Tabela 5.9). Entretanto, vale evidenciar de que o tempo despendido para executar a tarefa de seleção de atributos com a abordagem *wrapper* foi bem maior do que nas outras abordagens (filtro e ACP). Das 22 variáveis totais dez foram selecionadas pela abordagem *wrapper*. As variáveis selecionadas foram: EOCpOT, EOAT, GA, IMCP, LS, MB, EBITDA, ML, MO e TERFIN. Esta abordagem selecionou o dobro de variáveis em relação as outras duas abordagens empregadas (filtro e ACP), dessas dez variáveis seis tem origem somente no Balanço Patrimonial (EOCpOT, EOAT, IMCP, LS, EBITDA e TERFIN), três tem sua origem no Demonstrativo de Resultado do Exercício (MB, ML e MO) e uma tem sua origem tanto no Balanço Patrimonial quanto no Demonstrativo de Resultado do Exercício (GA). Esta abordagem foi a que apresentou os melhores resultados, sendo que há mais, entre as variáveis selecionadas, aquelas com origem nos dois demonstrativos contábeis, Balanço Patrimonial e Demonstrativo de Resultado do Exercício. Isto pode ser entendido que os desempenhos operacionais (MB, ML, MO e GA) exercem influência relevante na caracterização das empresas que podem a se tornar insolventes.

5.2.3 Aplicação das estratégias SEID e SEIDwS

Nesta seção a estratégia SEID desenvolvida para a predição de insolvências em empresas será aplicada na base de dados de painel. Na pratica, a aplicação completa do

SEID é obtida com o uso da votação majoritária em relação aos resultados dos modelos das sub-bases obtidas na definição da instância que está sendo avaliada. Desta forma, as sub-bases passam a representar um comitê de classificadores conforme descrito anteriormente.

Deve-se ressaltar que para a geração dos classificadores utiliza-se a validação cruzada em 10 partes, tanto para as sub-bases do SEID quanto para o SMOTE. Agora, com a utilização do SEID completo a validação será feita pelo método da substituição tanto para o SEID como para o SMOTE.

Da mesma forma, serão apresentados os resultados utilizando-se da estratégia SEIDwS. Neste caso, os atributos selecionados continuam os mesmos para cada sub-base. Porém, a votação majoritária deve aumentar a robustez na predição obtida para as instâncias avaliadas. O procedimento foi exposto na Figura 5.3.

No caso do SEIDwS, serão avaliados os melhores algoritmos de seleção determinados para as estratégias filtro e *wrapper* e também com o algoritmo ACP. Os resultados obtidos são mostrados na Tabela 5.13.

	SEIDwS_CONS+GS				SEIDwS_ACP				SEIDwS_WRAPPER			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	111	29	0,443	0,815	98	42	0,289	0,742	125	15	0,932	0,943
S	250	1220	0,897	0,815	440	1030	0,81	0,742	3	1467	0,993	0,943

Tabela 5. 13 Resultados referentes à base de dados balanceadas aplicando SEIDwS..

Pela tabela 5.13 se pode evidenciar, também, a influência da técnica de seleção de atributos *wrapper* para melhorar o resultado. Os resultados encontrados aplicando esta técnica apresentaram significativas melhoras em relação às técnicas de filtro e estatística.

5.2.4 Comparação dos resultados encontrados

Na tabela 5.14 é feita uma comparação dos melhores resultados encontrados, com SEID e o SEIDwS utilizando modelo *wrapper*.

	BASE ORIGINAL				SEIDwS_WRAPPER			
Classe	MC		F	AUC	MC		F	AUC
I	80	60	0,661	0,935	125	15	0,939	0,944
S	22	1448	0,972	0,935	3	1467	0,945	0,944

Tabela 5. 14– Comparação dos resultados.

Nos resultados encontrados na Tabela 5.14 as bases de dados de painel que obtiveram os melhores resultados na predição foram os modelos sujeitos ao balanceamento, ou seja, o SEIDwS, com vantagem para o modelo que utiliza a seleção de características. Destaca-se, principalmente nos modelos balanceados um ganho de eficácia na classificação na classe das insolventes, atendendo o interesse preponderante desta técnica.

5.2.6 Considerações sobre os resultados encontrados com as variáveis sequenciais e as de painel

Pelos resultados encontrados as variáveis seqüenciais se mostraram pouco mais eficientes do que as variáveis de painel na modelagem de previsão de insolvência de empresas brasileiras de capital aberto. As medidas F e AUC encontradas nos modelos elaborados com variáveis sequenciais foram melhores do que as mesmas medidas de avaliação encontradas com as variáveis de painel. Este fato indica que as variáveis sequenciais podem ser mais adequadas para modelagem em previsão de insolvência.

Já em relação à abordagem de seleção de atributos, o método *wrapper* foi mais eficiente nos dois tipos de variáveis. O mesmo aconteceu com emprego do classificador AD.

As variáveis selecionadas usando-se a base com dados sequenciais preponderam àquelas pertencentes ao grupo de liquidez. Provavelmente, isto ocorre devido ao pequeno período da composição da base de dados, caracterizando assim a maior debilidade financeira dos períodos finais de seu comprometimento de continuidade, prevalecendo, assim o diagnóstico daquela situação do que as causas.

Nas variáveis selecionadas usando-se a base com dados de painel, já aparecem aquelas que representam desempenhos operacionais das entidades, evidenciando motivos operacionais na descontinuidade da empresa. Talvez devido à extensão temporal das variáveis, essas variáveis passam a ser também representativas para previsão de insolvência.

Não se pode afirmar qual das duas abordagens pode compor melhor os modelos de previsão, pode-se afirmar que elas compõem bons modelos de previsão de insolvência de empresas brasileiras.

CAPÍTULO 6 Análise dos dados por setor econômico

Neste capítulo são apresentadas as análises estatísticas descritivas de dados das empresas de mesmo setor econômico. A seguir, descrevem-se os seis setores econômicos estudados nesta tese de acordo com a classificação do Bovespa no ano de 2007: (i) materiais básicos; (ii) consumo cíclico; (iii) consumo não cíclico; (iv) bens industriais; (v) construção e transportes e; (vi) tecnologia da informação e telecomunicações. Os motivos de serem estudados estes setores dizem respeito à facilidade no acesso às informações contábeis de empresas solventes e insolventes; capacidade de representar bem o ambiente econômico das empresas brasileiras de capital aberto durante o período de tempo estudado e a possibilidade de se fazer este estudo devido à qualidade e a quantidade de dados contábeis dessas empresas.

Vale aqui repetir a importância e os principais motivos da segmentação das empresas por setor econômico e o seu estudo. Para Iudícibus (2008, p. 91) empresas de mesmo setor econômico apresentam em seus demonstrativos contábeis semelhanças devido a suas estruturas patrimoniais e econômicas. Indicadores como de liquidez, endividamento e rentabilidade, por exemplo, devem apresentar valores bem próximos na sua média setorial. Empresas que apresentam índices bem distintos ao da média setorial no qual pertencem, devem apresentar situações com certas anomalias econômicas ou financeiras principalmente, portanto, em condições normais os seus indicadores também devem apresentar comportamentos ajustados. Silva (2006, p. 190) afirma que é possível comparar o índice financeiro de uma empresa com o mesmo índice relativo a outras empresas de mesma atividade econômica, para sabermos como está a empresa em relação as suas principais concorrentes ou mesmo em relação aos padrões do seu segmento de atuação. Para Caouette (1998, p. 129), o analista de crédito compara diversos índices contábeis do tomador em potencial com normas e tendências setoriais ou grupais pertinentes a estas variáveis. Portanto, com a segmentação econômica das empresas, tenta-se minorar erros cometidos nas classificações devido à presença de empresas com características patrimoniais e financeiras distintas num mesmo grupo.

6.1 Classificação das empresas por setor econômico

Apresenta-se agora uma classificação das empresas que compõem a base de dados em relação ao setor econômico a que pertencem. Pretende-se descrever tais setores de acordo com as características especificas que seus atributos tendem a assumir, tanto para empresas solventes quanto insolventes. Deve-se destacar que existem outros setores que não estão representados na base de dados utilizada.

6.1.1 Empresas do setor econômico de materiais básicos

Compõem este setor econômico empresas dos subsetores de embalagem, madeira e papel, materiais diversos, mineração, químicos, siderurgia e metalurgia. Neste setor econômico, as empresas apresentam, normalmente, valores proporcionalmente altos em seus ativos permanentes, sendo que em empresas, sobretudo dos setores de madeiras e papel, químicos, siderurgia e metalurgia, seus ativos permanentes são substanciados pelos ativos imobilizados (instalações, equipamentos, máquinas, etc).

Neste grupo de empresas há 30 instâncias representando as insolventes e 351 instâncias representando as empresas solventes, também deste setor. Nas Tabelas 6.1 e 6.2 são apresentadas as análises dos dados contábeis das empresas solventes e insolventes, respectivamente do setor econômico de materiais básicos.

Empresa Solventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOCpOT	0,982	0,000	0,288	0,423	0,397	67,99%
EOPL	1,317	-0,950	0,390	0,253	0,146	154,58%
EOAT	2,599	0,000	0,348	0,983	0,250	35,44%
ETAT	6,097	0,130	0,962	2,396	0,487	40,16%
ETPL	6,590	-3,707	1,383	1,422	0,739	97,29%
GA	2,784	0,078	0,650	1,171	0,559	55,49%
GAF	12,756	-6,521	2,133	2,790	1,711	76,48%
IMCP	3,872	-2,051	0,906	0,909	1,057	99,71%
LI	3,902	0,000	0,676	0,015	0,013	4430,97%
LS	5,144	0,022	0,922	0,020	0,094	4540,97%
LC	6,257	0,064	1,189	0,025	0,012	4749,82%
LG	6,141	0,058	1,136	0,024	0,072	4644,88%
TERFIN	2,954	-1,896	0,601	0,006	0,011	10871,57%
EBIT	3,032	-1,003	0,416	0,440	0,132	94,58%
EBTIDA	0,620	-1,003	0,227	0,510	0,150	44,57%
MB	2,591	-1,526	0,450	0,505	0,093	89,19%
ML	2,673	-1,607	0,471	0,512	0,079	91,90%
MO	2,591	-1,526	0,467	0,511	0,086	91,52%
ROA	0,295	-2,065	0,283	0,134	0,061	211,32%
ROI	0,508	-1,200	0,399	0,320	0,106	124,54%
ROE	0,344	-2,111	0,288	0,345	0,053	83,58%
RTA	0,295	-2,065	0,283	0,412	0,061	68,73%

Tabela 6. 1 Dados contábeis das empresas solventes do setor econômico materiais básicos

Empresas Insolventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOCpOT	3,884	0,476	1,088	1,596	1,268	68,15%
EOPL	3,605	-0,484	1,324	0,852	0,366	155,49%
EOAT	1,589	0,040	0,541	0,466	0,199	116,24%
ETAT	4,278	0,041	1,457	1,113	0,630	130,89%
ETPL	4,743	-1,307	1,854	1,297	1,129	142,93%
GA	1,478	0,160	0,408	0,896	0,979	45,59%
GAF	6,833	-24,053	10,444	-6,017	-1,385	173,57%
IMCP	19,083	-9,200	7,983	1,349	0,560	591,57%
LI	0,168	0,000	0,053	0,000	0,002	13503,53%
LS	0,773	0,020	0,279	0,004	0,037	7368,64%
LC	1,032	0,040	0,364	0,005	0,061	6781,63%
LG	0,680	0,080	0,251	0,004	0,031	7052,78%
TERFIN	0,302	-0,543	0,259	-0,001	-0,003	50151,99%
EBIT	0,339	-0,676	0,365	-0,268	-0,286	-136,13%
EBTIDA	1,074	-0,108	0,412	0,199	-0,014	207,19%
MB	0,462	0,064	0,139	0,263	0,239	52,89%
ML	0,493	-0,415	0,287	-0,070	-0,047	409,39%
MO	0,476	-0,235	0,226	-0,012	-0,064	1832,25%
ROA	0,468	-0,366	0,268	-0,043	-0,066	622,65%
ROI	0,813	-1,521	0,790	-0,423	-0,447	186,87%
ROE	0,538	-0,166	0,227	0,019	-0,041	1172,22%
RTA	0,495	-0,365	0,268	-0,036	-0,054	749,21%

Tabela 6. 2-Dados contábeis das empresas insolventes do setor econômico materiais básicos

Pelas Tabelas 6.1 e 6.2, verifica-se que há uma distinção de valores entre as empresas pertencentes à amostra das solventes em relação às empresas pertencentes à amostra das insolventes. Empresas insolventes apresentam valores de coeficiente de variação definido como o desvio padrão dividido pelas médias (CV) bem maiores do que os valores das empresas classificadas como solventes, caracterizando como altos os níveis de incertezas entre essas empresas, conseqüências de provável escassez de controles e planejamentos. Empresas sadias econômicas e financeiramente desse setor se caracterizam por valores medianos em sua liquidez, mas com valores não tão altos nas suas rentabilidades, sendo compensados por valores mais substanciais em suas margens de lucros (MB, ML e MO).

6.1.2 Empresas do setor econômico de consumo cíclico

Neste setor econômico estão incluídas empresas dos subsetores de comércio, hotéis, restaurantes, lazer, mídia, tecidos, vestuários, calçados e utilidades domésticas. Neste setor econômico, as empresas se caracterizam com valores proporcionalmente baixos em seus ativos permanentes e valores mais substanciais em seus ativos

circulantes e com ausências de valores significativos em seus ativos diferidos. Compõem este setor econômico 50 instâncias das empresas insolventes e 474 instâncias das empresas solventes. Nas Tabelas 6.3 e 6.4 são apresentados respectivamente, os dados contábeis das empresas solventes e insolventes, do setor econômico de consumo cíclico.

Empresa Solventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOcpOT	4,300	0,009	1,175	1,326	1,001	88,62%
EOPL	4,825	-0,584	1,431	1,010	0,602	141,67%
EOAT	0,717	0,008	0,238	0,272	0,339	87,44%
ETAT	20,876	0,003	2,890	0,996	0,595	290,29%
ETPL	6,148	-1,050	1,991	1,737	1,201	114,58%
GA	2,068	0,007	0,800	0,958	0,961	83,47%
GAF	9,273	-10,430	5,394	0,664	2,273	811,84%
IMCP	4,799	-6,991	2,919	0,416	1,189	701,05%
LI	0,623	0,006	0,138	0,093	0,031	148,54%
LS	2,413	0,004	0,656	0,739	0,529	88,86%
LC	2,422	0,004	0,656	0,925	0,890	70,94%
LG	11,940	0,007	3,099	1,671	0,686	185,48%
TERFIN	0,981	-1,056	0,402	0,072	0,010	559,30%
EBIT	0,374	-1,145	0,395	0,208	-0,182	189,50%
EBTIDA	4,068	-0,190	1,134	0,522	0,105	217,24%
MB	0,873	-0,365	0,264	0,351	0,302	75,39%
ML	0,216	-1,332	0,471	0,275	-0,064	171,52%
MO	0,184	-1,328	0,504	0,306	-0,090	164,42%
ROA	0,183	-1,199	0,261	0,101	-0,002	258,65%
ROI	0,376	-1,615	0,612	0,335	-0,011	182,85%
ROE	0,236	-1,198	0,230	0,059	-0,002	392,23%
RTA	0,209	-1,199	0,221	0,060	0,003	366,92%

Tabela 6. 3- Dados contábeis das empresas solventes do setor econômico de bens de consumo cíclico

	Empresas Insolventes					
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOcpOT	3,884	0,476	1,088	1,596	1,268	68,15%
EOPL	3,605	-0,484	1,324	0,852	0,366	155,49%
EOAT	1,589	0,040	0,541	0,466	0,199	116,24%
ETAT	4,278	0,041	1,457	1,113	0,630	130,89%
ETPL	4,743	-1,307	1,854	1,297	1,129	142,93%
GA	1,478	0,160	0,408	0,896	0,979	45,59%
GAF	6,833	-24,053	10,444	-6,017	-1,385	173,57%
IMCP	19,083	-9,200	7,983	1,349	0,560	591,57%
LI	0,168	0,002	0,053	0,039	0,021	135,04%
LS	0,773	0,020	0,279	0,379	0,371	73,69%
LC	1,032	0,040	0,364	0,537	0,606	67,82%
LG	0,680	0,080	0,251	0,356	0,305	70,53%
TERFIN	0,302	-0,543	0,259	-0,052	-0,032	501,52%
EBIT	0,339	-0,676	0,365	-0,268	-0,286	136,13%
EBTIDA	1,074	-0,108	0,412	0,199	-0,014	207,19%
MB	0,462	0,064	0,139	0,263	0,239	52,89%
ML	0,493	-0,415	0,287	-0,070	-0,047	409,39%
MO	0,476	-0,235	0,226	-0,012	-0,064	1832,25%
ROA	0,468	-0,366	0,268	-0,043	-0,066	622,65%
ROI	0,813	-1,521	0,790	-0,423	-0,447	186,87%
ROE	0,538	-0,166	0,227	0,019	-0,041	1172,22%
RTA	0,495	-0,365	0,268	-0,036	-0,054	749,21%

Tabela 6. 4- Dados contábeis das empresas insolventes do setor econômico de bens de consumo cíclico.

Neste setor econômico, os valores contábeis das empresas solventes e insolventes não apresentam distorções tão acentuadas como no grupo de empresas do setor econômico de materiais básico, mas mesmo assim os valores das empresas classificadas como insolventes podem ser considerados como deteriorados na liquidez, na rentabilidade, no endividamento e nas margens. Os valores dos coeficientes de variação não podem ser considerados com resultados tão dispares. Este é um setor que demanda uma liquidez relativamente baixa junto com suas rentabilidades e margens.

6.1.3 Empresas do setor econômico de consumo não cíclico

Neste setor econômico, estão incluídos empresas dos subsetores de agropecuária, alimentos processados, bebidas, comércio e distribuição, fumo, produtos de uso pessoal e de limpeza, e saúde. Neste setor econômico, as empresas se caracterizam com valores proporcionalmente baixos em seus ativos permanentes e valores mais substanciais em seus ativos circulantes e com baixos níveis de endividamento.

Compõem este setor 25 instâncias de empresas insolventes e 170 instâncias de empresas solventes. Nas Tabelas 6.5 e 6.6 são apresentadas respectivamente, as análises dos dados contábeis das empresas solventes e insolventes do setor econômico de consumo não cíclico.

Empresa Solventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOcpOT	0,008	1,009	0,321	0,387	0,389	83%
EOPL	7,552	5,665	1,166	0,332	0,302	351%
EOAT	0,000	15,522	2,410	0,769	0,245	313%
ETAT	0,081	18,834	2,826	1,153	0,539	245%
ETPL	4,020	7,253	1,525	1,159	0,985	132%
GA	0,008	2,200	0,546	0,817	0,849	67%
GAF	9,719	6,770	1,874	1,553	1,487	121%
IMCP	8,610	8,003	1,323	1,015	1,057	130%
LI	0,009	2,398	0,351	0,002	0,004	17577%
LS	0,003	6,104	0,844	0,010	0,090	8368%
LC	0,004	7,215	1,109	0,014	0,121	8010%
LG	0,014	7,524	1,202	0,011	0,087	10494%
TERFIN	1,087	0,786	0,662	0,002	0,004	35626%
EBIT	6,011	-2,049	0,839	0,070	0,051	1202%
EBTIDA	4,444	-0,513	0,654	0,070	0,076	939%
MB	9,369	-4,110	1,454	0,200	0,029	727%
ML	8,234	-9,253	1,470	0,111	0,027	1320%
MO	7,537	-4,110	1,339	0,203	0,028	658%
ROA	3,721	-0,354	0,580	0,104	0,025	558%
ROI	1,009	-0,759	0,357	0,017	0,044	2063%
ROE	3,722	-0,442	0,581	0,098	0,029	591%
RTA	3,243	-0,354	0,451	0,060	0,020	756%

Tabela 6. 5- Dados contábeis das empresas solventes do setor econômico de bens de consumo não-cíclico

.

Empresas Insolventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOcpOT	1,221	-4,362	2,112	-0,720	0,279	293,43%
EOPL	4,403	-0,917	1,726	0,473	-0,173	365,07%
EOAT	14,856	0,099	2,995	1,952	1,179	153,43%
ETAT	17,209	0,378	3,219	2,208	1,418	145,81%
ETPL	4,414	-8,255	3,683	-0,858	-1,026	429,11%
GA	1,185	0,001	0,458	0,475	0,262	96,53%
GAF	9,060	-1,913	3,617	1,676	0,308	215,79%
IMCP	1,425	-0,572	0,792	0,445	-0,015	178,05%
LI	0,162	0,000	0,056	0,052	0,027	106,70%
LS	0,990	0,002	0,374	0,436	0,400	85,73%
LC	1,439	0,002	0,521	0,590	0,531	88,36%
LG	0,992	0,002	0,340	0,356	0,350	95,56%
TERFIN	0,450	-1,303	0,506	-0,310	-0,047	163,25%
EBIT	0,327	-12,621	4,556	-1,910	-0,091	238,47%
EBTIDA	0,351	-8,208	2,990	-1,174	0,103	254,82%
MB	0,604	-7,537	1,498	-0,141	0,133	1063,51%
ML	0,221	-7,536	1,907	-1,101	-0,131	173,23%
MO	0,244	-7,537	2,504	-1,492	-0,139	167,82%
ROA	0,058	-4,180	1,508	-0,964	-0,141	156,51%
ROI	6,794	-1,230	2,694	0,500	-0,785	538,51%
ROE	0,075	-4,216	1,479	-1,058	-0,497	139,80%
RTA	0,058	-3,243	0,669	-0,354	-0,141	188,92%

Tabela 6. 6- Dados contábeis das empresas insolventes do setor econômico de bens de consumo não-cíclico

Os resultados das empresas deste setor econômico apresentaram algumas semelhanças com as empresas do setor econômico cíclico, liquidez baixa rentabilidade baixa, endividamento baixo e margens também baixam. As medianas encontradas para as empresas insolventes sofrem uma variação bem mais acentuadas do que as das empresas solventes e os coeficientes de variações não apresentaram distorções substanciais entre os grupos de empresas.

6.1.4 Empresas do setor econômico de bens industriais

Neste setor econômico estão incluídos empresas dos subsetores de comércio, equipamentos elétricos, máquinas e equipamentos, material de transporte e serviços. Neste setor econômico, as empresas se caracterizam com valores proporcionalmente altos em seus ativos permanentes e valores pouco substanciais em seus ativos circulantes e com níveis razoáveis de endividamento. Nas Tabelas 6.7 e 6.8 são apresentadas respectivamente, as análises dos dados contábeis das empresas solventes e insolventes do setor econômico de bens industriais.

Empresas Solventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOCpOT	0,979	0,009	0,290	0,457	0,455	63,47%
EOPL	6,998	-4,065	1,053	0,399	0,186	263,87%
EOAT	7,105	0,009	0,730	0,397	0,280	184,17%
ETAT	2,771	0,047	0,521	0,730	0,565	71,39%
ETPL	33,080	-9,001	4,463	1,357	0,864	328,95%
GA	2,772	0,001	0,508	0,795	0,764	63,93%
GAF	29,203	-4,244	4,323	3,125	1,659	138,34%
IMCP	16,976	-7,474	2,559	1,133	0,786	225,92%
LI	1,454	0,000	0,229	0,146	0,049	156,75%
LS	3,193	0,005	0,627	0,844	0,749	74,26%
LC	4,592	0,008	0,945	1,285	1,089	73,55%
LG	3,557	0,001	0,743	0,959	0,798	77,41%
TERFIN	2,347	-3,246	0,588	-0,238	-0,133	-246,89%
EBIT	3,693	-7,043	0,890	0,019	0,095	4757,95%
EBTIDA	31,315	-4,568	2,279	0,230	0,125	990,02%
MB	9,150	-7,433	1,099	0,104	0,006	1057,96%
ML	2,537	-8,091	0,971	0,195	0,014	497,18%
MO	2,787	-7,433	0,938	0,179	0,023	524,35%
ROA	0,241	-7,900	1,023	0,207	0,016	493,80%
ROI	1,505	-0,998	0,493	0,154	0,029	319,35%
ROE	0,305	-10,877	1,269	0,238	0,019	534,00%
RTA	0,241	-1,641	0,213	0,030	0,016	714,90%

Tabela 6. 7- Dados contábeis das empresas solventes do setor econômico de bens industriais

Empresas Insolventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOCpOT	3,325	2,504	0,923	0,141	0,455	656,10%
EOPL	2,297	0,425	5,631	-0,525	0,399	-1072,91%
EOAT	0,160	0,110	5,663	-0,198	0,397	-2857,74%
ETAT	0,954	0,741	0,976	0,724	0,565	134,76%
ETPL	14,523	2,858	13,727	4,019	1,357	341,52%
GA	0,748	0,190	3,016	-0,178	0,639	-1692,11%
GAF	14,339	-9,336	10,759	5,951	3,125	180,80%
IMCP	10,695	1,332	10,787	5,732	2,259	188,18%
LI	0,068	0,000	6,030	2,839	0,229	212,37%
LS	0,683	0,050	5,971	3,251	0,749	183,65%
LC	0,842	0,100	14,518	-4,017	0,945	-361,43%
LG	0,200	0,090	11,500	-3,187	0,774	-360,83%
TERFIN	0,224	0,129	7,966	2,402	-0,133	331,68%
EBIT	1,037	-0,173	26,332	0,346	0,095	7620,74%
EBTIDA	0,134	-0,920	20,434	0,265	0,230	7719,87%
MB	-0,384	-0,583	13,375	-2,252	0,104	-593,90%
ML	-0,418	-0,477	6,432	-1,794	0,195	-358,46%
MO	-0,099	-0,333	7,141	-1,973	0,179	-361,89%
ROA	-0,389	-5,168	5,004	-1,467	0,207	-341,07%
ROI	-0,019	-0,348	1,331	0,777	0,493	171,27%
ROE	-0,278	-0,421	6,375	0,906	0,305	703,72%
RTA	-0,071	-0,333	3,691	0,104	0,030	3563,91%

Tabela 6. 8- Dados contábeis das empresas insolventes do setor econômico de bens indústria

Pode ser visto nas Tabelas 6.7 e 6.8 que o coeficiente de variação da maioria das variáveis apresenta valores bem distantes entre as empresas solventes e insolventes. Empresas deste setor econômico, quando entram num processo de deterioração tem essa etapa bem acentuada e acelerada.

Compõem a base de dados deste setor 10 instâncias caracterizando empresas insolventes e 202 instâncias caracterizando empresas solventes.

6.1.5 Empresas do setor econômico de construções e transportes

Neste setor econômico estão incluídas empresas dos subsetores de construção e engenharia e transportes. Neste setor econômico, as empresas se caracterizam com valores proporcionalmente altos em seus ativos permanentes e valores também bem substanciais em seus ativos circulantes e com níveis razoáveis de endividamento.

Compõem a base de dados deste setor 15 instâncias caracterizando empresas insolventes e 230 instâncias caracterizando empresas solventes. Nas Tabelas 6.9 e 6.10 são apresentadas a análise dos dados contábeis das empresas solventes e insolventes, respectivamente do setor econômico de construções e transportes.

Empresa Solventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOCpOT	1,000	0,000	0,330	0,345	0,267	95,56%
EOPL	9,443	-4,117	1,063	0,387	0,169	274,40%
EOAT	0,778	0,000	0,182	0,218	0,173	83,65%
ETAT	2,212	0,084	0,316	0,569	0,559	55,49%
ETPL	41,199	-9,785	4,632	1,792	0,928	258,51%
GA	1,897	0,000	0,455	0,469	0,330	97,01%
GAF	78,860	-16,385	7,250	3,685	2,119	196,76%
IMCP	45,467	-45,840	7,099	0,948	0,840	749,16%
LI	3,282	0,000	0,486	0,045	0,041	1076,17%
LS	6,433	0,005	1,353	0,139	0,097	974,19%
LC	7,099	0,006	1,462	0,158	0,119	924,41%
LG	6,553	0,036	1,216	0,144	0,108	846,26%
TERFIN	9,234	-8,360	1,126	0,008	0,004	14632,45%
EBIT	6,787	-6,023	1,168	0,042	0,044	2804,44%
EBTIDA	11,569	-7,208	1,794	0,194	0,022	923,88%
MB	6,861	-7,034	1,151	0,067	0,012	1728,72%
ML	9,866	-8,057	1,632	0,117	0,013	1392,52%
MO	6,861	-10,804	1,568	0,213	0,012	736,54%
ROA	0,281	-1,771	0,156	0,007	0,007	2215,67%
ROI	2,999	-1,000	0,426	0,084	0,013	510,04%
ROE	0,214	-1,767	0,156	0,011	0,008	1404,40%
RTA	0,281	-0,506	0,097	0,003	0,005	3189,92%

Tabela 6. 9 - Dados contábeis das empresas solventes do setor econômico de construção e transportes

Empresas Insolventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOCpOT	1,664	0,310	0,590	1,097	1,317	53,81%
EOPL	1,199	0,477	0,321	0,914	1,066	35,09%
EOAT	0,733	0,035	0,292	0,355	0,297	82,33%
ETAT	1,015	0,382	0,259	0,690	0,675	37,53%
ETPL	2,078	0,618	0,624	1,443	1,633	43,25%
GA	0,298	0,004	0,133	0,194	0,280	68,42%
GAF	1,033	-3,059	1,784	-1,327	-1,956	-134,40%
IMCP	3,734	-1,710	2,316	1,227	1,657	188,78%
LI	0,028	0,000	0,008	0,008	0,002	109,44%
LS	0,493	0,170	0,129	0,309	0,272	41,84%
LC	0,322	0,012	0,137	0,204	0,279	66,89%
LG	0,690	0,020	0,303	0,435	0,592	69,58%
TERFIN	0,021	-0,571	0,243	-0,257	-0,219	-94,30%
EBIT	6,550	-0,906	3,432	1,872	-0,021	183,32%
EBTIDA	7,842	-0,903	3,843	2,703	1,165	142,18%
MB	0,438	0,076	0,152	0,290	0,355	52,28%
ML	0,093	-0,038	0,038	0,009	0,003	439,68%
MO	-0,012	-0,996	0,459	-0,358	-0,058	-128,06%
ROA	0,020	-0,040	0,021	-0,007	0,000	-287,30%
ROI	1,027	-0,031	0,492	0,327	0,000	150,46%
ROE	0,062	-0,016	0,028	0,004	-0,007	775,24%
RTA	0,046	-0,010	0,017	0,013	0,010	129,81%

Tabela 6. 10- Dados contábeis empresas insolventes do setor econômico de construção e transportes

Para esse setor os valores que ficaram mais diferenciados foram os das medianas. As empresas insolventes apresentam medianas bem superiores ou bem inferiores aos das empresas solventes evidenciando as grandes variações ocorridas nas empresas insolventes.

6.1.6 Empresas do setor econômico de tecnologia da informação e telecomunicações

Neste setor econômico estão incluídas empresas dos subsetores de computadores e equipamentos, programas e serviços, telefonia fixa e móvel. Neste setor econômico, as empresas se caracterizam com valores proporcionalmente altos em seus ativos permanentes, sobretudo nos ativos diferidos e valores também bem substanciais em seus ativos circulantes e com altos níveis de endividamento.

Compõem a base de dados deste setor 10 instâncias caracterizando empresas insolventes e 43 instâncias caracterizando empresas solventes. Nas Tabelas 6.11 e 6.12 são apresentadas as análises dos dados contábeis das empresas solventes e insolventes, respectivamente do setor econômico de tecnologia da informação e telecomunicações.

Empresa Solventes						
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOcpOT	0,942	0,008	0,248	0,345	0,316	71,88%
EOPL	2,070	0,021	0,640	0,956	0,865	66,88%
EOAT	0,670	0,017	0,179	0,367	0,364	48,70%
ETAT	0,812	0,158	0,169	0,531	0,609	31,89%
ETPL	4,324	0,188	0,797	1,388	1,556	57,41%
GA	1,536	0,009	0,439	0,532	0,469	82,68%
GAF	10,346	-0,424	1,538	2,163	1,915	71,10%
IMCP	2,961	0,552	0,475	1,428	1,313	33,29%
LI	6,273	0,007	0,923	0,441	0,290	209,34%
LS	3,345	0,080	0,721	1,012	0,897	71,19%
LC	3,512	0,080	0,754	1,144	1,115	65,92%
LG	2,177	0,186	0,353	0,746	0,672	47,34%
TERFIN	4,566	-1,236	0,915	0,093	-0,007	981,45%
EBIT	0,347	-0,969	0,233	0,010	0,041	2246,77%
EBTIDA	0,624	-0,180	0,241	0,144	0,074	166,82%
MB	1,784	-7,603	1,804	-0,543	0,008	-332,45%
ML	0,202	-2,128	0,572	-0,172	0,013	-332,97%
MO	0,307	-2,234	0,622	-0,177	0,021	-352,50%
ROA	0,082	-0,235	0,066	-0,003	0,012	-2362,46%
ROI	0,216	-0,848	0,200	-0,024	0,031	-849,85%
ROE	0,104	-0,199	0,065	0,001	0,017	4471,31%
RTA	0,082	-0,235	0,060	0,000	0,013	24257,92%

Tabela 6. 11- Dados contábeis das empresas solventes do setor econômico de tecnologia da informação e telecomunicações

	Empresas Insolventes					
Atributos	Máximo	Mínimo	DP	Média	Mediana	CV
EOcpOT	-0,016	-0,241	0,116	-0,126	-0,123	-91,97%
EOPL	-0,021	-1,012	0,520	-0,514	-0,511	-101,09%
EOAT	1,851	0,350	0,784	1,099	1,101	71,34%
ETAT	0,475	0,280	0,092	0,373	0,377	24,64%
ETPL	-0,830	-1,541	0,367	-1,186	-1,186	-30,91%
GA	0,598	0,440	0,074	0,519	0,519	14,29%
GAF	0,160	-0,173	0,101	-0,126	-0,154	-80,27%
IMCP	1,055	-1,830	0,888	-1,231	-1,442	-72,10%
LI	0,038	0,001	0,013	0,014	0,019	88,03%
LS	0,973	0,070	0,453	0,430	0,094	105,38%
LC	0,022	0,010	0,005	0,015	0,012	31,91%
LG	0,260	0,020	0,119	0,117	0,030	101,76%
TERFIN	0,063	-0,833	0,461	-0,388	-0,385	-118,65%
EBIT	-0,228	-1,829	0,828	-1,030	-1,028	-80,38%
EBTIDA	0,257	-1,327	0,741	-0,626	-0,751	-118,30%
MB	0,469	0,168	0,145	0,315	0,318	46,01%
ML	-0,949	-3,012	1,049	-1,998	-1,981	-52,54%
MO	-0,979	-3,012	1,049	-2,005	-1,995	-52,33%
ROA	-0,449	-0,595	0,070	-0,522	-0,522	-13,42%
ROI	0,373	-0,930	0,700	-0,319	-0,314	-219,29%
ROE	-0,525	-0,988	0,215	-0,781	-0,763	-27,56%
RTA	-0,552	-1,325	0,387	-0,944	-0,939	-40,93%

Tabela 6.12- Dados contábeis das empresas insolventes do setor econômico de tecnologia da informação e telecomunicações

Nas Tabelas 6.11 e 6.12 podem ser constatadas as diferenças entre os atributos das empresas solventes em relação às empresas insolventes e essas diferenças são maiores na coluna das medianas. Também neste setor as medianas são mais distorcidas para as empresas insolventes.

6.2 Análise de classificadores com bases de dados original por setor econômico.

Apresenta-se, nesta seção, uma análise das bases de dados referentes aos seis setores econômicos representados. Tais bases não receberão nenhum tratamento especifico no que tange ao desbalanceamento inerente de suas instâncias.

Inicialmente, será determinado qual classificador apresenta um melhor desempenho na predição de insolvência para este setor especifico. Para tal, foram utilizadas as seguintes estratégias: regressão logística (RL), máquina de vetor suporte (SVM), *multilayerperceptron* (MLP) e árvore de decisão (AD), da mesma forma dos capítulos anteriores. Em uma segunda etapa, o classificador que apresentar melhor resultado para este setor será utilizado em um procedimento de seleção de características utilizando a abordagem *wrapper* tendo sido definido o GS como o algoritmo de busca. A parametrização dos algoritmos é mantida igual aquela utilizada no capitulo anterior.

Pretende-se, então, avaliar a relevância, coerência e importância dos atributos ou variáveis econômicas selecionadas para cada setor econômico. Desta forma, pretende-se avaliar a existência ou não de comportamentos diferenciados em relação às variáveis econômicas envolvidas no processo para predição de insolvência de setores econômicos diferenciados.

6.2.1 Análise de classificadores com a base de dados original do setor econômico de materiais de básicos.

Neste grupo de empresas há 30 instâncias representando as insolventes e 351 instâncias representando as empresas solventes, também deste setor. Na Tabela 6.13 são apresentados os testes dos classificadores utilizando a base de dados das empresas do setor econômico de materiais básicos.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	25	5	0,757	0,907	22	8	0,83	0,928	25	5	0,847	0,923	24	6	0,873	0,942
S	11	340	0,977	0,907	1	350	0,987	0,928	4	347	0,987	0,923	1	350	0,99	0,942

Tabela 6. 13-Resultados dos classificadores no treinamento da base de dados original referente ao setor econômico de materiais básicos

Para a base de dados do setor econômico de empresas de materiais básicos, em relação a matriz de confusão, a medida F e área ROC apresentaram pouca diferença entre os classificadores. Entretanto, SVM e AD foram aqueles que foram capazes de melhor classificar tanto as empresas solventes quanto as empresas insolventes. O método AD obteve um resultado superior, portanto sendo utilizado como algoritmo indutor no processo de seleção de atributos.

As variáveis selecionadas empregando-se as técnicas estudadas foram: EOCpOT, EOAT, GAF, MB, TERFIN. Uma avaliação mais pormenorizada da correlação destas variáveis com o setor que representam será feita no próximo capitulo.

6.2.2 Análise de classificadores com a base de dados original do setor econômico de empresas de consumo cíclico.

Compõem a base de dados deste setor econômico 50 instâncias das empresas insolventes e 474 instâncias das empresas solventes. Na Tabela 6.14 são apresentados os testes dos classificadores utilizando a base de dados das empresas do setor econômico de consumo cíclico.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	41	9	0,811	0,921	31	19	0,738	0,813	41	9	0,877	0,923	42	8	0,884	0,942
S	10	464	0,98	0,921	3	471	0,977	0,813	3	471	0,987	0,923	3	471	0,988	0,942

Tabela 6. 14- Resultados dos classificadores no treinamento da base de dados original referente ao setor econômico de empresas de consumo cíclico

Na base de dados das empresas do setor econômico de empresas de consumo cíclico, os resultados obtidos apresentam algumas diferenças entre eles. RL classifica bem a maioria das empresas insolventes, mas apresenta erros significativos nas empresas solventes (9) evidenciando sua boa capacidade de classificar a classe minoritária. MPL erra na classificação de algumas (9) empresas insolventes, mas erra pouco nas empresas solventes (3), obtém uma medida F melhor do que o classificador RL e obtém um AUC pouco superior ao RL. SVM teve um desempenho inferior ao RL e ao MPL para esta base de dados. AD foi o classificador de melhor desempenho na medida F entre os estudados para esta base de dados. Na MC pode-se verificar que são oito instâncias insolventes classificada incorretamente. Mesmo diante disso AD foi o classificador que obteve os melhores resultados para esta base de dados referente ao setor econômico de consumo cíclico.

As variáveis mais bem selecionadas pelas técnicas estudadas foram: EOCpOT, GAF, LC, MB, EBITDA.

6.2.3 Análise de classificadores com a base de dados original do setor econômico de empresas de consumo não cíclico.

Compõem a base de dados deste setor 25 instâncias de empresas insolventes e 168 instâncias de empresas solventes. Na Tabela 6.15 são apresentados os testes dos classificadores utilizando a base de dados das empresas do setor econômico de empresas de consumo não cíclico.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	19	6	0,745	0,927	20	5	0,889	0,9	20	5	0,851	0,86	22	3	0,88	0,942
S	7	163	0,962	0,927	0	170	0,986	0,9	2	168	0,98	0,86	3	167	0,982	0,942

Tabela 6. 15- Resultados dos classificadores no treinamento da base de dados original referente ao setor econômico de empresas de consumo não cíclico

Na base de dados referente ao setor econômico de consumo não cíclico os melhores resultados ficaram com os classificadores RL e AD na área ROC. RL classificou corretamente 19 das empresas insolventes, mas errou em 6 empresas solventes. Já AD errou em 3 empresas insolventes e errou em apenas 3 nas solventes. A medida F apresentou valores bem próximos de AD e SVM na classe das insolventes, 0,88 e 0,889, respectivamente.

Assim aplicando o AD no processo de seleção de atributos, têm-se as seguintes variáveis: EOPL, ETPL, GAF, ROI, EOCpOT.

6.2.4 Análise de classificadores com a base de dados original do setor econômico de empresas de bens industriais.

Compõem a base de dados deste setor 10 instâncias caracterizando empresas insolventes e 202 instâncias caracterizando empresas solventes.

Na Tabela 6.16 são apresentados os testes dos classificadores utilizando a base de dados das empresas do setor econômico de empresas de bens industriais.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	7	3	0,467	0,848	6	4	0,75	0,8	7	3	0,778	0,928	7	3	0,737	0,923
S	13	189	0,959	0,848	0	202	0,99	0,8	1	201	0,99	0,928	2	200	0,988	0,923

Tabela 6. 16- Resultados dos classificadores no treinamento da base de dados original referente ao setor econômico de empresas de bens industriais

Para o setor econômico de bens industriais, os resultados para os classificadores não apresentaram distorções relevantes. Todos os classificadores tiveram nas classes das empresas insolventes erros parecidos e foram bem eficazes na classificação das empresas insolventes e solventes. O classificador MLP obteve as melhores medidas F e valores de AUC.

Assim aplicando o MLP no processo de seleção de atributos, têm-se as seguintes variáveis: EOCpOT, LS, EBITDA, TERFIN.

6.2.5 Análise de classificadores com a base de dados original do setor econômico de empresas de construções e transportes.

Compõem a base de dados deste setor 15 instâncias caracterizando empresas insolventes e 230 instâncias caracterizando empresas solventes. Na Tabela 6.17 são apresentados os testes dos classificadores utilizando a base de dados das empresas do setor econômico de empresas de construções e transportes.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	12	3	0,533	0,845	12	3	0,8	0,89	12	3	0,857	0,927	9	6	0,667	0,811
S	18	212	0,953	0,845	3	227	0,987	0,89	1	229	0,99	0,927	3	227	0,981	0,811

Tabela 6. 17- Resultados dos classificadores no treinamento da base de dados original referente ao setor econômico de empresas de construção e transportes

No setor econômico de empresas de construção e transportes os classificadores que obtiveram os melhores medidas foram MLP.

Assim aplicando o MLP no processo de seleção de atributos, têm-se as seguintes variáveis: EOCpOT, LI, EBIT.

6.2.6 Análise de classificadores com a base de dados original do setor econômico de empresas de tecnologia da informação e telecomunicações.

Compõem a base de dados deste setor 10 instâncias caracterizando empresas insolventes e 43 instâncias caracterizando empresas solventes. Na Tabela 6.18 são

apresentados os testes dos classificadores utilizando a base de dados das empresas do setor econômico de empresas de tecnologia da informação e telecomunicações.

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	8	2	0,695	0,845	9	1	0,9	0,938	12	3	0,857	0,927	8	2	0,842	0,93
S	5	38	0,915	0,845	1	42	0,977	0,938	1	229	0,99	0,927	1	42	0,965	0,93

Tabela 6. 18- Resultados dos classificadores no treinamento da base de dados original referente ao setor econômico de empresas de tecnologia da informação e telecomunicações

Para o setor econômico de empresas de tecnologia da informação e de telecomunicações, os classificadores MLP e SVM apresentaram os melhores resultados e, por conseguinte, apresentaram melhores capacidades de interpretar essa base de dados.

Assim aplicando o SVM no processo de seleção de atributos, têm-se as seguintes variáveis: EOCpOT, GAF, ROA, ROE.

O desempenho dos classificadores para a base de dados originais completa, onde não há segmentação em relação aos setores econômicos, foi liderado pela AD e SVM. Deve-se ressaltar que como ocorreu na base completa, à predição das empresas solventes apresentou melhor resultado em relação à predição das empresas insolventes para as bases segmentadas. Credita-se tal desempenho ao fato das bases por segmento também serem bases desbalanceadas, assim como a base original.

Os resultados relativos à seleção de variáveis por setor indicam claramente a dependência da relevância das variáveis econômicas de acordo com o setor que representam. Assim, a análise por setor, baseada nas variáveis selecionadas, permite uma avaliação mais acurada da insolvência das empresas deste setor com a utilização das variáveis mais representativas.

6.3 Análises por setor econômico aplicando sub-bases e o algoritmo SEID

Nesta seção, será feita uma análise da predição por setor econômico utilizando sub-bases balanceadas. Para cada um dos 6 setores econômicos representados serão gerados 3 sub-amostras pelo algoritmo SEID. O procedimento é referente a Figura 5.1.

A seguir, são apresentados os resultados das predições para cada setor econômico com a aplicação do SEIDwS. O procedimento realizado nesta seção foi apresentado na Figura 5.2. A dimensão das sub-bases é de 30 x 30, ou seja, 30 instâncias pertencentes às empresas solventes e 30 instâncias pertencentes às empresas insolventes. O tamanho reduzido das instâncias definidas como insolventes dos diversos setores econômicos determinaram a dimensão de 30 x 30. O classificador utilizado como indutor foi aquele que obteve as melhores medidas para cada setor econômico apresentado na seção anterior.

6.3.1 Bases de dados de empresas do setor econômico de materiais básicos

Na tabela 6.19 são apresentados os resultados das métricas geradas em cada uma das sub-amostras balanceadas, conforme Figura 5.2. A Tabela mostra que na classe minoritária em todos os novos modelos, a classificação é bem sucedida, classificando corretamente todas as instâncias.

	SB1_MB30WRAPPER+GS				SB2_MB30WRAPPER+GS				SB3_MB30WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	30	0	0,968	0,984	30	0	0,968	0,984	30	0	0,937	0,967
S	2	349	0,997	0,984	2	349	0,997	0,984	4	347	0,994	0,967

Tabela 6. 19- Resultados referentes ao setor econômico de materiais básicos.

Na Tabela 6.20 é apresentado o resultado da aplicação do SEIDwS, e comparado com os resultados da base original do setor.

	BASE ORIGINAL				SEIDwS			
Classe	MC		F	AUC	MC		F	AUC
I	25	5	0,847	0,937	30	0	0,968	0,984
S	4	347	0,987	0,937	2	349	0,997	0,984

Tabela 6. 20- Comparação dos resultados referentes ao setor econômico de materiais básicos da base original com a aplicação da técnica SEIDwS

6.3.2 Bases de dados das empresas do setor econômico de consumo cíclico

Na tabela 6.21 são apresentados os resultados em cada uma das sub-amosra balanceadas, conforme Figura 5.2. A Tabela mostra que na classe minoritária em todos os novos modelos a classificação é bem sucedida, classificando corretamente todas as instâncias.

	SB1_CC30WRAPPER+GS				SB2_CC30WRAPPER+GS				SB3_CC30WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	50	0	0,925	0,958	50	0	0,909	0,950	50	0	0,793	0,908
S	8	466	0,991	0,958	10	464	0,989	0,950	26	448	0,971	0,908

Tabela 6. 21- Resultados referentes ao setor econômico de consumo cíclico.

Na Tabela 6.22 é apresentado o resultado da aplicação do SEIwS comparado com a base original do setor.

	BASE ORIGINAL				SEIDwS			
Classe	MC		F	AUC	MC		F	AUC
I	42	9	0,884	0,942	50	0	0,947	0,991
S	3	471	0,988	0,942	9	465	0,99	0,991

Tabela 6. 22-Comparação dos resultados referentes ao setor econômico de consumo cíclico da base original com a aplicação da técnica SEIDwS.

6.3.3 Bases de dados das empresas do setor econômico de consumo não cíclico

Na tabela 6.23 são apresentados os resultados em cada uma das sub-amostras balanceadas, conforme Figura 5.2. A Tabela mostra que na classe minoritária em todos os novos modelos a classificação é bem sucedida, classificando corretamente todas as instâncias.

	SB1_NC30WRAPPER+GS				SB2_NC30WRAPPER+GS				SB3_NC30WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	25	0	0,862	0,947	25	0	0,877	0,950	25	0	0,877	0,950
S	8	162	0,976	0,947	7	163	0,978	0,950	7	163	0,978	0,950

Tabela 6. 23-Resultados referentes ao setor econômico de consumo não cíclico.

Na tabela 6.24 é apresentado o resultado da aplicação do SEIDwS, conforme Figura 5.3. A Tabela mostra que houve uma melhora na classificação das empresas solventes em relação aos três modelos mostrados na tabela 6.24.

	BASE ORIGINAL				SEIDwS			
Classe	MC		F	AUC	MC		F	AUC
I	22	3	0,88	0,942	25	0	0,963	0,992
S	3	167	0,982	0,942	6	164	0,97	0,992

Tabela 6.24-Resultados referentes ao setor econômico de consumo não cíclico com a aplicação do SEIDwS com o resultado original.

6.3.4 Base de dados das empresas do setor econômico de bens industriais

Na Tabela 6.25 são apresentados os resultados após a aplicação do SEID, conforme Figura 5.2. A Tabela mostra que na classe minoritária em todos os novos modelos a classificação é bem sucedida, classificando corretamente todas as instâncias.

	SB1_BI30WRAPPER+GS				SB2_BI30WRAPPER+GS				SB3_BI30WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	10	0	0,952	0,981	10	0	0,952	0,981	10	0	0,8	0,943
S	1	201	0,998	0,981	1	201	0,998	0,981	5	197	0,987	0,943

Tabela 6. 25-Resultados referentes ao setor econômico de bens industriais.

Na Tabela 6.26 é apresentado o resultado da aplicação do SEIDwS com a base original, conforme Figura 5.3.

	BASE ORIGINAL				SEIDwS			
Classe	MC		F	AUC	MC		F	AUC
I	7	3	0,778	0,928	10	0	0,952	0,981
S	1	201	0,99	0,928	1	201	0,998	0,981

Tabela 6. 26-Resultados referentes ao setor econômico de bens industriais com a aplicação do SEIDwS com o resultado original

6.3.5 Bases de dados das empresas do setor econômico de construções e transportes

Na Tabela 6.27 são apresentados os resultados, conforme Figura 5.2. Esta Tabela mostra que na classe minoritária, em todos os novos modelos, a classificação é bem sucedida, classificando corretamente todas as instâncias.

	SB1_CT30WRAPPER+GS				SB2_CT30WRAPPER+GS				SB3_CT30WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	15	0	0,857	0,956	15	0	0,75	0,930	15	0	0,909	0,968
S	5	225	0,989	0,956	10	220	0,978	0,930	3	227	0,993	0,968

Tabela 6. 27-Resultados referentes ao setor econômico de construção e transporte.

Na Tabela 6.28 é apresentado o resultado da aplicação do SEIDwS com a base original, conforme Figura 5.3.

	BASE ORIGINAL				SEIDwS *WRAPPER*			
Classe	MC		F	AUC	MC		F	AUC
I	12	3	0,857	0,927	15	0	0,882	0,989
S	1	229	0,99	0,927	4	226	0,991	0,989

Tabela 6. 28-Resultados referentes ao setor econômico de construções e transporte com a aplicação do SEIDwS com o resultado original

6.3.6 Base de dados das empresas do setor econômico de tecnologia da informação

Na Tabela 6.29 são apresentados os resultados, conforme Figura 5.2. A Tabela mostra que na classe minoritária em todos os novos modelos a classificação é bem sucedida, classificando corretamente todas as instâncias.

	SB1_TI30WRAPPER+GS				SB2_TI30WRAPPER+GS				SB3_TI30WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	10	0	0,8	0,942	10	0	0,909	0,957	10	0	0,909	0,957
S	5	38	0,938	0,942	2	41	0,976	0,957	2	41	0,976	0,957

Tabela 6. 29-Resultados referentes ao setor econômico da tecnologia da informação e telecomunicações.

Na Tabela 6.30 é apresentado o resultado da aplicação do SEIDwS, e comparado com os resultados da base original do setor.

	BASE ORIGINAL				SEIDwS			
Classe	MC		F	AUC	MC		F	AUC
I	9	1	0,9	0,938	10	0	0,952	0,974
S	1	42	0,977	0,938	1	42	0,948	0,974

Tabela 6. 30-Resultados referentes ao setor econômico da tecnologia da informação com a aplicação do SEIDwS com o resultado original

Os resultados mostraram, em geral, que o balanceamento proporcionado pelo algoritmo SEID aumenta bastante a capacidade de previsão de empresas insolventes, independente dos setores analisados. Nota-se também que a utilização do SEIDwS aumenta a robustez na previsão das empresas solventes em relação as sub-base balanceadas. De uma forma geral, os resultados de predição por setor apresentaram uma qualidade de predição superior em relação à utilização de uma base de dados completa, com todos os setores representados. Apesar de alguns setores apresentarem poucas instâncias, as estratégias das sub-bases desbalanceadas do SEID evitaram distorções de nível de predição entre as classes.

6.4 Comparação das técnicas SEIDwS e SMOTE na base de dados completa e segmentadas por setores econômicos

Nesta subseção são testadas as bases de dados completas e as segmentadas por setores econômicos utilizando a técnica apresentada no algoritmo SEIDwS e comparados com os gerados pelo algoritmo SMOTE.

Na Tabela 6.30 são apresentados os resultados referentes aos testes feitos nas bases de dados dos setores econômicos estudados, comparando a técnica de balanceamento apresentada nesta tese mais a técnica de votação majoritária com a técnica do SMOTE (já explicitada em seções anteriores). No teste com o SMOTE, o *k* (número de vizinhos mais próximos) utilizado foi igual a 5.

		Técnicas de balanceamento			
	Classe	SEIDwS		SMOTE	
		F	AUC	F	AUC
Base de dados original sem segmentação	I	0,939	0,944	0,911	0,993
	S	0,945	0,944	0,991	0,993
Base de dados por setor econômico					
Materiais básicos	I	0,968	0,984	0,979	0,987
	S	0,997	0,984	0,989	0,987
Consumo cíclico	I	0,947	0,991	0,986	0,993
	S	0,99	0,991	0,986	0,993
Consumo não cíclico	I	0,963	0,992	0,91	0,993
	S	0,97	0,992	0,911	0,993
Bens industriais	I	0,952	0,981	0,975	0,984
	S	0,998	0,981	0,995	0,984
Construção e Transportes	I	0,882	0,989	1	1
	S	0,991	0,989	1	1
Tecnologia da informação e telecomunicações	I	0,952	0,974	1	1
	S	0,948	0,974	1	1

Tabela 6. 31-Comparação do SEIDwS e o SMOTE com bases segmentas por setores econômicos.

A tabela 6.31 evidencia a importância da segmentação econômica para se obter melhores resultados na classificação das empresas insolventes. Na base sem segmentação, os valores encontrados da medida F para as insolventes, sempre foram inferiores, em relação aos encontradas nas bases segmentadas.

Na comparação com o SMOTE algoritmo adotou-se o valor de 5 para o número de vizinhos mais próximos. Deve-se ressaltar que também foi aplicada a seleção de atributos quando se utiliza o algoritmo SMOTE.

Em relação aos algoritmos de balanceamento testados, os resultados ficaram bastante próximos, sendo que o SEIDwS apresentou resultados melhores para alguns setores e o SMOTE para outros setores econômicos.

6.5 Considerações Finais

Estudos de previsão de insolvência feitos com dados agrupados por setores econômicos geram melhores resultados em suas classificações tanto para as empresas insolventes quanto para as solventes Isso se deve a grande homogeneidade econômica existente entre os dados das empresas pertencentes ao mesmo setor econômico. Apesar dessas evidências, poucos são os estudos publicados na literatura específica que realizam estas discriminações na etapa de elaboração das bases de dados com o objetivo principal de prever insolvência de empresas. Há um acentuado ganho na qualidade dos resultados obtidos nessas classificações como conseqüência do agrupamento dos dados econômico-financeiros por setores econômicos. Isso pode ser evidenciado nas bases de dados em que não foi realizado o balanceamento. Pelo estudo aqui realizado, fica evidente a importância da necessidade da setorização econômica das empresas para se obter uma melhor acurácia nas classificações. Até mesmo na etapa de seleção de características, os resultados foram distintos para cada setor econômico. Diante disso, pode ser considerado como uma etapa de pré-processamento essa discriminação econômico setorial quando se estuda previsão de insolvência em empresas.

CAPÍTULO 7 Regras de classificação para empresas por setores econômicos

Neste capítulo serão apresentadas as regras de classificação geradas de acordo com cada grupo de empresas pertencentes aos seus respectivos setores econômicos. Estas regras estão de acordo com a base de dados aqui estudada facilitando no entendimento de uma provável dificuldade financeira que a empresa venha a enfrentar em um futuro.

7.1 Regras de classificação das empresas do setor econômico materiais básicos

São apresentadas aqui as regras de classificação referentes as empresas do setor econômico de materiais básicos mostrando as possibilidades financeiras, de acordo com as variáveis empregadas, que podem levar as empresas deste setor econômico à um estágio de insolvência.

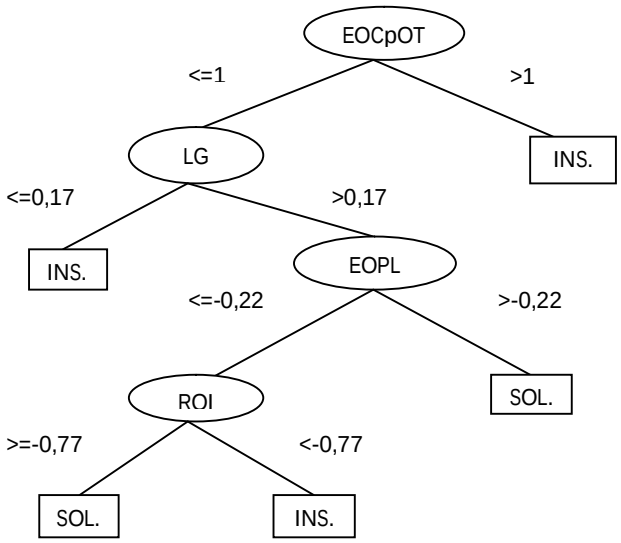

Figura 7. 1-Regras de classificação das empresas do setor econômico materiais básicos.

Pelas regras geradas para as empresas do setor econômico de materiais básicos, aquelas empresas que apresentarem a variável EOCpOT (Endividamento Oneroso de Curto Prazo sobre Oneroso Total) com um valor superior a 1 (100%) podem ser classificadas como uma empresa com grandes possibilidades de vir a tornar insolvente. Empresas com altos valores nessa variável apresentam obrigações financeiras de curto prazo bastante elevadas em relação às obrigações financeiras totais. Nessas empresas há um acentuado comprometimento financeiro que, de acordo com as regras, pode levar a empresa à insolvência.

A variável seguinte a ser analisada é a LG (liquidez geral) que é uma variável com natureza de solvência (relação da capacidade que a empresa tem de pagar todas suas obrigações de curto e longo prazo). As regras encontradas indicam que as empresas de materiais básicos com índices de Liquidez Geral inferior ou igual a 0,17 apresentam características de serem insolventes já aquelas empresas com índices de LG superiores a 0,17 apresentam características de serem solventes.

EOPL (Endividamento Oneroso sobre Patrimônio Líquido) é a variável que, pela regra, deve ser analisada após a LG. EOPL representa a proporção de endividamento financeiro em relação ao capital próprio. O valor limite é de -0,22, ou seja, empresas com valores maiores do que este podem ser consideradas como empresas solventes. Já empresas com valores abaixo disso, somente se poderá tirar alguma conclusão após a análise da variável ROI. Para essa variável o valor limite encontrado é -0,77, portanto, valores acima disso indicam que a empresa pode ser considera como solvente, caso contrário, ela pode ser considerada insolvente.

Empresas deste setor econômico devem ser entendidas como aquelas em que as variáveis de endividamento (EOCpOT e EOPL), liquidez geral (LG) junto com a de rentabilidade (ROI), possuem alto poder de discriminação entre as empresas solventes e insolventes, podendo ser concluído que as variáveis que meçam desempenhos operacionais não aparecem como tão relevantes nesta discriminação, não há variável com origem no Demonstrativo de Resultado do Exercício. Os resultados indicam que para empresas deste setor econômico, a influência do desempenho financeiro prepondera sobre o operacional na caracterização da continuidade destas empresas.

7.2 Regras de classificação das empresas do setor econômico de consumo cíclico

Nesta subseção são apresentadas as regras de classificação referentes as empresas do setor econômico de consumo cíclico mostrando as possibilidades financeiras, de acordo com as variáveis usadas, que podem levar as empresas deste setor econômico à um estágio de insolvência.

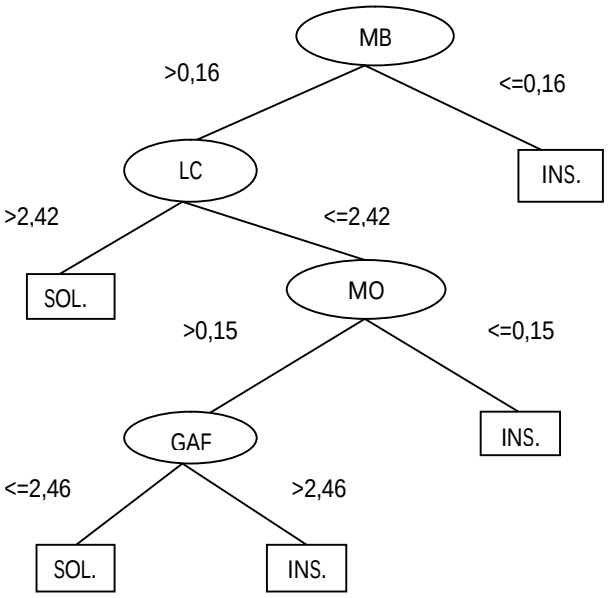

Figura 7. 2- Regras de classificação das empresas do setor econômico de consumo cíclico

Pelas regras obtidas para as empresas pertencentes ao setor econômico de consumo cíclico, a variável inicial para análise na caracterização da empresa em solvente ou insolvente é a MB (Margem bruta). Pelas regras, se MB for inferior a 0,16 a empresa pode ser considerada insolvente, já para valores superiores a 0,16, deve-se analisar outras variáveis. A variável seguinte a ser analisada é a LC (Liquidez corrente). Se esta variável for superior a 2,42 a empresa pode ser classificada como solvente, caso contrário, deve-se analisar a variável seguinte que é a MO (Margem operacional). MO com valores menores ou iguais a 0,15, a empresa pode ser caracterizada como insolvente, MO com índices superiores a 0,15 há necessidade de verificação de outra

variável, GAF (Grau de alavancagem financeira). Empresas com valores de GAF superiores a 2,46, são consideradas insolventes, caso contrário, elas podem ser classificadas como empresas solventes.

As empresas deste setor econômico que não atentarem para as variáveis referentes à avaliação de desempenhos operacionais (MB e MO), com atuações nas suas estruturas de custos, tecnologia e competitividade, eficiência das equipes de vendas, audácia e criatividade e adequação de seu planejamento estratégico, certamente, poderão apresentar dificuldades na sua continuidade. Em empresas pertencentes a este setor econômico, os atributos operacionais devem prevalecer, porém não se esquecendo da liquidez (LC) e da alavancagem operacional (GAF).

7.3 Regras de classificação das empresas do setor econômico de consumo não cíclico

Nesta subseção são apresentadas as regras de classificação referentes às empresas do setor econômico de consumo não cíclico mostrando as possibilidades financeiras que podem levar as empresas deste setor econômico a um estágio de insolvência, de acordo com as variáveis empregadas.

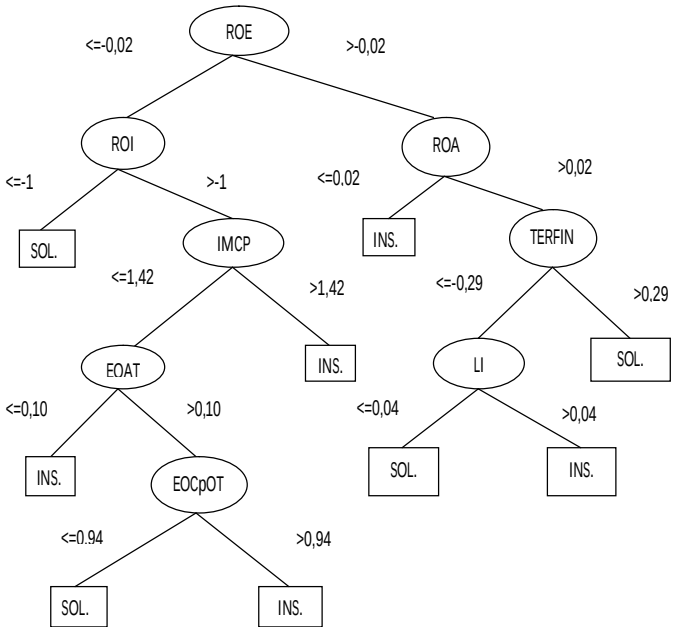

Figura 7. 3-Regras de classificação das empresas do setor econômico de consumo não cíclico

As regras geradas na amostra das empresas do setor econômico de consumo não cíclico apresentam a variável ROE (Retorno sobre o capital) como a primeira ser analisada visando classificar a empresa como solvente ou insolvente. Se o seu valor for menor ou igual a -0,02 a variável seguinte a ser analisada é ROI (Retorno sobre investimento), entretanto, se o valor de ROE for maior do que -0,02 a próxima variável a ser analisada é ROA (Retorno sobre o investimento operacional). Neste setor econômico, diferentemente dos setores anteriormente analisados, a primeira variável não permite obter uma conclusão final em relação à classe da empresa. Somente a partir das variáveis ROI e ROA é que as empresas podem ser classificadas como solventes ou insolventes. Se ROI for menor ou igual a -1 a empresa pode ser classificada como solvente. Se ROA for menor ou igual a 0,02 a empresa pode ser classificada como insolvente. Com ROI maior do que -1 a próxima variável a ser analisada será IMCP (Grau de imobilização do capital próprio), se ela for maior do que 1,42 a empresa pode ser classificada como insolvente, caso contrário, deve-se avaliar EOAT (Endividamento oneroso sobre ativo total). Se EOAT for menor do que 0,10 a empresa será insolvente. Para EOAT maior que este valor, a próxima variável a ser analisada será EOCpOT (Endividamento oneroso de curto prazo sobre oneroso total). Para esta variável o limite é 0,94, ou seja, se maior que este valor, a empresa pode ser classificada como insolvente, já se EOCpOT for menor a empresa pode ser classificada como solvente.

Partindo da variável ROA, se o seu valor for menor ou igual a 0,02 a empresa pode ser considerada insolvente, caso contrário, a variável seguinte a ser analisada será TERFIN (Termômetro financeiro). Para valores desta variável maiores que 0,29 a empresa é solvente, caso ocorra o contrário, a variável LI (Liquidez imediato) definirá a classe da empresa. Para LI maior do que 0,04 a empresa é insolvente, já para LI menor ou igual a 0,04 a empresa será considerada solvente.

Pelas regras extraídas na amostra de empresas deste setor econômico, as variáveis de rentabilidade (ROE, ROA e ROI) são as determinantes para caracterizar a empresas em prováveis solventes ou insolventes. Para estas empresas é preponderante se acautelar na combinação de sua lucratividade, da qualidade e eficiência na utilização do Ativo e da estratégia de financiamento destes ativos utilizados na consecução do seu objeto empresarial.

7.4 Regras de classificação das empresas do setor econômico de bens industriais

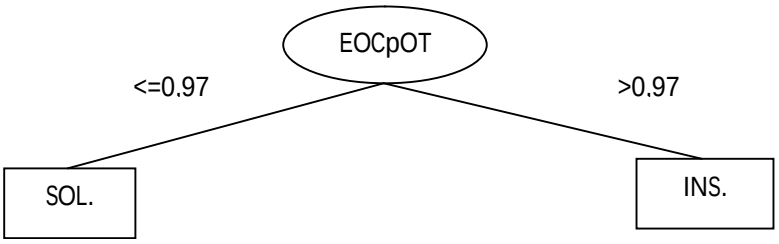

Figura 7. 4- Regras de classificação das empresas do setor econômico de bens industriais

A regra gerada pela amostra das empresas do setor econômico de bens de capital apresenta apenas uma variável determinante na caracterização de empresas solventes ou insolventes. Para valores da variável EOCpOT iguais ou menores que 0,97 a empresa pode ser classificada como solvente, para valores maiores que 0,97 a empresa pode ser classificada como insolvente. Cabe ressaltar do tamanho reduzido dessa amostra podendo, por conseguinte, gerar um comprometimento na qualidade das regras geradas.

7.5 Regras de classificação das empresas do setor econômico de construções e transportes

Nesta subseção são apresentadas as regras de classificação referentes às empresas do setor econômico de construções e transportes mostrando as possibilidades financeiras que podem levar as empresas deste setor econômico a um estágio de insolvência, de acordo com as variáveis utilizadas no modelo.

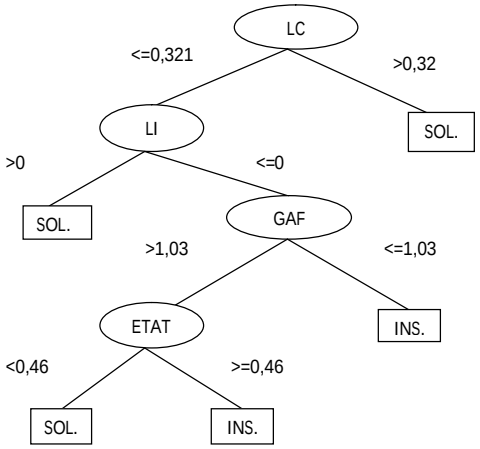

Figura 7. 5- Regras de classificação das empresas do setor econômico de construções e transportes

Para as regras de classificação das empresas pertencentes ao setor econômico de construções e transportes, a variável LC (Liquidez corrente) se apresenta como o primeiro índice delimitador para a classificação da empresa em solvente ou insolvente. Empresas com LC maior do que 0,32 podem ser classificadas como solventes, caso contrario, as empresas devem recorrer a outra variável de liquidez, a liquidez imediata (LI). As empresas com LI maior do que 0 são classificadas como solventes, já aquelas com índices menores ou iguais a 0 deverão se classificadas após a análise da variável GAF. Se a empresa apresenta um GAF menor ou igual a 1,03 ela será classificada como insolvente, entretanto se ela tiver um GAF maior do que 1,03 a variável seguinte a ser analisada será ETAT (Endividamento total sobre ativo total). Empresas com ETAT maior do que 0,46 serão classificadas como solventes, enquanto empresas com ETAT menor ou igual a 0,46 serão classificadas como insolventes.

As empresas deste setor econômico devem preservar mais os aspectos patrimoniais, evidenciando os capitais aplicados (bens e direitos) e a origem dos mesmos (obrigações). Para essas empresas, a capacidade de quitar suas obrigações de curto prazo (LC e LI), aliado a composição das obrigações (GAF e ETAT), sendo determinantes na sua continuidade. Estas empresas são bem mais dependentes da conservância de uma relação saudável entre suas contas patrimoniais.

7.6 Regras de classificação das empresas do setor econômico de tecnologia da informação e telecomunicações

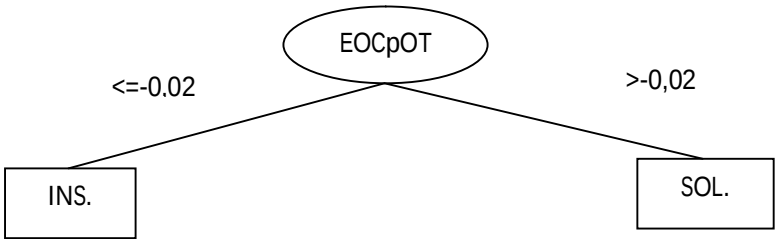

Figura 7. 6- Regras de classificação das empresas do setor econômico de tecnologia da informação e telecomunicações

A regra gerada pela amostra das empresas do setor econômico de bens de tecnologia da informação e telecomunicações apresenta apenas uma variável determinante na caracterização de empresas solventes ou insolventes, a variável EOCpOT. As empresas deste setor econômico com valor de EOCpOT igual ou menor que -0,02 podem ser classificadas como insolventes, caso contrário podem ser classificadas como solventes. Cabe ressaltar do tamanho reduzido dessa amostra podendo, por conseguinte, gerar um comprometimento na qualidade das regras geradas.

CAPITULO 8 Conclusões e futuros estudos

O primeiro problema estudado nesta tese diz respeito à origem dos dados a serem utilizados, podendo vir dos demonstrativos contábeis ou do mercado. Nesta tese foram apresentados alguns artigos publicados na literatura específica nos quais evidenciaram a maior precisão dos resultados quando da utilização de dados originados nos demonstrativos contábeis. Com esses dados, os modelos preditivos obtiveram melhores valores de acurácia na predição do que nos modelos elaborados com dados do mercado quando da avaliação da capacidade de continuidade de uma entidade. Somente quando da adoção de várias modificações em relação aos modelos tradicionais, os modelos elaborados com dados do mercado apresentaram melhores resultados conforme evidenciado em Hillegelst *et al*., (2004). O trabalho de Famá (1970) já questionava a eficiência do mercado em razão da reação à informação e, por conseguinte a capacidade dos dados de sua origem na composição de modelos referentes à predição. Mensah (1984), Reisz e Perlich (2007), Agarwal e Taffler (2008) e Chen e Cheng (2009) evidenciaram em seus estudos a preferência nos modelos de predição de insolvência com dados originados nos registros contábeis através dos resultados obtidos nestes modelos em relação aos elaborados com dados de mercado. Para Haugen e Baker (2008) há evidências de que os mercados de ações dos EUA são altamente ineficientes podendo, assim ser inferido de que, também o mercado brasileiro apresente altas ineficiências com relação às informações e consequentemente ao uso desses dados na elaboração de modelos preditores da saúde financeira de empresas. Altman (2005) levantou algumas questões referentes aos modelos elaborados com dados originados nos livros contábeis e no mercado. Nos dois modelos há aspectos que contribuem para a ineficiência dos resultados, tal estudo foi realizado no mercado mexicano após a crise de 1994 e o autor infere os resultados para os mercados emergentes.

Problemas referentes à previsão de insolvência de empresas sempre se deparam com a desigualdade na proporção das amostras envolvidas no estudo. A quantidade de empresas que apresentam insolvência em ambientes econômicos normais é bem inferior ao de empresas saudáveis financeiramente. Além dessa consideração, é importante também salientar a escassez de informações confiáveis das empresas insolventes devido a vários fatores. Diante disso, as amostras com dados de empresas insolventes são pequenas, gerando assim um problema de desbalanceamento entre a amostra de empresas solventes e a amostra de empresas insolventes. A classificação da amostra de

empresas insolventes acaba sendo prejudicada devido ao problema do desbalanceamento entre amostras. Esta tese estudou e propôs um método de balanceamento para tratar este tipo de problema. Os resultados apresentados pelo método proposto foram competitivos com o método de balanceamento publicado na literatura conhecido como SMOTE. Através do método proposto, empresas pertencentes a amostra das insolventes são bem caracterizadas chegando a obter erros do tipo I e tipo II pouco significativos. Uma grande vantagem do método proposto por esta tese - SEID – em relação ao SMOTE é que o método processa os dados reais obtidos nas bases de dados. Diferente do SMOTE que, através de sua metodologia gera bases com dados artificiais e utiliza somente *over-sampling*, o SEID emprega em suas bases somente os dados reais obtidos nos demonstrativos contábeis das empresas.

Na comparação dos resultados encontrados nesta tese com outros estudos (Tsai, Tsai e Wu, Nanni e Lumini) em bases de dados (UCI), os resultados também se mostraram bem eficientes e competitivos.

Para que previsão de insolvência apresente resultados mais confiáveis e consistentes, é necessária a segmentação econômica dos dados, através de seus indicadores contábeis. Empresas do mesmo setor econômico apresentam estruturas contábeis e patrimoniais bem semelhantes. Com dados de empresas pertencentes estritamente ao mesmo setor econômico, a classificação das empresas em insolventes e solventes apresenta maior eficiência. Portanto, é importante despender atenção na elaboração das amostras com dados de acordo com as empresas de mesmo setor econômico e os resultados deste estudo evidenciam isso. Em vários estudos sobre o tema de previsão de insolvência de empresas, os erros ocorridos talvez possam ser creditados ao não reconhecimento da setorização econômica dos dados.

Ainda na fase de pré-processamento para previsão de insolvência, o reconhecimento da necessidade de sistematizar a seleção de atributos ficou bem evidenciado nesta tese. Modelos elaborados com dados contábeis, devido à própria natureza dos dados, demandam um estudo mais detalhado na seleção de atributos. Características do sistema legal nacional, maneiras de se financiar, relacionamento entre o Fisco e a Contabilidade, tipo de empresas, tipo de regime ou critério de estrutura e apresentação de demonstrações contábeis são alguns dos principais motivos que determinam a necessidade de estudos mais detalhados na etapa de seleção de atributos. Além dos fatores citados pode ser acrescida a importância de se conhecer, para empresas de mesmo setor econômico, aquelas variáveis com maior capacidade de

discriminar empresas com maiores probabilidade de virem a apresentar problemas na sua saúde financeira. Os resultados apresentados na tese evidenciaram essas diferenças e influência da seleção de atributos para previsão de insolvência de empresas.

O melhor entendimento do processo de insolvência de empresas, sobretudo quando segmentado por setor econômico, é relevante no estudo de previsão de insolvência. Nesta tese foi apresentado que a premissa da segmentação das empresas por setores econômicos é preponderante e facilita na geração de melhores resultados contribuindo para obtenção de regras mais caracterizadas de acordo com o setor econômico no qual a empresa pertence. Nos resultados ficou evidente as diferenças nas regras para as empresas segmentadas setorialmente, regras que obtiveram nos testes, valores bem significativos, o que as habilita a uma melhor interpretação no processo de insolvência de empresas. Em relação às regras, ficou também evidente a importância do tamanho da amostra na obtenção de regras com melhores capacidades interpretativas, facilitando no desenvolvimento da habilidade do analista em predizer aquelas empresas com maiores probabilidades de vir a apresentar problemas em sua saúde financeira. O algoritmo utilizado para extração das regras foi o C4.5 devido a sua acessibilidade, eficiência e interpretabilidade além de ter sido o algoritmo com os melhores resultados nos testes de avaliação de desempenho de um classificador.

Quatro foram os classificadores pesquisados visando uma melhora classificação das empresas em solventes e insolventes, (i) regressão logística, (ii) a rede *multilayerperceptron*, (iii) máquina de vetor suporte e, (iv) arvore de decisão. Nos testes realizados que obtiveram os melhores resultados com as amostras das empresas foram os classificadores empregando *multilayerperceptron* e árvore de decisão. Tais resultados ficam mais evidenciados quando da segmentação econômica das empresas. Amostras de empresas segmentadas apresentaram resultados mais eficientes. A pesquisa do melhor classificador também foi útil para determinar o classificador utilizado na técnica de *wrapper* para selecionar atributos. A pesquisa evidenciou as diferentes características existentes entre os dados mesmo que esses tenham origem nos demonstrativos contábeis das empresas. Pode-se afirmar, de acordo com os resultados desta pesquisa, que para a classificação de empresas solventes e insolventes, a característica do classificador é determinante para se obter resultados mais eficazes mesmo considerando que neste estudo não tenha sido feito uma análise mais apurada dos índices que compõem os classificadores.

Na elaboração de modelos de previsão de insolvência, os dados podem ser temporais ou de painel. Nesta tese foram feitas análises comparativas dos efeitos dos dois tipos de dados na modelagem. Modelos que utilizam dados temporais apresentam a capacidade de prever o período ou o tempo no qual a empresa pode vir a se tornar insolvente. Neste tipo de dados, as informações referentes ao período (ano) da insolvência sempre são as que apresentam maiores entropias devido à "deterioração" contábil mais acentuado da empresa, entretanto esses dados, na prática, são impossíveis de serem aplicados porque representam um período em que a empresa foi declarada insolvente, portanto não há previsão efetivamente. Outra questão que depõe contra esses dados temporais diz respeito ao tamanho da amostra, sobretudo quando se está estudando com demonstrativos contábeis de empresas brasileiras de capital aberto. O número de empresas de capital aberto é pequeno comparado com estudos de outros países. Como já foi apresentado em capítulos anteriores, com esses dados, o modelo apresenta resultados nos testes inferiores aos modelos com dados de painel. Com o uso destes dados, as instâncias são os dados das empresas e, as variáveis, conseqüentemente, serão temporais. Já a modelagem com dados de painel ocorre das amostras apresentarem dimensões superiores. As instâncias passam a ser os períodos (anos). Os classificadores utilizados apresentaram melhores eficiências e, conseqüentemente, os resultados ficaram mais confiáveis. Com esses dados foi viável estudar previsão de insolvência segmentando as empresas economicamente (devido ao tamanho das amostras) e os modelos apresentaram diferenças facilitando um melhor entendimento sobre cada setor econômico.

O algoritmo apresentado neste trabalho pode ser descrito como sendo formado por um comitê de classificadores (agregação) para melhorar a capacidade de classificar as empresas, sobretudo as insolventes. A metodologia do comitê de classificadores consistiu em elaborar sub-bases de dados em que os dados originais das empresas insolventes (por ser em número bem inferior aos das solventes) sempre estiveram presentes nos novos bancos de dados. A metodologia se completa com a aplicação do *majority voting* (maioria dos votos). Através dessa técnica os possíveis erros de classificação das insolventes diminuem. Os resultados mostraram a eficiência da técnica na classificação das empresas, sobretudo as insolventes. Na validação dos resultados foram comparados os dados gerados pelo algoritmo proposto e o SMOTE (apresentado em capítulos anteriores). A comparação evidenciou a capacidade competitiva do algoritmo proposto podendo ser considerado como um bom algoritmo com fins de

balanceamento e posterior avaliação utilizando a votação majoritária (*majority voting)*. Cabe ressaltar a influência do pequeno tamanho da amostra nos resultados em dois dos setores econômicos: construção e transportes junto com tecnologia da informação e telecomunicações. Estes setores apresentaram resultados bem eficazes e semelhantes muito mais devido à pequena dimensão das amostras do que a qualidade dos dados.

Esta tese teve como foco principal caracterizar empresas que apresentam grandes possibilidades de se tornarem insolventes. Todo o estudo visou caracterizar melhor possíveis empresas insolventes utilizando e propondo novas técnicas. É importante salientar a respeito deste estudo que no ambiente econômico brasileiro há uma grande dificuldade na obtenção de dados em demonstrativos contábeis confiáveis de empresas insolventes. A cultura empresarial e gerencial no Brasil ainda demanda grandes evoluções no aspecto de controle e elaboração de bancos de dados qualificados e confiáveis. Diferentemente de outras economias, onde ocorre uma facilitação ao acesso de dados de empresas insolventes, no Brasil esse acesso aos dados contábeis é uma das etapas mais árduas neste tema. Talvez isso ocorra devido ao caráter eminentemente estratégico e legal das informações, mas parece que o motivo principal recaia nas questões de cultura empresarial e gerencial do ambiente de negócio no Brasil. Tem-se que evidenciar a permanente, mas lenta mudança desse ambiente talvez devido ao aumento do nível de competitividade entre os negócios e a constante mudança do ambiente produtor e consumidor. Outro fator preponderante das mudanças diz respeito às técnicas contábeis. Estas vêm acompanhando as mudanças nos ambientes econômicos para melhor adequar aos seus objetivos que é a de controlar e planejar entidades. Em relação às técnicas de *DM* estas vêm também sendo permanentemente incrementadas com novas e mais eficientes técnicas visando adequar melhor a modelagem e a descoberta de novos conhecimentos ao ambiente real dos negócios.

A principal e relevante limitação desta pesquisa se refere ao tamanho da amostra, citado anteriormente. O reduzido tamanho da amostra deve-se à escassez de informações contábeis confiáveis sobre empresas insolventes. Tais empresas quando começam a entrar no estado concordatário, costumam deixar de enviar informações à CVM. Por isso a amostra de insolvente fica reduzida, restringindo, conseqüentemente, a amostra total de solventes e insolventes limitando o estudo.

Uma opção que poderia ter a conseqüência de aumentar a amostra seria o uso de dados obtidos no SERASA. Entretanto, preferiu-se usar dados da CVM por este ser o órgão normativo do sistema financeiro brasileiro e também mais acessível e confiável.

Pesquisas em modelagem de previsão de insolvência vêm crescendo e se desenvolvendo cada vez mais. Novas técnicas de inteligência computacional, bancos de dados diferentes, novas variáveis sempre poderão ser implementadas e desenvolvidas visando melhorar os resultados. Essas técnicas devem procurar sempre acompanhar às constantes mudanças ocorridas no ambiente de negócios dentre essas mudanças pode ser citado a nova lei de falência. Essa nova lei que entrou em vigor no meio do ano de 2005 e pode ter gerado alterações nos resultados de uma futura modelagem de previsão de insolvência de empresas brasileiras de capital aberto.

Referência Bibliográfica

ABDOU, HUSSEIN A. "Genetic programming for credit scoring: The case of Egyptian public sector banks". *Expert Systems with Applications*, v. 36, Issue 9, November 2009, p. 11402-11417.

AHN B. S.; CHO S.S.; KIM C. Y. "The integrated methodology of rough set theory and artificial neural network for business failure prediction". *Expert Systems with Applications*, v. 18, Issue 2, February 2000, p. 65-74.

AGARWAL, VINEET; TAFFLER, RICHARD. "Comparing the performance of market-based and accounting-based bankruptcy prediction models". *Journal of Banking & Finance*, v. 32, Issue 8, August 2008, p. 1541-1551.

ALTMAN, E.I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy". *Journal of Finance*, v. 23, 1968, p. 589-609.

ALTMAN, E. I.; HALDEMAN, R.G.; NARAYANAN, P. "Zeta Analysis: A new model to identify bankruptcy risk of corporations", *Journal of Banking and Finance*, v. 1, 1977, p. 29–54.

ALTMAN, Edward I; BAIDYA, Tara K. N.; DIAS, Luiz Manoel Ribeiro. "Previsão de problemas financeiros em empresas." *Revista de Administração de Empresas*, v. 19, jan./mar., 1979, p. 17-28.

ALTMAN, Edward I. "An emerging market credit scoring system for corporate bonds". *Emerging Markets Review*, v. 6, Issue 4, December 2005, p. 311-323.

ALTMAN, Edward I.; GIANCARLO, Marco; FRANCO, Varetto. "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)". *Journal of Banking & Finance*, v. 18, Issue 3, may. 1994, p. 505-529.

ARANAZ, Magdalena Ferran. *SPSS para Windows – Programacion y análisis estadístico*. Madri: McGraw-Hill, 1996.

ATIYA Amir. F. "Bankruptcy prediction for credit risk using neural network: a survey and new results". *IEEE transactions on neural networks*, v. 12 nº 4, July 2001.

BALCAEN, Sofie; OOGHE, Hubert. "35 Years of studies on business failure: on overview of the classical statistical methodologies and their related problems". *The British Accounting Review*, v. 38, Issue 1, March, 2006, p. 63-93.

BATISTA, G.E.; PRATI, R.C.; MONARD, M.C.; "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data**".** *SIGKDD Explorations*, v. 6 Issue 1, 2004, p. 20–29.

BELHAOUI, A. R. *Accounting Theory*. 4. Ed. Londres: Thomson Learning, 2000.

BHARATH, S.; SHUMWAY, T. “Forecasting default with the KMV-Merton model”. *Working paper, University of Michigan*, 2004.

BHAVANI, Raskutti; ADAM, Kowalczyk. “Extreme re-balancing for SVMs: a case study”. *SIGKDD Explorations* 6(1): 2004, p. 60-69.

BLACK, F.; SCHOLES, M., “The pricing of options and corporate liabilities”**.** *Journal of Political Economy***,** v. 7, 1973, p. 637-654.

BLÖCHLINGER, A., LEIPPOLD, M.,. “Economic benefit of powerful credit scoring”**.** *Journal of Banking and Finance,* v. 30, 2006, p. 851-873.

BODIE, Zvi. *Finanças*. Porto Alegre: Editora Bookman Companhia. 2001.

BORITZ J. Efrim Kennedy Duane B. “Effectiveness of neural network types for prediction of business failure”**.** *Expert Systems with Applications,* v. 9, Issue 4, 1995, p. 503-512.

BRAGA-NETO, U.; HASHIMOTO, R.; DOUGHERT, Edward R. Nguyen, DANH V.; CARROLL, Rymond J. “Is cross-validation better than resubstitution for ranking genes?” Vol. 20 nº 2, 2004, p. 253-258. *DOI: 10.1093/bioinformatics/btg399*.

BRAGANÇA, Luiz Augusto de; BRAGANÇA, Sérgio Luiz de. **“**Rating” previsão de concordatas e falências no Brasil”. *VII Congresso ABAMEC*/1984.

BREALEY, Richard; MYERS, Stewart C. *Princípios de finanças empresariais.* Editora Mcgraw-Hill de Portugal, 1996.

CANBAS S, A.; CABUK, S.B.; KILIC, “Prediction of commercial bank failure via multivariate statistical analysis of financial structure: The Turkish case”, *European Journal of Operational Research*, v. 166, 2005, p. 528–546

CAOUETTE, John B.; ALTMAN, Edward I.; NARAYANAN, P. *Gestão do risco de crédito: o próximo grande desafio financeiro*. Rio de Janeiro: Ed. Qualitymark., 1999.

CARMO, M. E. M. *A concordata das companhias de capital aberto: um estudo preditivo utilizando modelos de análise fatorial.* Dissertação de Mestrado, Departamento de administração, PUC-RIO. Rio de Janeiro: PUC, 1987.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. **“**SMOTE: Synthetic minority over-sampling technique”. *Journal of Artificial Intelligence Research*, v. 16, 2002, p. 321-357.

CHAWLA, N. V.; JAPKOWICZ, N. KOLCZ, A. “Editorial: Special issue on learning from imbalanced data sets”**.** *SIGKDD Explorations*, v. 6, Issue 1, 2004, p. 1-6.

CHEN, Hsueh-Ju; HUAND, Shaio Yan; LIN, Chi-Shie. “Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach”**.** *Expert Systems with Applications,* v. 36, Issue 4, may. 2009, p. 7710-7720.

CHEN, You-Shyang; CHING-HSUE, Cheng. “Evaluating industry performance using extracted RGR rules based on feature selection and rough sets classifier”. *Expert Systems with Applications,* v. 36, Issue 5, july 2009, p. 9448-9456.

CHEN,Wei-Sen; DU, Yin-Kuan. “Using neural networks and data mining techniques for the financial distress prediction model”. *Expert Systems with Applications,* v. 36, Issue 2, Part 2, march 2009, p. 4075-4086.

CHEN, Wun-Hwa; SHIH, Jen-Ying. “A study of Taiwan's issuer credit rating systems using support vector machines”. *Expert Systems with Applications*, v. 30, Issue 3, april 2006, p. 427-435.

CHIEN-Chiang Lee; JUN-DE, Lee; CHI-CHUAN, Lee. “Stock prices and the efficient market hypothesis: Evidence from a panel stationary test with structural breaks”**.** *Japan and the World Economy*, v. 22, Issue 1, january 2010, p. 49-58.

CHYE, Koh Hian; CHIN Tan We. “Credit scoring using data mining techniques”. *Singapore Management Review*, v. 26, Nº 2, july 2004.

CLARK P.; MATWIN, S. “Using qualitative models to guide induction learning”, *Proceedings of ICML-93,1993 p*. 49-56.

CORNETT, M. M. ;MARCUS, A. J.;SAUNDERS, A.;TEHRANIAN, H., “The impact of institutional ownership on corporate operating performance”**.** *Journal of Bank & Finance***,** Volume 31, Issue 6, 2007, p. 1771-1794.

CRISTIANINI N.; SHAWE-TAYLOR J. “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods**”**, *Cambridge University Press*, Cambridge, 2000.

DING, Yongsheng; SONG, Xinping; ZEN, Yueming. “Forecasting financial condition of Chinese listed companies based on support vector machine”. *Expert Systems with Applications,* Volume 34, Issue 4, may. 2008, p. 3081-3089.

DIETRICH J. R.; KAPLAN, R.S. “Empirical analysis of the loan classification decision”, *The Accounting Review*, Volume 57, Issue 1, jan., 1982, p. 18–38.

DAMODARAN, Aswath. *Finanças corporativas aplicadas*. Tradução Jorge Ritter. Porto Alegre: Bookman, 2002.

DASH, M.; LIU, H. “Consistency-based search in feature selection”. *Artificial Intelligence,* Volume 151 n.1-2, p.155-176, december 2003.

DEVROYE, L., GYORFI, L.; LUGOSI, G. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York 1996.

DIMITRAS A. I.;SLOWINSKI R.; SUSMAGA, R. Zopounidis, C. “Business failure prediction using rough sets”**.** *European Journal of Operational Research*, v. 114, Issue 2, 16 april 1999, p. 263-280.

DOMINGOS, P. “Metacost: A general methods for makimg classifiers cost-sensitive”. *In: proceeding of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, 1999. pp. 155-164 San Diego, CA. ACM Press.

DULIBA, K.. “Contrasting neural nets with regression in predicting performance”. In: *Proceedings of the 24th Hawaii international conference on system sciences, v., IEEE Press*, Alamitos, CA 1991, p. 63–170.

ELIZABETSKY, Roberto. *Um modelo matemático para decisão no banco comercial.* (Trabalho apresentado ao Departamento de Engenharia de Produção da Escola Politécnica da USP). São Paulo: USP, 1976.

ELLIOT, B.; ELLIOT, J. *Financial accounting and reporting*. 6ª. ed. Prentice Hall Europe, 2002.

ESTABROOKS, Andrew; JO, Taecho; JAPKOWICZ, Nathalie. “A Multiple resampling methods for learning from imbalanced data sets”. *Computacional Intelligence*, v. 20, number 1, february 2004, p.18-36..

ETEMADI, Hossein; ROSTAMY, ALI Asghar Anvary; DEHKORDI, Hassan Farajzadeh. “A genetic programming model for bankruptcy prediction: Empirical evidence from Iran”. *Expert Systems with Applications*, v. 36, Issue 2, Part 2, march 2009, p. 3199-3207.

FAMA, Eugene F. “Efficient capital markets: a review of theory and empirical work”. *Journal of Finance*, v. 25, 1970, p.383-417.

FAMA, Eugene F. “The information in the term structure”. *Journal of Financial Economics*, v. 13 Issue 4, 1984, p.509-528.

FAYYAD U., G.; PIATESKY-SHAPIRO; SMYTH P. “From data mining to knowledge discovery in databases”. *AI Magazine*, v. 17, Issue 3, 1996, p. 37-54.

FIORAMANTI, Marco. “Predicting sovereign debt crises using artificial neural networks: A comparative approach”. *Journal of Financial Stability*, v 4, Issue 2, june 2008, p. 149-164.

FLETCHER D.; GOSS E. “Forecasting with neural networks: An application using bankruptcy data”, *Information and Management*, v. 24 1993, p. 159–167.

FRANK, Eibe; WITTEN, Ian H. “Generating Accurate Rule Sets Without Global Optimization”. In: *Proceedings of the Fifteenth international Conference on Machine Learning* (July 24 - 27, 1998). J. W. Shavlik, Ed. Morgan Kaufmann Publishers, San Francisco, CA, 144-151.

FREITAS A. A. *Data mining and knowlwdge discovery with evolutionary algorithms.* Springer-Verlag Berlin Heidelberg New York, 1998.

FRYDMAN, Halina; EDWARD, Altman I.; KAO, Duen-Li. "Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress". *The Journal of Finance,* v. 40, nº 1, mar., 1985, p. 269-291.

GARY M. Weiss; KATE Mccarthy; BIBI, Zabar. "Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?", In: *Proceedings of the 2007 International Conference on Data Mining*, CSREA Press, 35-41.

GARY M. Weiss. "Mining with Rarity: A Unifying Framework", *SIGKDD Explorations*, v. 6, Issue 1, 2004, p.7-19.

GUOZHONG AN, "The Effects of Adding Noise During Backpropagation Training on a Generalization Performances", *Neural Computation*, v.8, p. 643-674, 1996.

HA, T. M.; BUNKE, H. "Off-line, Handwritten Numeral Recognition by Pertubation Method". *Pattern Analysis and Machine Intelligence*, v.19, may. 1997, p. 535-539.

HALL, Mark A.; HOLME, Geoffrey. "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", *IEEE transactions on knowledge and data engineering*, VOL. 15, NO. 3, MAY/JUNE 2003, p. 1437-1447.

HAN J.; KAMBER, M. *Data mining: concepts and techniques*. Morgan Kaufmann, San Francisco, 2ª edition, 2006.

HSIAO, Cheng. *Analysis of painel data*. Cambridge: Cambridge University Press, 1986.

HASLEM J.; SCHERAGA, A.; BEDINGFIELD, C.A.; JAMES, P. "An Analysis of the Foreign and Domestic Balance Sheet Strategies of the U.S. Banks and Their Association to Profitability Performance", *Management International Review*. First Quarter, Wiesbaden, 1992.

HAUGEN, Robert A.; BAKER, Nardin L.; CASE Closed (November 20, 2008). *The handbook of portfolio construction: contemporary applications of Markowitz techniques*, John B. Guerard Jr., ed., Forthcoming. Available at SSRN: Disponível em: <http://ssrn.com/abstract=1306523>. Acesso em: 21 agos. 2006.

HAYDEN E., 2003, *Are credit scoring models sensitive with respect to default definitions? Evidence from the Austrian Market.* Dissertation Paper, Department of Business Administration, University of Vienna, Austria, p. 1-43.

HAYKIN, Simon. *Redes neurais: Princípios e prática*. Trad. Paulo Martins Engel. – 2. ed. – Porto Alegre: Bookman, 2001.

HILLEGEIST, S.; KEATING, E.; CRAM, D.; LUNDTEDT, K. "Assessing the probability of bankruptcy". *Review of Accounting Studies*, v. 9, p. 5-34, 2004.

HOLTE, R. C.; ACKER, L.E.; PORTER, E.B.W. “Concept Learning and the Problem of Small Disjuncts”, In: *Proceeding of the Eleventh Joint International Conference on Artificial Intelligence*, p. 813-818, 1989.

HORNGREN, Charles; HARRISON, Walter T.; BAMBER, Linda Smith. *Accounting* – 4 ed. Prentice Hall, New Jersey, 1999.

HORTA, Rui Américo Mathiasi. *Utilização de indicadores contábeis na previsão de insolvência: Análise empírica de uma amostra de empresas comerciais e industriais brasileiras.* Dissertação apresentada no programa de mestrado em Ciência Contábeis da Universidade Estadual do Rio de Janeiro em 2001.

HUA, Zhongsheng; WANG, Yu; XU, Xiannoyan; ZHANG, Bin; LIANG, Liang. “Predicting corporate financial distress based on integration of support vector machine and logistic regression”. *Expert Systems with Applications*, v. 33, Issue 2, aug. 2007, p. 434-440.

HUANG. Zan Chen; HSINCHUN, Hsu; CHIA-JUNG, Chen, WUN-HWA; WU, Soushan. “Credit rating analysis with support vector machines and neural networks: a market comparative study”. *Decision Support Systems*, v. 37, Issue 4, sep. 2004, p. 543-558.

HUANG., C. L., CHEN M. C., WANG., C.J. “Credit scoring with a data mining approach based on support vector machines”. *Expert Systems with Applications*, v. 33, Issue 4, nov. 2007, p. 847-856.

HUNG, Chihli; CHEN, Jing-Hong. “A selective ensemble based on expected probabilities for bankruptcy prediction”. *Expert systems with applications*, 2009, v. 36, Issue 3, apr. 2009, p. 3297-5309.

IUDÍCIBUS, Sérgio de (Org.). *Manual de contabilidade das sociedades por ações: aplicável às demais sociedades* – 2. Reimp. – São Paulo: Atlas, 2007.

IUDÍCIBUS, Sérgio de. *Análise de Balanços*. 9ª Ed. São Paulo: Atlas, 2008.

JAPKOWICZ N.; STEPHEN, S., “The Class Imbalance Problem: A Systematic Study”. *Intelligent Data Analysis*, v. 6, Number 5, p. 429-450, nov. 2002.

JAPKOWICZ, N. “The class imbalance problem: Significance and strategies”**.** Proceedings of the 2000. *International Conference on Artificial Intelligence* (IC-AI'2000), v. 1, p. 111-117.

JO, T.; JAPKOWICZ N. “Class Imbalances versus Small Disjuncts”, *SIGKDD Explorations* 6(1), June 2004, p. 40-49.

JONES S.; HENSHER, D.A. “Predicting firm financial distress: A mixed logit model”, *Accounting Review*, v. 79, Issue 4, 2004, p. 1011–1038.

KANITZ, Stephen Charles. *Como prever falências.* São Paulo: Mc Graw-Hill do Brasil, 1978.

KASZNAR, Istvan Karoly. *Falências e Concordatas de Empresas: modelos teóricos e estudos empíricos.* Dissertação de Mestrado submetido à congregação da Escola de Pós-Graduação em Economia (EPGE) do Instituto de Economia – FGV/RJ. Rio de Janeiro: FGV, 1986.

KARELS G. V.; PRAKASH A J., "Multivariate normality and forecasting of business bankruptcy", *Journal of Business Finance and Accounting*, v. 14, Issue 4, 1987.

KEASEY K.; WATSON R., 1991, *Financial distress models: a review of their usefulness.* British journal of Management, Vol. 2, n. 2, july 1991, p. 89-102.

KIM, Hong Sik; SOHN, So Young. "Support vector machines for default prediction of SMEs based on technology credit". *European Journal of Operational Research*, In Press, Corrected Proof, Available online 1 april 2009.

KIM, Myoung-Jong; Han, Ingoo. "The discovery of experts decision rules from qualitative bankruptcy data using genetic algorithms". *Expert Systems with Applications*, v. 25, Issue 4, nov. 2003, p. 637-646.

KOLARI J.; GLENNON, D.; SHIN, H.; CAPUTO, M. "Predicting large US commercial bank failures", *Journal of Economics and Business*, 54 (32 1) (2002) 361–387.

KRIEGEL H-P.; BORGWARDT K. M.; KRÖGER, P.; PRYAKHIN, A.; SCHUBERT M.; ZIMEK A., "Future trendes in data mining". *Data Mining and Knowledge Discovery*, v. 15, Number 1 / august, 2007.

KOHAVI, R..; JOHN, G. H. "Wrappers for feature subset selection". *Artif. Intell., v.*97, 1997. p.273-324.

KUBAT, M.; MATWIN, Stan. "Addressing the curse of imbalanced training sets: One-sided selection". In: *Proceedings of the Forteenth International Conference on Machine Leaning*, 1997, pp. 179-186. Nashiville, Tenesse. Morgan Kaufmann.

KUMAR, P. RAVI; RAVI, V. "Bankruptcy prediction in Banks and firms via statistical and intelligent techniques – A review". *European Journal of Operational research*, v. 180, 2007, p. 1-28.

LACHERR. C.; COATS, P. K.; SHARMA, S.C.; FANT, L. F. "A neural network for classifying the financial health of a firm", *European Journal of Operations Research*, v. 85, 1995, p. 53-65.

LAITINEN, E.K. "Traditional versus operating cash flow in failure prediction". *Journal of Business Finance and Accounting*, v. 21, n. 2, march 1994, p. 195-217.

LAITINEN T.; KANKAANPÄÄ M. “Comparative analysis of failure prediction methods: the Finnish case”. *The European Accounting Review*, v. 8, n. 1, 1999. p. 67-92.

LAM, L.; SUEN, C. Y. “Application of majority voting to pattern recognition: an analysis of its behavior and performance”. *IEEE Transactions on Systems, Man and Cybernetics* – Part A, 27(5), 1997, p.553–568.

LEE, Young-Chan. “Application of support vector machines to corporate credit rating prediction”. *Expert Systems with Applications*, v. 33, Issue 1, july 2007, p. 67-74.

LEE, H. D.; MONARD, M. C.; BARANAUSKAS, J. A. *Empirical Comparasion of Wrapper and Filter Approaches for Feature Subset Selection*. Technical report 94, ICMC-USP. Disponível em: <ftp//ftp.icmc.usp.br/pub/BIBLIOTECA/Rt_94>. Acesso em: 21 de ago. 2008.

LEE, Kun Chang; HAN, Ingoo; KWON, Youngsig. “Hybrid neural network models for bankruptcy predictions”. *Decision Support Systems*, v. 18, Issue 1, sep. 1996, p. 63-72.

LEE, Tian-Shyug; CHEN, I-Fei. “A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines”. *Expert Systems with Applications*, v. 28, Issue 4, may. 2005, p. 743-752.

LENSBERG, T.; EILIFSEN, A.; MCKEE, T. E. “Bankruptcy theory development and classification via genetic programming”. *European Journal of Operational Research*, v.169, jun. 2008, p.677–697.

LI HUI, JIE SUN. “Majority voting combination of multiple case-based reasoning for financial distress prediction”. *Expert Systems with Applications*, v.36, apr. 2009, p. 4363-4373.

LIU B.; W. HSU. “Post-analysis of learned rules,” *AAAI-96*, p. 828-834, 1996.

LIU B. M.; HU A.; HSU, W. "Multi-level organization and summarization of the discovered rules," *KDD-2000*, 2000.

LIU, H. ; SETIONO, R. (1996). “A probabilistic approach to feature selection – a filter solution”. In: *Proc. of the 13th Int. Conf. on Machine Learning*, pages 319-327, Bari, Italy.

LIU B.; HSU, W.; CHEN, S. “Analyzing discovered classification rules using general impressions,” *KDD-97*, 1997.

LIU B.; HSU, W; MA, Y. “Pruning and summarizing the discovered associations.” *KDD-99*, p. 125-134, 1999.

LIU, H.; MOTODA, H. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Massachusetts, 1998.

LOPES, Alexsandro Broedel; MARTINS, Eliseu. *Teoria da contabilidade: uma nova abordagem* – 2ª. Reimpressão – São Paulo: Atlas, 2007.

MAHONEY J. J.; MOONEY, R. J. **"**Comparing methods for refining certainty-factor rule bases," *Proceedings of ICML-94*, pp. 173-180, 1994.

MALOOF., Marcos. A. "Learning when data sets are imbalanced and when costs are unequal and Unknown". *Workshop on Learning from Imbalanced Data Sets II, ICML,* Washington C, 2003.

MANLY, Bryan J. F. *Métodos Estatísticos Multivariados – Uma Introdução*. 3ª edição - Porto Alegre: Artmed Editora S.A., 2005.

MARTENS, D.; BAESENS, B. GESTEL; VANTHIENEN, T. V. "Comprehensible credit scoring models using rule extraction from support vector machines". *European Journal of Operational Research*, v. 183, 2007, p. 1466-1476.

MARTIN, D. "Early warning of bank failure: A logit regression approach", *Journal of Banking and Finance*, v.1, 1977, p. 249–276.

MATARAZZO, Dante Carmine. *Análise financeira de balanços: abordagem básica e gerencial.* 6ª Ed. São Paulo: Atlas, 2003.

MATIAS, Alberto Borges. *Contribuição às técnicas de análise financeira: um modelo de concessão de crédito.* (Trabalho apresentado ao Departamento de Administração da Faculdade de Economia e Administração da USP.) São Paulo: [s.n.], 1978, p. 82, 83, 90.

MATLAB. http://www.mathworks.com.

MCCARTHY, John. "A Basis for a Mathematical Theory of Computation"**.** *Studies in Logic and the Foundations of Mathematics,* v. 26, 1959, p. 33-70.

MC LEAY S., OMAR A., "The sensitivity of prediction models tot the non-normality of bounded an unbounded financial ratios". *British Accounting Review*, v. 32, 2000, p. 213-230.

MENSAH, Y.M.,. "Na examination of the stationarity of multivariate bankruptcy prediction models: A methodological study". *Journal of Accounting Reseach*, v. 22, 1984, p.380-395.

METZ, C. E. "Statistical Analysis of ROC Data in Evaluating Diagnostic Performance". In: *Multiple Regression Analysis: Applications in the health sciences.* Number 13, edited by Donald E. Herbert and Raymond H. Myers, 1986, p. 365-384. American Institute of Physics.

MIN, Jae H. Lee; YOUNG-CHAN. "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters". *Expert Systems with Applications,* v. 28, Issue 4, may. 2005, p. 603-614.

MIN,Sung-Hwan.; LEE, Jumin,;HAN. Ingoo. “Hybrid genetic algorithms and support vector machines for bankruptcy prediction”. *Expert Systems with Applications,* v. 31, Issue 3, oct. 2006, p. 652-660.

MOONEY R. J. **“**Induction over the unexplained: using overly-general theories to aid concept learning,” *Machine Leaning*, 1993, v. 10, p. 79-110.

NANNI, Loris,; LUMINI, Alessandra. “An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring”. *Expert Systems with Applications,* v. 36, Issue 2, Part 2, mar. 2009, p. 3028-3033.

NWOGUGU, Michel. “Decision-making, risk and corporate governance: New dynamic models/algorithms and optimization for bankruptcy decisions”. *Applied Mathematics and Computation*, v. 179, Issue 1, 1 aug. 2006, p. 386-401.

ODOM, M.; SHARDA, R. “A neural network model for bankruptcy prediction”**.** In: *Proceedings of the international joint conference on neural networks*, Vol. 2, IEEE Press, Alamitos, CA, 1990, p. 163–168.

OH, S.B. “On the relationship between majority vote accuracy and dependency in multiple classifier systems”. Pattern Recognition Letters, 2003, 24,359–363.

OHLSON, J.A. “Financial ratios and the probabilisitic prediction of bankruptcy”. *Journal of Accounting Research*, v. 18, 1980, p.109-131.

OLSEN, J.P., 1996, “Restructuring of Distressed Bank Debt: Some Empirical Evidence from the UK,” *IFA Friday Workshop Paper.*

ORTEGA J.; D. FISHER. **“**Flexibly exploiting prior knowledge in empirical learning,” In: *Proceedings of IJCAI-95*, p. 1041-1047, 1995.

PADMANABHAN B.; TUZHILIN, A. **“**A belief-driven method for discoveringunexpected patterns”. *KDD-98*, 1998, p. 74-100.

PADMANABHAN B.; TUZHILIN, A. “Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns”. *KDD-2000*, 2000.

PARK, Cheol-Soo; HAN, Ingoo. “A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction”. *Expert Systems with Applications,* v.23, Issue 3, oct. 2002, p. 255-264.

PAZZANI M.; KIBLER, D. “The utility of knowledge in inductive learning,” *Machine learning*, v. 9, p. 57-94, 1992.

PENDHARKAR, Parag C. “A threshold-varying artificial neural network approach for classification and its application to bankruptcy prediction problem”. *Computers & Operations Research,* v. 32, Issue 10, oct. 2005, p. 2561-2582.

PIATESKY-SHAPIRO G. C. MATHEUS; SMYTH, P.; UTHURUSAMY, R. “*KDD-93: progress and challenges”, AI magazine,* Fall (1994), p. 77-87, 1994.

PIATESKY-SHAPIRO G.; MATHEUS, C. "The interestingness of deviations," *KDD-94***,** p. 25-36, 1994.

PILA, A. D. *Seleção de Atributos relevantes para Aprendizado de Máquina utilizando a Abordagem de Rough Sets*. Dissertação de mestrado, ICMC-USP. Disponível em: http://www.teses.usp.br/teses/disponíveis/55/55134/tde-13022002-153921/publicadissertacao_ADP.pdf.Acesso em: 19 abr. 2006.

PINDYCK, Robert S.; RUBINFELD, Daniel L. *Econometria – Modelos & Previsões.* Rio de Janeiro: Elsevier, 2004.

PIRAMUTHU, Selwyn. "Evaluation feature selection methods for learning in data mining applications"**.** *European Journal of Operational Research*, v. 156, 2004, p. 483-494.

PIRAMUTHU S. "On preprocessing data for financial credit risk evaluation". *Expert Systems with Applications***,** v. 30, 2006, p.489-497.

PLATT H.D.; PLATT, M.B.,. "Predicting corporate financial distress: reflections on choise-based sample bias". *Journal of Economics and Finance*, v. 26, nr.2, Summer 2002, p. 184-199.

POLIKAR, Robi. "Ensemble Based Systems in Decision Making". *IEEE Circuits and Systems Magazine.* Third Quarter. 2006.

PREMACHANDRA, I.M.; BHABRA, Gurmeet Singh, Sueyoshi Toshiyuki. "DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique". *European Journal of Operational Research, v. 193, Issue 2*, 1 mar. 2009, p. 412-424.

PROVOST, F.; FAWCETT, T. "Robust classification for imprecise environments**"**. *Machine Learning*, v.42, 2001, p.203-231.

PYLE D. *Data preparation for data mining.* Morgan Kaufmann Publishers Inc. 1999.

QUILAN J. R., *Induction of decision trees*, Machine Learning, v.1, 1986, v. 81-106.

QUILAN J. R. *C4.5: program for machine learning*. Morgan Kaufmann, 1992.

QUILAN J. R., *Programs for Machine Learning*, Morgan Kaufman Publishers Inc., San Francisco, CA, USA, 1993.

RAVI, V.; KURNIAWAN, H.; THAI, Peter Nwee Kok,; KUMAR, P. Ravi. "Soft computing system for bank performance predition". *Applied Soft Computing*, v. 8, jan. 2008, p.305-315.

REISZ, Alexander S., PERLICH C. "A market-based framework for bankruptcy prediction". *Journal of Financial Stability,* v. 3, Issue 2, jul. 2007, p. 85-131.

REZENDE, Solange Oliveira (Org.). *Sistemas inteligentes: fundamentos e aplicações.* Barueri, SP: Manole, 2005.

ROSS, Stephen A.; WESTERFIELD, Randolph, W. Jaffe,; JEFFREY F. *Administração Financeira – Corporate Finance* – São Paulo: Editora Atlas, 2ª edição, 2002.

ROY, J.; COSSET, C. "The determinants of country risk ratings", *Journal of International Business studies,* First Quarter (1990), p. 135–139.

RUSSEL, Stuar J.; NORVIG, Peter. *Inteligência Artificial*: tradução da segunda edição – Rio de Janeiro: Elsevier, 2004 – 2ª Reimpressão.

SALCEDO-SANZ, SANCHO. FERNÁNDEZ-VILLACAÑAS, JOSÉ LUIS. SEGOVIA-VARGAS, MARIA JESÚS. BOUSOÑO-CALZÓN, CARLO. "Genetic programming for the prediction of insolvency in non-life insurance companies". *Computers & Operations Research, v. 32, Issue 4*, apr. 2005, p. 749-765.

SALCHENBERGER L. M.; CINAR, E. M.; LASH, A. "Neural networks: A new tool for predicting thrift failures*", Decision Sciences,* v. 23, 1992, p. 899–916.

SANTOS, Samuel Cruz Dos. *Um modelo de análise discriminante múltipla para previsão de inadimplência em empresa*. Dissertação de Mestrado, Departamento de Administração, PUC/RJ. Rio de Janeiro: PUC, 1996.

SAUDAGARAN, S. M. *International accounting: a user perspective*. 2. Ed. Cincinnati: South Western, 2004.

SAUNDERS, A. ALLEN; DELONG, G. L. "Issues in the credit risk modeling of retail markets"**.** *Journal of Bank & Finance,* v. 28, 2004.

SAUNDERS, A.; CORNETT M. M. "Financial Institutions Management: A Risk Management Approach"**.** *Paperback edition.* McGraw-Hill/Irwin, 2008.

SAUNDERS, ANTHONY. *Medindo o risco de crédito*. Rio de Janeiro: Qualitymark Editora, 2000.

SHARDA R.; WILSON, R. L. "Neural network experiments in business-failure forecasting: Predictive performance measurement issues, International". *Journal of Computational Intelligence and Organizations,* v. 1, Issue 2, 1996, p. 107-117.

SHEN, C.; LIU, B. *Generating Classification Rules According to User's Existing Knowledge,* 3 Science Drive 2, Singapore, 117543, 2001.

SHIN, Kyung-Shik; LEE, Yong-Joo; KIM, Hyun-Jung."An application of support vector machines in bankruptcy prediction model". *Expert Systems with Applications,* v. 28, Issue 1, jan. 2005, p. 127-135.

SHIN, Kyung-Shik; LEE, Yong-Joo. "A genetic algorithm application in bankruptcy prediction modeling". *Expert Systems with Applications, v.e 23, Issue 3*, oct 2002, p. 321-328.

SILVA, José Pereira da. *Gestão e análise de risco de crédito*. 5ª Ed. São Paulo: Atlas, 2006.

SILBERSCHATZ, A.; TUZHILIN, A. "What makes patterns interesting in knowledge discovery systems," *IEEE Transactions on Knowledge and Data Engineering*, v.8, no. 6, p. 970-974, 1996.

SOMOL P.; BAESENS B.; PUDIL P.; VANTHIENEN J., "Filter-versus Wrapper-based Feature Selection for Credit Scoring", *International Journal of Intelligent Systems*, v. 20, Number 10, 2005, p. 985-999.

STEIN, R. "The relationship between default prediction and lending profits: Integrating ROC analysis and loan princing". *Journal of Banking and Finance*, v. 29, 2005, p. 1213-1236.

SUN JIE, LI HUI. "Data mining method for listed companies' financial distress prediction". *Knowledge-Based Systems*, 2008, v. 21, p. 1-5.

SUN, JIE, LI, HUI. "Data mining method for listed companies' financial distress prediction". *Knowledge-Based Systems*, v. 21, 2008, p. 1-5.

TAFFLER, R.J., "Empirical models for the monitoring of UK corporations". *Journal of Banking and Finance*, v. 8, 1984, p.199-227.

TAFFLER R.J.; AGARWAL V. "Do statistical failure prediction models work ex ante or only ex post?" Paper read in: *the Deloitte & Touche Lecture Series on credit risk*, University of Antwerp, February 2003, Belgium.

TAM K. Y.; KIANG, M. Y. "Managerial applications of neural networks: The case of bank failure predictions", *Management Science*, v. 38, Issue 7, 1992, p. 926-947.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. 2006. *Introduction to Data Mining*. Pearson Education, Inc.Boston USA.

TAY Francis E. H.; SHEN, Lixiang. "Economic and financial prediction using rough sets model", *European Journal of Operational Research*, v. 141, Issue 3, sep. 2002, p. 641-659.

TSENG, Fang-mei; LIN, Lin. "A quadratic interval logit model for forecasting bankruptcy". *The International Journal of Management Science*, v. 33, Issue 1, feb. 2005, p. 85-91.

TIMMERMANN, Allan; GANGER, Clive, W. J. "Efficient market hypothesis and forecasting". *International Journal of Forecasting*, v. 20, 2004, p. 15-27.

TRIPPI R.R.; TURBAN, E. *Neural Networks in Finance and Investment: Using Artifcial Intelligence to Improve Real-World*. Performance, Probus, Chicago, IL, 1993.

TSAI., C. F.; WU J. W. "Using neural network ensembles for bankruptcy prediction and credit scoring". *Expert Systems with applications*, v. 34, Issue 4, may. 2008, p. 2639-2649.

TSAI., C. F., "Feature selection in bankruptcy prediction". *Knowledge-Based Systems*, v. 22, aug. 2008, p. 120–127.

VAPNIK,V.N. *Statistical Learning Theory*, John Wiley, New York, 1998.

VARETTO, Franco. "Genetic algorithms applications in the analysis of insolvency risk". *Journal of Banking & Finance*, v. 22, Issues 10-11, oct. 1998, p. 1421-1439.

XU, L., Krzyzak, A.; SUEN, C. Y. "Methods for combing multiple classifiers and their applications to handwriting recognition". *IEEE Transactions on Systems, Man and Cybernetics*, v. 22, Issue 3, may./jun.1992, p. 418–435.

WARD T.J.; FOSTER B.P., "A note on selecting a response measure for financial distress". *Journal of Business Finance and Accounting*, v. 24, nr. 6, jul.1997, p. 869-879.

WEST, R. C, "A factor analytic approach to bank condition", *Journal of Banking and Finance*, v. 9, jun.1985, p. 253–266.

WEST, David; DELLAN Scott; QIAN Jingxia. "Neural network ensemble strategies for financial decision applications". *Computers & operations research*, v. 32, 2005.

WESTON, J. FRED; BRIGHAM, EUGENE. F. *Fundamentos da Administração Financeira*. Editora Makron Books, São Paulo, 10ª edição, 2000.

WILCOX, J.W., "A prediction of business failure using accounting data, empirical research in accounting: Selected studies", *Journal of Accounting Research*, Suplement to v. 11, 1973, p. 163–179.

WILSON, R. L.; SHARDA, R. "Bankruptcy prediction using neural networks", *Decision Support Systems*, v. 11, jun. 1994, p. 545-557.

WITTEN, Ian .H.; FRANK, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems, 2ª ed. 2005.

WU, C.H.; FANG, W. C.; GOO, Y. J. *Variable selection method affects SVM approach in bankruptcy prediction*. Advances in intelligent Systems Research, 2006. Disponível em: http://www.atlantis-press.com/php/download_paper.php?id=114. Acesso em: 09/10/2008.

YOUN, Hyewon; ZHENG GUYU, L.; WAUNG, S.; LAI, K. K. “Predicting Korean lodging firm failures: An artificial neural network model along with a logistic regression model”. *International Journal of Hospitality Management*, Available online 26, jul. 2009.

YU, L. WAUNG; LAI, K. K. “Credit risk assessment with a multistage neural network ensemble learning approach”. *Expert Systems with Applications*, v. 34, fev. 2008, p. 1434-1444.

ZAHEDI F., “A meta-analysis of financial application of neural networks”, *International Journal of Computational Intelligence and Organizations*, v. 1, Issue 3, 1996, p. 164-178.

ZHANG, Guoqiang Zhang; MICHAEL Y. HU; EDDY, Patuwo W.; DANIEL C. Indro. “Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis**”**. *European Journal of Operational Research,* v. 116, jul. 1999, p. 16-32.

ZMIJEWSKI, M., “Methodological issues related to the estimation of financial distress prediction models”, *Journal of Accounting Research*, v. 22, Supplement. 1984, p. 59–82.