

Comparação de Técnicas de Seleção de Atributos para Previsão de Insolvência de Empresas Brasileiras no Período 2005-2007

Autoria: Rui Américo Mathias Horta, Frederico A. de Carvalho, Francisco José dos Santos Alves, Marcelino José Jorge

O objetivo deste estudo é comparar técnicas de seleção de atributos aplicados em previsão de insolvência empresas brasileiras não financeiras de capital aberto. Foram aplicadas três abordagens de seleção de atributos, sendo duas de aprendizado de máquina – Filtro e Wrapper – e uma de estatística multivariada – Análise de Componentes Principais. Posteriormente é feita uma avaliação da seleção dos atributos; para esta avaliação são empregados três algoritmos de classificação – Regressão Logística, Árvore de Decisão e Máquina de Vetor Suporte. A principal contribuição do artigo é destacar e considerar explicitamente a fase de seleção das variáveis preditivas através de abordagens com metodologias distintas. No levantamento do referencial teórico desta pesquisa, constatou-se a prática de utilizar variáveis previamente relacionadas em pesquisas anteriores (Kanitz, 1978; Pereira, 1982; Bragança e Bragança, 1984; Sanvicente e Minardi, 2000; Brito, Neto e Corrar, 2009). Este procedimento pode desconsiderar fatores culturais e institucionais, tais como os hábitos quanto a tomar crédito ou as legislações tributária, fiscal e societária em cada país. A partir do final da década de 90, questões tais como o aparecimento de novas técnicas de modelagem, a expansão dos mercados de capitais, os impactos dos mercados imperfeitos e das informações assimétricas e as constantes mudanças no ambiente econômico das empresas trouxeram de volta o interesse pela análise e previsão da insolvência de empresas, ocasionando inúmeras pesquisas no mundo (Altman, Marco e Varetto, 1994; Back, Laitinen e Kaisa, 1996; Eisenbeis, 1997; Lennox, 1999; Härdle, Moro e Schäfer, 2005; Wu, Fang e Goo, 2006).

Esta pesquisa empírica foi do tipo descritivo e quantitativo, compreendendo as empresas classificadas no SERASA e na BOVESPA como “solventes” e “insolventes” (“concordatárias”, “em recuperação judicial”, e “falidas”), e abrangendo o período de 2005 a 2007.

Os resultados indicam, para a amostra utilizada, que a abordagem wrapper é mais eficiente do que filtro e análise de componentes principais. Além disso, a abordagem wrapper obteve melhores classificações nas três técnicas - regressão logística (89,88%), árvore de decisão (93,45%) e máquina de vetor suporte (97,02%). A utilização de modelos preditivos de insolvência, construídos pela aplicação de Data Mining, é uma dentre várias formas de avaliar o risco de uma instituição sem depender, apenas, da avaliação subjetiva do analista. Na prática da análise de risco, e tendo em mente os recursos computacionais disponíveis, esses modelos preditivos podem ser incorporados como procedimentos analíticos para avaliar a probabilidade de insolvência. Neste sentido, então, tornam-se interessantes para bancos, investidores, governos, auditores, gerentes, fornecedores, empregados e muitos outros, para que possam avaliar, com razoável antecedência, se há ou haverá problemas de insolvência, em andamento ou futuros.

1. INTRODUÇÃO

Prever com exatidão se uma empresa vai tornar-se insolvente é talvez impossível, visto o próprio ambiente de incerteza do mercado e tantas outras situações adversas a que todas as empresas estão sujeitas. Entretanto, é possível identificar, com auxílio de modelos de previsão de insolvência, aquelas organizações com maiores probabilidades de falhar em futuro próximo, permitindo, assim, identificar e apontar medidas corretivas em tempo hábil.

Nos anos recentes, uma revolução de idéias e de novas tecnologias computacionais – tais como inteligência computacional, comitê de classificadores, seleção de atributos, balanceamento de bancos de dados, dentre outras – tem ampliado a maneira pela qual o risco de crédito é medido e gerido. Algumas outras justificativas podem ser acrescidas para explicar o interesse sobre esse assunto:

- (i) nas últimas décadas o ambiente econômico geral tem mudado com enorme velocidade e tem experimentado tendências para baixo (dificuldades financeiras);
- (ii) as margens das taxas de juros, ou *spreads*, têm-se tornado muito estreitas, ou seja, a compensação de risco-retorno advinda de empréstimos piorou;
- (iii) o aumento nas incertezas financeira em quase todas as economias mais desenvolvidas do mundo tem mostrado que valores de imóveis e de ativos físicos são muito difíceis de prever e de realizar através de liquidações. Quanto mais fracos e incertos forem os valores das garantias reais, mais arriscada se torna a concessão de empréstimos;
- (iv) empresas insolventes acabam gerando envolvimento em vários setores econômicos e grandes custos para todos, especialmente os sociais;
- (v) com a evolução da disponibilidade de dados e das técnicas computacionais há um impulso nas possibilidades de pesquisa que concernem à previsão de insolvência. Dentre essas novas técnicas computacionais, destaca-se *Data Mining* (DM), que se tornou proeminente no final da década de 90, e apresentou forte ênfase na combinação entre conjuntos de dados para capturar padrões que são muito sutis ou muito complexos para serem detectados somente por analistas de dados.

O objetivo deste estudo é comparar técnicas de seleção de atributos aplicados em previsão de insolvência de empresas brasileiras não financeiras de capital aberto. Foram aplicadas três abordagens de seleção de atributos, sendo duas de aprendizado de máquina – Filtro e *Wrapper* – e uma de estatística multivariada – Análise de Componentes Principais. Posteriormente é feita uma avaliação da seleção dos atributos; para esta avaliação são empregados três algoritmos de classificação – Regressão Logística, Árvore de Decisão e Máquina de Vetor Suporte.

A principal contribuição do artigo é destacar e considerar explicitamente a fase de seleção das variáveis preditivas através de abordagens com metodologias distintas. Alguns autores (Shirata, 2001; Wu, Fang e Goo, 2006; Piramuthu, 2006) têm chamado atenção para a importância do processo de seleção de atributos, que, na maior parte dos estudos sobre previsão de insolvência, nem sempre é claramente discutido, o que dificulta o entendimento sobre como se chegou às variáveis utilizadas. Em pesquisa recente utilizando dados brasileiros, Horta e Alves (2008), realizaram estudo com o mesmo objetivo, mas valendo-se de dados mais restritos, coletados para o período de 2004 a 2006. O presente trabalho, além de diferir pelo período estudado, aplica técnicas de classificação originadas da inteligência computacional, ensejando abordagens, comparações e conclusões distintas.

2. REFERENCIAL TEÓRICO

Os modelos de previsão de insolvência oferecem aos analistas e aos gestores de crédito uma ferramenta avançada, isenta de influências subjetivas e que lhes possibilita obter uma classificação quantitativa sobre o sucesso quanto à “saúde financeira” das empresas. Sua aplicabilidade, inicialmente, são as operações de curto prazo, considerando que a insolvência está mais relacionada à perda da capacidade de endividamento do que ao desempenho operacional.

Apesar de sua longa história na literatura especializada (Fitzpatrick, 1932; Winakor e Smith, 1935), o estudo do “insucesso” de empresas com base em indicadores obtidos a partir dos demonstrativos contábeis tomou ímpeto nos anos 70 (Deakin, 1972; Edmister, 1972; Kanitz, 1978), em seguida aos trabalhos pioneiros de Beaver (1967) e de Altman (1968).

No Brasil, a análise quantitativa da insolvência de empresas com objetivos preditivos desenvolveu-se de modo significativo a partir dos anos 80 (Pereira, 1982; Bragança e Bragança, 1984) prosseguindo então com Sanvicente e Minardi (2000), Horta (2001) e Horta e Alves (2008), entre outros.

No levantamento do referencial teórico desta pesquisa, constatou-se a prática de utilizar variáveis previamente relacionadas em pesquisas anteriores (Kanitz, 1978; Pereira, 1982; Bragança e Bragança, 1984; Sanvicente e Minardi, 2000, Brito, Neto e Corrar, 2009). Este procedimento pode desconsiderar fatores culturais e institucionais, tais como os hábitos quanto a tomar crédito ou as legislações tributária, fiscal e societária em cada país.

A partir do final da década de 90, questões tais como o aparecimento de novas técnicas de modelagem, a expansão dos mercados de capitais, os impactos dos mercados imperfeitos e das informações assimétricas e as constantes mudanças no ambiente econômico das empresas trouxeram de volta o interesse pela análise e previsão da insolvência de empresas, ocasionando inúmeras pesquisas no mundo (Altman, Marco e Varetto, 1994; Back, Laitinen e Kaisa, 1996; Eisenbeis, 1997; Lennox, 1999; Härdle, Moro e Schäfer, 2005; Wu, Fang e Goo, 2006).

Apesar de as novas técnicas de modelagem se terem tornado mais acessíveis, existem vários problemas, relatados na literatura específica, no que se refere à aplicação de alguns desses métodos, especialmente os mais tradicionais, na previsão de insolvência (Balcaen e Ooghe, 2006). Alguns desses problemas pertencem a tópicos tais como a dicotomia da variável dependente, a seletividade da amostra, a não-estacionaridade e os dados instáveis nas séries disponíveis, o uso de informações contábeis anuais; a seleção das variáveis independentes e a dimensão temporal.

Com o uso de técnicas de DM alguns desses problemas começaram a reduzir-se, conforme será detalhado nas seções seguintes.

2.1 Data mining (DM)

A concepção de DM ainda largamente aceita por diversos pesquisadores na área foi inspirada por Fayyad, Piatetsky-Shapiro e Smyth (1996) ao afirmarem que: “Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados”.

O processo de identificação de padrões em DM é dividido em três grandes etapas, conforme Rezende (2005): pré-processamento, extração de padrões e pós-processamento.

Neste estudo é dada ênfase a uma das etapas do pré-processamento, o problema da seletividade das variáveis independentes utilizando algumas das técnicas de DM.

2.1.1 Seleção de atributos

A seleção de atributos constitui uma etapa de fundamental importância em DM, sendo frequentemente entendida e realizada como uma etapa de pré-processamento. Os objetivos da seleção de atributos para previsão de insolvência são: (i) desenvolvimento de modelos compactos, (ii) uso e refinamento do modelo de classificação ou predição na avaliação e (iii) identificação de índices financeiros relevantes (Piramuthu, 2006).

Os algoritmos usados para seleção de atributos podem ser separados em duas atividades principais: busca do subconjunto de atributos e avaliação dos subconjuntos de atributos encontrados, tal como pode ser visto na Figura 1.

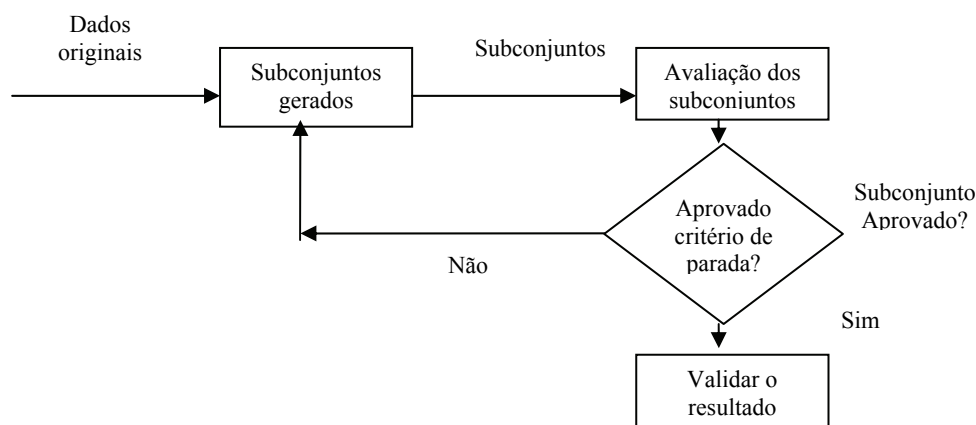


Figura 1. - Passos na Seleção de Atributos.
Fonte: Liu e Motoda, 1998.

Na primeira atividade – buscar um subconjunto de atributos - seleciona-se um subconjunto de variáveis relevantes com o apoio de um algoritmo de busca. Na segunda - avaliar o subconjunto de atributos selecionados – mede-se quão bom é um determinado atributo dentro do subconjunto selecionado segundo um critério de avaliação. No presente trabalho foi utilizado o critério da consistência, ou seja, 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).

2.1.1.1 Abordagem Filtro

A abordagem filtro introduz um processo de separação, que ocorre antes da aplicação do algoritmo de aprendizagem propriamente dito. Em outras palavras, a idéia é separar (“filtrar”) atributos irrelevantes, segundo algum critério, antes do aprendizado ocorrer, como ilustra a Figura 2.

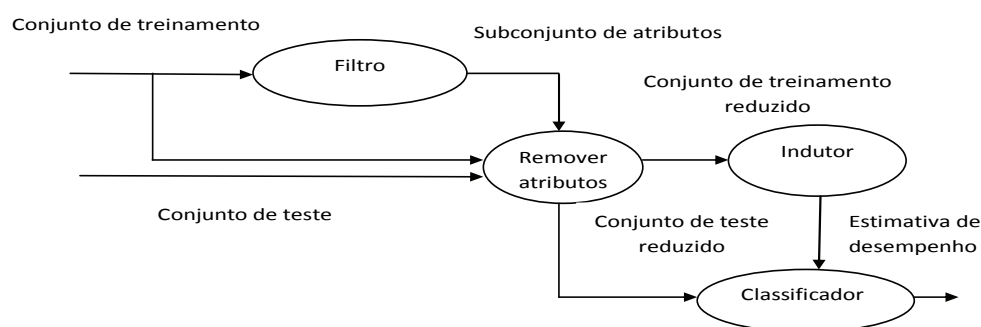


Figura 2 – Abordagem Filtro
Fonte: Freitas, 1998.

A abordagem filtro é independente do algoritmo de aprendizado, que receberá somente o subconjunto de atributos relevantes (“importantes”) selecionados pelo filtro. Nesta abordagem a meta é selecionar um subconjunto de atributos que preserve a informação pertinente ao conjunto completo de atributos (Freitas, 1998). Os algoritmos de busca do melhor subconjunto de atributos utilizados na abordagem filtro foram: *Genetic Selection* (GS) e outro *GreedyStepwise* (SP).

GS é realizado através de algoritmos genéticos (AGs), que são algoritmos de otimização global baseados em analogias aos mecanismos de seleção natural e da genética. Eles empregam uma estratégia de busca paralela e estruturada, embora aleatória, direcionada à busca de pontos de “alta aptidão”, ou seja, pontos em que a função a ser otimizada (minimizada ou maximizada) tem valores relativamente baixos ou altos. Apesar de aleatórios, os AGs não são buscas aleatórias não-direcionadas, pois exploram informações passadas (“históricas”) para encontrar novos pontos de busca para os quais são esperados melhores desempenhos. Embora possam parecer simplistas do ponto de vista metafórico, esses algoritmos são suficientemente complexos para fornecer mecanismos poderosos e robustos de busca adaptativa (Rezende, 2005).

O método SP seleciona as variáveis passo a passo, admitindo que uma variável selecionada em uma etapa possa ser eliminada em etapa posterior. Outra vantagem deste método é que, no processo de seleção, pode-se começar a partir de qualquer equação construída com base em um particular subconjunto de variáveis independentes, incluindo o que contém todas elas (Witten e Frank, 2005).

O algoritmo de avaliação de atributos na seleção aqui utilizado foi o CFS (seleção de atributos baseada em correlação). O CFS (Hall e Holme, 2002) utiliza a seguinte métrica para fazer a avaliação:

$$U(X, Y) = 2.0 * \left[\frac{H(X) + H(Y) - H(X, Y)}{H(X) + H(Y)} \right]. \quad (1)$$

Depois de calcular uma matriz de correlação, o algoritmo CFS aplica uma estratégia heurística de busca para encontrar um bom subconjunto de atributos de acordo com o critério:

$$Merit_s = \frac{k \overline{r_{cf}}}{\sqrt{k + k(k-1)r_{ff}}}, \quad (2)$$

onde $H(X)$ e $H(Y)$ em (1) são os atributos; $Mérit_s$ é o “mérito” de um subconjunto S_k contendo k atributos; r_{cf} é a média da correlação entre classes, e r_{ff} , a média da correlação entre os atributos.

2.1.1.2 Abordagem *Wrapper*

A abordagem *wrapper* ocorre externamente ao algoritmo básico de aprendizagem, mas utiliza este algoritmo como uma espécie de “caixa preta” para analisar, a cada iteração, o subconjunto de atributos até ali selecionados. Em outras palavras, os métodos *wrapper* geram um subconjunto candidato de atributos selecionado do conjunto de treinamento, e utilizam a precisão resultante do algoritmo de aprendizado para avaliar o subconjunto de atributos em questão. Esse processo é repetido para cada subconjunto de atributos até que o critério de parada determinado pelo usuário seja satisfeito, como está indicado na Figura 3. Esta abordagem avalia os atributos usando estimativas de precisão providas por algoritmos de aprendizado pré-determinados (Freitas, 1998).

No presente estudo foram testados três algoritmos de aprendizagem: um com origem na Estatística (Regressão Logística - RL), outro com origem em regras de indução (Árvore de Decisão - AD) e o terceiro apoiado em inteligência computacional (Máquina de Vetor Suporte - MVS). Estes três algoritmos de aprendizado serão explicados nas seções posteriores.

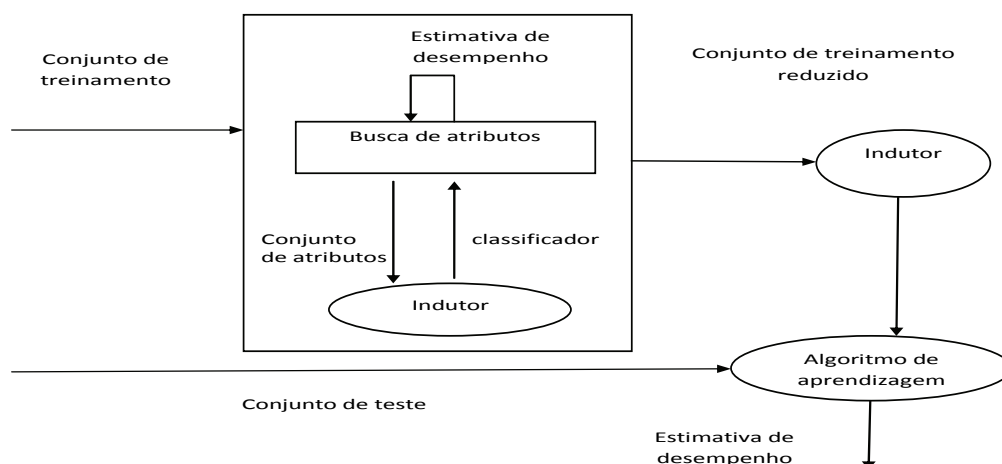


Figura 3 – Abordagem *Wrapper*
Fonte: Freitas, 1998

2.2.1.3 Análise de Componentes Principais (ACP)

A ACP é um dos métodos estatísticos multivariados (Manly, 2008). O objetivo da análise é encontrar combinações lineares de p variáveis X_1, X_2, \dots, X_p para produzir índices Z_1, Z_2, \dots, Z_k que sejam não correlacionados e que reproduzam totalmente a informação (variabilidade) presente nos dados. Em outros termos, a idéia básica do método é transformar p variáveis (tipicamente correlacionadas) em $k < p$ combinações lineares não correlacionadas.

3. METODOLOGIA

Esta pesquisa empírica foi do tipo descritivo e quantitativo, compreendendo as empresas classificadas no SERASA e na BOVESPA como “solventes” e “insolventes” (“concordatárias”, “em recuperação judicial”, e “falidas”), e abrangendo o período de 2005 a 2007.

3.1 Dados baseados no mercado *versus* dados baseados nos demonstrativos contábeis

Em finanças corporativas os estudos empíricos podem apoiar-se em dados contábeis, dados de mercado ou utilizando os dois conjuntamente, segundo a disponibilidade em cada caso. Em pesquisa recente, Agarwal e Taffler (2008) compararam as alternativas de utilizar os dados baseados no mercado *vis à vis* os dados baseados nos livros contábeis, concluindo que modelos de previsão de insolvência baseados em dados contábeis produzem benefícios econômicos significativamente maiores do que os produzidos pelos modelos baseados nos dados de mercado. Anteriormente Blöchliger e Leippold (2006) haviam observado que as diferenças nos erros são economicamente significativas a favor dos dados contábeis. Assim a alternativa escolhida para este trabalho foi a de utilizar dados corporativos baseados nos demonstrativos contábeis.

3.2 A amostra

Para o objetivo deste estudo são necessários dois tipos de amostras. O primeiro tipo refere-se a um conjunto de empresas chamadas de “problemáticas”, isto é, que apresentaram problemas de insolvência em um determinado período. Tal conjunto é aqui construído a partir de empresas classificadas, no SERASA ou na BOVESPA, como “concordatária”, “em recuperação judicial” ou “falida”. O segundo tipo diz respeito a um conjunto de empresas

“saudáveis” no sentido precisamente contrário ao do primeiro grupo, ou seja, empresas que não apresentaram problemas de insolvência no tempo estudado.

Na amostra do primeiro grupo, foram consideradas empresas chamadas de “problemáticas” no período de 2005 a 2007. Na amostra do segundo grupo foram incluídas empresas “saudáveis” em proporção ao primeiro grupo da seguinte forma: para cada empresa do primeiro grupo, foram relacionadas duas empresas financeiramente saudáveis (empresas “solventes”) com tamanho do ativo equivalente e pertencentes ao mesmo setor de atividade econômica, tentando, ainda, repetir a localização geográfica. Com base nos critérios expostos, foram finalmente selecionadas 56 empresas “insolventes” e 112 “solventes”, totalizando 168 empresas, ou seja, 504 observações nos três anos do período.

3.3 Coleta dos dados

Neste estudo, tratando-se de pesquisa descritiva, o método de coleta de dados foi a pesquisa documental sobre dados em fontes secundárias. Segundo Gil (2002, p.45), existe uma diferença entre pesquisa documental e bibliográfica:

A diferença essencial entre ambas está na natureza das fontes. Enquanto a pesquisa bibliográfica se utiliza fundamentalmente das contribuições dos diversos autores sobre determinado assunto, a pesquisa documental vale-se de materiais que não receberam ainda tratamento analítico, ou que ainda podem ser elaborados de acordo com os objetivos da pesquisa.

Os dados coletados são de natureza quantitativa, concentrando-se nos três últimos balanços e demonstrativos de resultado dos anos de 2003 a 2007 das empresas em estudo, junto ao SERASA e à BOVESPA.

O período escolhido foi o intervalo de tempo entre 2005 a 2007, de modo a dispor de informações que fossem, ao mesmo tempo, mais recentes e, supostamente, menos afetadas pela nova lei de falências, que entrou em vigor em junho de 2005. Analisaram-se, então, dois demonstrativos contábeis - Balanço Patrimonial e Demonstrativo de Resultado do Exercício – relativos ao ano do pedido de concordata ou falência, mais os dois anos precedentes ao pedido. Esta providência vai facilitar a inclusão de variáveis defasadas nas equações dos modelos considerados.

A coleta de dados consistiu em levantar 20 (vinte) indicadores econômico-financeiros anuais para cada uma das empresas selecionadas. Não foram incluídos dados dos balanços consolidados, objetivando-se estudar as empresas singularmente. Os grupos de indicadores econômico-financeiros são exatamente as variáveis usualmente empregadas para análise das demonstrações contábeis - liquidez, endividamento, rentabilidade e lucratividade (Iudícibus, 1998; Pereira, 2006) - e estão listadas de forma abreviada no Anexo 1.

3.4 Tratamento dos dados

Inicialmente buscou-se analisar as demonstrações contábeis das empresas que compõem o estudo, para os anos de 2003 a 2007. Os índices extraídos desta análise receberam passaram por procedimentos de pré-processamento, tais como seleção de atributos e de técnicas de *data mining*.

A seleção de atributos é uma forma de selecionar um subconjunto de atributos relevantes com o objetivo de caracterizar empresas solventes e insolventes. Já a técnica de DM que foi aplicada é a de classificação, feita através de três classificadores com metodologias distintas: Regressão Logística, Árvore de Decisão, e Máquina de Vetor Suporte.

Foi utilizado o *software* livre WEKA versão 3.5.6., que possibilita elaborar várias aplicações envolvendo *data mining*.

3.5 Avaliação das abordagens de seleção e classificação

Para avaliar os resultados das abordagens de seleção de atributos foram aplicados três algoritmos classificadores - regressão logística, árvore de decisão e máquina de vetor suporte.

Regressão logística é um caso particular de regressão linear (múltipla) em que a variável dependente é o logaritmo natural da probabilidade de ocorrência de um dado evento, que, na equação de regressão a ser estimada, é uma função linear das variáveis independentes (preditoras). Esta regressão é bem adequada para dados que apresentem características da distribuição de Poisson (ou dos “eventos raros”) e são comumente modelados usando regressão de Poisson (Han e Kamber, 2006). No caso presente, como o número de falências é geralmente muito menor que o número de sucessos, pode-se aplicar a idéia de “evento raro”. *Árvore de decisão* toma como entrada um objeto ou situação descritos por um conjunto de atributos e retorna uma “decisão” – o valor de saída previsto de acordo com a entrada (Han e Kamber, 2006). *Máquina de vetor suporte* é uma classe de funções não-lineares (chamadas *kernel*) que mapeiam os vetores de entrada em um espaço de mais alta dimensão (espaço de características), em que um hiperplano de separação será então obtido como solução para um problema de classificação (Rezende, 2005), por exemplo, a classificação (“separação”) em “solventes” e “insolventes”.

As métricas aqui utilizadas para avaliar *a posteriori* uma dada classificação foram: Matriz de Confusão (*confusion matrix*), Validação Cruzada, Medida F e Área ROC. A *Matriz de Confusão* é simplesmente uma tabela em que são representados os T_p (verdadeiros positivos), T_n (verdadeiros negativos), F_p (falsos positivos), F_n (falsos negativos). A *Validação Cruzada* permite que todos os dados da base de dados sejam utilizados para treinamento e teste; neste trabalho foram adotados 10 subconjuntos. A *Medida F* é a média entre precisão e *recall*. Nos três casos estas métricas medem a capacidade de reconhecer (e distinguir) os exemplos negativos e positivos. Finalmente, a área ROC é um gráfico bidimensional em que o eixo horizontal representa a taxa de erro da classe negativa ($1 - S_{pec}$) e o eixo vertical contém os valores de sensibilidade. O desempenho do classificador é medido pela área sob a curva ROC (Witten e Frank, 2005).

4. APRESENTAÇÃO E ANÁLISE DE RESULTADOS

Nesta seção são apresentados os resultados da primeira etapa da modelagem, a saber, a seleção de variáveis que são candidatas a compor o modelo de previsão, bem como os resultados da modelagem propriamente dita, etapa em que se verifica o poder da seleção das variáveis e depois, a segunda etapa, o poder preditivo do modelo. Além disso, são apresentados os resultados de natureza descritiva, que ajudam a visualizar e a comparar a evolução de alguns indicadores selecionados para cada grupo de empresas.

4.1 Resultado da Seleção

No Quadro 1 aparecem as variáveis preditoras selecionadas pelos respectivos métodos. Vale ressaltar que os valores subscritos nas variáveis se referem aos valores das variáveis naquele período de tempo (ano). Por exemplo, “Liquidez imediata₁” se refere a dois anos antes do ano da declaração da insolvência da entidade, porque é o primeiro (“1”)ano da série.

4.1.1 Análise descritiva - Evolução de indicadores previsores nas empresas insolventes selecionados pelas duas abordagens

A Tabela 1, a seguir, mostra a evolução dos “melhores indicadores” selecionados pela utilização das abordagens antes mencionadas, nos 3 últimos períodos anteriores à insolvência, para o caso do grupo das empresas insolventes.

	Wrapper			Filtro		ACP
	RL	AD	MVS	GS	SP	
1. Saldo de tesouraria sobre o ativo total ₂	X	X		X	X	X
2. Saldo de tesouraria sobre o ativo total ₃	X	X	X	X	X	X
3. Liquidez imediata ₁		X	X			
4. Liquidez imediata ₂						X
5. Liquidez imediata ₃		X	X		X	X
6. Liquidez seca ₂		X	X			X
7. Liquidez seca ₃	X		X	X	X	X
8. Liquidez corrente ₂				X		
9. Margem operacional ₂	X			X		
10. Margem operacional ₃					X	
11. Margem operacional após resultado financeiro ₁	X					X
12. Margem operacional após resultado financeiro ₃	X	X	X	X	X	X
13. Margem líquida ₁	X					
14. Margem líquida ₃	X			X	X	
15. Margem bruta ₁			X			
16. Margem bruta ₃			X	X	X	
17. Rentabilidade operacional sobre ativo total ₃	X	X	X	X	X	
18. Rentabilidade líquida sobre o ativo total ₂	X		X	X		
19. Rentabilidade líquida sobre o ativo total ₃	X			X		
20. Rentabilidade Líquida sobre patrimônio líquido ₃	X			X	X	X
21. Endividamento oneroso sobre patrimônio líquido ₂		X				X
22. Endividamento total sobre ativo total ₂	X					
23. Endividamento total sobre patrimônio líquido ₂		X				X
24. Endividamento total sobre patrimônio líquido ₃			X	X	X	
25. Giro do ativo ₁	X					X
26. Giro do ativo ₂	X		X			X

Quadro 1 – Quadro-resumo dos atributos selecionados com as respectivas técnicas.

Fonte: Os autores.

Legenda:

Wr/RL/AD/MVS = Abordagem de seleção de atributos *wrapper* tendo como algoritmos de aprendizagem regressão logística (RL), árvore de decisão (AD) e máquina de vetor suporte (MVS).

Fil/SG/SP = Abordagem de seleção de atributos filtro tendo como algoritmo de busca *search genetic* (SG) e como algoritmo de busca *greedy stepwise* (SP).

ACP = Análise de componentes principais.

Para o indicador referente à Liquidez Seca (LS), verifica-se que suas médias vão-se deteriorando ao longo dos anos. Além disso, os desvios-padrões aumentam para depois tomar algum fôlego, indicando uma relativa homogeneidade das características. A variação mais significativa se refere à passagem do período T-1 para o período T (81,84%). Esse resultado sugere a falta de resultados positivos nas ações para administrar os estoques, valendo-se de prováveis acúmulos de estoques, indicando perda na capacidade de melhorar suas vendas.

Variável	Período T-2			Período T-1			Período T		
	Média	DP	CV (%)	Média	DP	CV (%)	Média	DP	CV (%)
LS	0,6203	0,6138	98,52	0,5225	0,661	126,52	0,435	0,541	124,36
MORF	0,30	0,51	170,00	-0,51	0,257	-50,39	-0,57	1,35	-236,84
STAT	-0,193	0,381	200,27	-0,211	0,425	201,421	-0,297	0,4638	126,79
ROAT	-0,1098	0,3093	281,72	-0,2545	0,5583	219,37	-0,291	0,4272	143,77

Tabela 1 – Evolução dos indicadores no grupo das insolventes.

Fonte: Elaborada pelos autores

Legenda: Média = Média aritmética dos índices; DP = Desvio-padrão; CV = Coeficiente de variação (DP/Média).

No caso da variável Margem Operacional após o Resultado Financeiro (MORF), os índices apresentam comportamentos diferentes - as médias vão-se deteriorando ano a ano, mas os desvios-padrões apresentam movimento flutuante, havendo uma elevação e logo após uma queda em seu valor.

A evolução da variável Saldo de Tesouraria sobre Ativo Total (STAT) é coerente com a situação de empresas em fase pré-falimentar (Pereira, 2006), uma vez que suas médias crescem negativamente durante os períodos precedentes à insolvência. Uma empresa nessa situação apresenta excessiva dificuldade para gerenciar sua tesouraria, talvez devido à enxurrada de múltiplos compromissos, levando ao sacrifício das atividades operacionais na tentativa de atender obrigações de curtíssimo prazo.

Para a variável indicativa de Rentabilidade - Operacional sobre Ativo Total (ROAT) - o percurso das médias se manteve coerente, com diminuição dos valores no decorrer dos períodos. O que se observa neste estudo é a de perda gradativa de rentabilidade das empresas insolventes na amostra aqui utilizada. Há maior deterioração durante a transição do período T-2 para T-1 ($\Delta ROAT = 144,36\%$), sugerindo aumento de ineficiência operacional e, conseqüentemente, financeira, conduzindo a empresa ao estado de insolvência mais agudo.

No caso do índice de giro do ativo, ocorre acentuada seguida de posterior melhora relativa, podendo-se supor que houve alguma desmobilização visando melhorar a estrutura de capital das empresas. No entanto, pode-se argumentar, também, que houve perda da eficiência operacional se o índice for analisado em conjunto com o índice Margem Operacional após o Resultado Financeiro (MORF). Pode-se supor, ainda, que nas empresas insolventes estudadas ocorreram acentuados comprometimentos na capacidade operacional, conseqüentemente influenciando a liquidez.

De acordo com o comportamento das variáveis selecionadas, pode-se deduzir a ocorrência de perdas da capacidade operacional nas empresas estudadas, conduzindo-as à insolvência.

4.1.2 Análise descritiva - Evolução de indicadores previsores nas empresas solventes selecionados pelas duas abordagens

Na Tabela 2, em que não ocorrem variações tão relevantes, o que chama atenção são os valores encontrados para o coeficiente de variação. A comparação destes valores com os das empresas insolventes mostra que são inferiores àqueles, indicando que a dispersão nos índices das empresas insolventes é maior do que nas empresas solventes.

Variável	Período T-2			Período T-1			Período T		
	Média	DP	CV (%)	Média	DP	CV (%)	Média	DP	CV (%)
LS	0,85	0,7334	86,28	0,862	0,6521	75,64	0,8928	0,6689	74,9
MORF	0,21	1,79	873,17	0,89	6,42	721,34	1,33	9,998	751,72
STAT	0,0018	0,227	12611,1	-0,0007	0,0967	1401,44	-0,012	0,1738	1448,33
ROAT	0,0015	0,277	18466,7	-0,0019	0,1598	8410,52	-0,006	0,087	1500,00

Tabela 2 – Evolução dos indicadores no grupo das empresas solventes.
Fonte: Elaborada pelos autores

O inverso ocorreu com os índices das empresas solventes; cujos coeficientes de variação são bem mais comportados do que os correspondentes no grupo das insolventes. Aliás, a explicação para essas diferenças deve começar nas médias, que são negativas e estão declinando nas empresas insolventes. Com a capacidade operacional razoavelmente comprometida, a busca de rentabilidades razoáveis é abandonada, passando a ser uma estratégia secundária, resultando na aceleração do processo de insolvência.

4.2 Resultados das técnicas aplicadas segundo diferentes classificadores

São apresentados nesta seção os resultados obtidos com a aplicação de técnicas de seleção, com o propósito de desenvolver modelos de insolvência compostos por variáveis selecionadas simultaneamente nas abordagens filtro, *wrapper* e análise de componentes principais, e a seguir determinar os efeitos dessas abordagens segundo os classificadores estudados.

Na abordagem *wrapper* somente serão apresentados resultados dos casos em que o algoritmo indutor for o mesmo do classificador, visto que os resultados encontrados quando do uso do algoritmo indutor diferente do classificador foram pouco significativos. Além disso, os resultados apresentados foram gerados pelo algoritmo de busca GS devido ao desempenho relativamente melhor do que o do GP.

4.2.1 Classificação de resultados de previsão dos grupos (subconjuntos de atributos selecionados pelas abordagens *wrapper*, filtro e ACP; classificador aplicado sendo a Regressão Logística)

Técnica de seleção	ALGORITMO DE APRENDIZAGEM - RL							
	WR RL		FIL/GS		FIL/SP		ACP	
Grupo de origem	I	S	I	S	I	S	I	S
Insolventes - I	45	11	35	21	38	18	41	15
Solventes - S	6	106	10	102	11	101	13	99
Medida F	0,841	0,926	0,693	0,868	0,724	0,874	0,745	0,876
Área Roc	0,915	0,915	0,859	0,859	0,868	0,868	0,854	0,854
Classificação correta no grupo de origem	89,88%		81,54%		82,73		83,33%	

Tabela 3 – Classificação de resultados de previsão dos modelos aplicando as variáveis selecionadas pelas técnicas de seleção estudadas para o classificador Regressão Logística.

Fonte: Elaborada pelos autores.

Na Tabela 3 pode ser constatado que a abordagem *wrapper* obteve os melhores resultados tanto na Matriz de Confusão, quanto para a Medida F, Área ROC e na classificação correta.

4.2.2 Classificação de resultados de previsão dos grupos (subconjuntos de atributos selecionados pelas abordagens *wrapper*, filtro e ACP; classificador sendo a Árvore de Decisão)

Técnica de seleção	ALGORITMO DE APRENDIZAGEM - AD							
	WR AD		FIL/GS		FIL/SP		ACP	
Grupo de origem	I	S	I	S	I	S	I	S
Insolventes - I	51	5	50	6	51	5	50	6
Solventes - S	6	106	8	104	8	104	11	101
Medida F	0,903	0,951	0,877	0,937	0,887	0,941	0,855	0,922
Área Roc	0,942	0,942	0,953	0,953	0,958	0,958	0,931	0,931
Classificação correta no grupo de origem	93,45%		91,66%		92,26		89,88%	

Tabela 4 – Classificação de resultados de previsão dos modelos aplicando as variáveis selecionadas pelas técnicas de seleção estudadas para o classificador Árvore de Decisão.

Fonte: Elaborada pelos autores.

Na Tabela 4 pode ser constatado que *wrapper* apresentou o melhor resultado. A melhor classificação correta do grupo de origem foi de 93,45% (última linha, segunda coluna).

4.2.3 Classificação de resultados de previsão dos grupos composta pelos subconjuntos de atributos selecionados pelas abordagens *wrapper*, filtro e ACP aplicando o classificador Máquina de Vetor Suporte.

Técnica de seleção	ALGORITMO DE APRENDIZAGEM - MVS							
	WR MVS		FIL/GS		FIL/SP		ACP	
Grupo de origem	I	S	I	S	I	S	I	S
Insolventes - I	52	4	34	22	41	15	0	56
Soventes - S	1	111	3	109	5	107	0	112
Medida F	0,954	0,978	0,731	0,897	0,804	0,915	0	0,8
Área Roc	0,96	0,96	0,79	0,79	0,844	0,844	0,5	0,5
Classificação correta no grupo de origem	97,02%		85,11%		88,09		66,66%	

Tabela 5 – Classificação de resultados de previsão dos modelos aplicando as variáveis selecionadas pelas técnicas de seleção estudadas para o classificador Máquina de Vetor Suporte.

Fonte: Elaborada pelos autores.

Na Tabela 5 pode-se verificar que a abordagem *wrapper* foi a que obteve os melhores resultados, sendo que a melhor classificação correta do grupo de origem alcançou 97,02%.

4.2.4 Quadro comparativo e análise dos resultados dos modelos compostos pelos subconjuntos de atributos selecionados e avaliados pela abordagem filtro e *wrapper*:

A Tabela 6 resume os resultados do trabalho e pode oferecer uma indicação mais prática para escolha das técnicas aqui consideradas, pois que permite comparar os resultados gerados em cada alternativa de especificação dos modelos estimados e avaliados pelas abordagens filtro, *wrapper* e análise de componentes principais. Conforme mostra a tabela, os modelos correspondentes à abordagem *wrapper* apresentaram melhores índices de classificação para cada um dos classificadores estudados, podendo, neste sentido, ser considerado robusto (quer dizer, independente do classificador). Cabe ressaltar que, conforme era de esperar segundo a literatura, o melhor desempenho ocorreu no caso do classificador Máquina de Vetor Suporte, que produziu 97,02% de classificação correta no grupo de origem, as melhores medidas Fs e a melhor área ROC.

Abordagem da avaliação de subconjuntos de atributos selecionados	Classificador	Nº de acertos	Nº de erros	Acertos %	Erros %
Filtro – GS	AD	154	14	91,66	8,34
Filtro – SP	AD	155	13	92,26	7,74
ACP	AD	151	17	89,88	10,12
<i>Wrapper</i> RL	RL	151	17	89,88	10,12
<i>Wrapper</i> AD	AD	157	11	93,45	6,55
<i>Wrapper</i> MVS	MVS	163	5	97,02	2,98

Tabela 6: Classificação dos melhores resultados de previsão para cada técnica de seleção.

Fonte: Elaborada pelos autores.

O objetivo deste estudo foi precisamente comparar três abordagens de seleção de atributos em um banco de dados contendo dados contábeis de empresas brasileiras. Os resultados obtidos mostram que a abordagem *wrapper* obteve os melhores resultados na avaliação da seleção de subconjuntos de atributos para cada um dos classificadores aplicados, sendo GP o mais eficiente quando se usou essa abordagem. Quando da utilização de um algoritmo de avaliação (aprendizagem) diferente do algoritmo classificador os resultados não foram aceitáveis.

Dentre os algoritmos de busca utilizados pela abordagem Filtro, o SP foi o que forneceu melhores resultados.

Dentre os classificadores o que alcançou melhor resultado foi Máquina de Vetor Suporte, com 97,02% de acertos na classificação correta no grupo de origem, confirmando o sucesso recente que lhe atribuem os pesquisadores da área.

Em todos os classificadores de avaliação a abordagem *wrapper* obteve os melhores resultados, podendo-se concluir que as variáveis contábeis selecionadas foram capazes de realizar bem a tarefa de classificação. Em termos de usabilidade, pode-se esperar que a abordagem *wrapper* seja sempre viável para avaliar subconjuntos de atributos que irão compor os modelos preditivos que se quiser comparar. Outra conclusão importante se refere ao fato de que a redução da dimensionalidade pode e deve ser praticada como um dos passos na etapa de pré-processamento, quando do uso de dados contábeis brasileiros. Foram demonstradas, também, a importância da explicitação da etapa de avaliação da seleção de atributos na busca de melhores resultados e da utilidade de técnicas de *data mining*, em comparação com abordagens mais tradicionais, para a previsão de insolvência usando dados contábeis de empresas brasileiras.

5. COMENTÁRIOS CONCLUSIVOS

O trabalho utilizou-se de dados contábeis, relativos a vinte indicadores econômico-financeiros anuais de empresas selecionadas, no período de 2003-2007. Os grupos de indicadores econômico-financeiros são os mesmos usualmente empregados na análise das demonstrações contábeis - liquidez, endividamento, rentabilidade e lucratividade (Iudícibus, 1998; Pereira, 2006).

Quanto aos resultados, pode-se concluir que, dentre as variáveis selecionadas nas três abordagens testadas, cinco foram simultaneamente “apontadas” pelos dois métodos estudados (a saber, as variáveis LS, MORF, STAT, ROAT), sugerindo que são aquelas com melhor capacidade para caracterizar as classes de empresas solventes e insolventes. Nesta medida tornam-se, então, fortes candidatas a compor um modelo de previsão de insolvência.

Com o apoio das técnicas de *Data Mining*, foram comparadas três abordagens de seleção de atributos na elaboração de modelos de previsão de insolvência, buscando, por meio do uso de diferentes técnicas, destacar a importância da fase de seleção das variáveis preditivas. Pode-se argumentar que a atenção a esta etapa de seleção de variáveis, assunto pouco discutido na literatura sobre modelos de previsão de insolvência, corresponde a uma contribuição o ponto central desse trabalho.

Para além da relevância acadêmica, este resultado assume importância ainda maior ao considerar as modificações ocorridas nos mercados, que se tornaram mais competitivas e passaram a fortalecer e valorizar empresas que toleraram rentabilidades menores para manter sua continuidade.

A utilização de modelos preditivos de insolvência, construídos pela aplicação de *Data Mining*, é uma dentre várias formas de avaliar o risco de uma instituição sem depender, apenas, da avaliação subjetiva do analista. Na prática da análise de risco, e tendo em mente os recursos computacionais disponíveis, esses modelos preditivos podem ser incorporados como procedimentos analíticos para avaliar a probabilidade de insolvência. Neste sentido, então, tornam-se interessantes para bancos, investidores, governos, auditores, gerentes, fornecedores, empregados e muitos outros, para que possam avaliar, com razoável antecedência, se há ou haverá problemas de insolvência, em andamento ou futuros. Por exemplo, pode-se ponderar os diversos índices selecionados de modo a obter um modelo que construa um índice de risco de crédito. A ponderação poderá ser feita com técnicas de *Data Mining*, sugerindo que a combinação de *Data Mining* com indicadores contábeis seja uma ferramenta útil para prever concordatas de empresas, podendo, ademais, ser utilizada para estabelecer *scores* associados ao risco de crédito.

Apesar de a qualidade dos dados contábeis ser, ainda, muitas vezes questionada em termos de utilização na construção de modelos de previsão de falência, os resultados aqui obtidos foram significativos e satisfatórios, atingindo cerca de 90% ou mais de acerto na classificação *a posteriori* e evidenciando, assim, o conteúdo de informação que esses dados proporcionam, ainda que seja para previsões. É possível que, com o emprego de alguns desses dados, seja cada vez melhor a precisão de estudos preditivos sobre a saúde financeira de uma empresa.

Estando limitados a uma particular amostra, os resultados deste estudo não permitem afirmar que a abordagem *wrapper* será sempre mais eficaz para quaisquer outras bases de dados, a despeito de que, aqui, para esta base de dados, tenha obtido os melhores resultados. Aliás, dada a natureza empírica deste estudo, não seria mesmo possível esperar resultados com tamanha força. Desde logo, no entanto, os resultados encorajam prosseguir na elaboração de novos estudos comparativos, já agora com uma expectativa acerca do desempenho relativo daquela abordagem.

Finalmente, em estudos futuros, a exemplo do que vem ocorrendo em outros campos de aplicação, cabe recomendar a utilização de outras técnicas de seleção de atributos e de outros classificadores, bem como de outras e novas técnicas de DM na obtenção de regras de classificação, tais como comitê de classificadores ou balanceamento do banco de dados, entre outras.

REFERÊNCIA BIBLIOGRÁFICA

AGARWAL, V.; TAFFLER, R. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking and Finance*, v. 32, Issue 8, p.1541-1551, aug. 2008.

ALTMAN, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, v. 23, n.4, p. 589-609, sep. 1968.

ALTMAN, E. I., G. MARCO, G.; VARETTO, F. Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience), *Journal of Banking and Finance*, v. 18, Issue 3, p. 505-529, may. 1994.

BACK, B., Laitinen, T., KAISA, S. Neural Networks and Genetic Algorithms for Bankruptcy Predictions. *Expert Systems With Applications*, v. 11, Issue 4, p. 407-413, 1996.

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

BEAVER, W. H. Financial Ratios as Predictors of Failure, *Journal of Accounting Research - Empirical research in accounting: selected studies*. *Journal of Accounting Research* 4 (Suppl.), p. 71-111, 1967.

BLÖCHLINGER, A., LEIPPOLD, M. Economic benefit of powerful credit scoring. *Journal of Banking and Finance*, v. 30, Issue 3, march, p. 851-873, 2006.

BRAGANÇA, L. A. de; BRAGANÇA, S. L. Previsão de concordatas e falências no Brasil, *Anais do VII Congresso ABAMEC*, 1984.

BRITO, G. A.S., NETO, A. A., CORRAR, L.J. Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil. *Revista Contabilidade e Finanças*, v. 20, n. 51, set./dez., p. 28-43, 2009.

DEAKIN, E. B. A Discriminant Analysis of Prediction of Business Failure, *Journal of Accounting Research*, v. 10, n. 1 spring, p. 167-179, 1972.

EDMISTER, R. O. An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction, *Journal of Financial and Quantitative Analysis*, v. 7, n. 2, Supplement: Outlook for the Securities Industry mar., p. 1477-1493, 1972.

EISENBEIS, R. A. Pitfalls in the application of discriminant analysis in business, finance, and economics, *The Journal of Finance*, v. 32, n. 3 jun. p. 875-900, 1977.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. Em *Advances in Knowledge Discovery and Data Mining: an overview*. American Association for Artificial Intelligence, p. 1-30, 1996.

Fitzpatrick P.J.: A Comparison of ratios of successful industrial enterprises with those of failed firms. *Certified Public Accountant*, October, November and December, p.598-605, p. 656-662, p. 727-731, 1932.

FREITAS A. A. Data mining and knowledge discovery with evolutionary algorithms. Berlin / Heidelberg / New York: Springer-Verlag, 1998.

GIL, A. C. *Como elaborar Projetos de Pesquisa*. 4 ed. São Paulo: Atlas, 2002.

HALL, M. A; HOLME, G. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. XX, NO. Y, MONTH , 2002.

HAN, J; KAMBER, M. *Data mining: concepts and techniques*. 2 ed. San Francisco: Morgan Kaufmann, 2006.

HÄRDLE, W.; MORO R.; A. SCHÄFER D. Predicting Bankruptcy with Support Vector Machines. SFB 649 Discussion Paper 2005-009. Disponível em <http://sfb649.wiwi.hu-berlin.de> [acesso em 10/03/2006].

HAYKIN, SIMON. *Redes neurais: Princípios e prática*. Trad. Paulo Martins Engel. – 2. ed. – Porto Alegre: Bookman, 2001.

HORTA, R. A. M. Utilização de indicadores contábeis na previsão de insolvência – análise empírica de uma amostra de empresas comerciais e industriais brasileiras. Rio de Janeiro, 2001. Dissertação (Programa de Mestrado em Ciências Contábeis), UERJ, 2001.

_____; ALVES, F. J. S. Técnicas de *Data Mining* na seleção de atributos para previsão de insolvência: aplicação e avaliação usando dados brasileiros recentes. XXXII EnANPAD, Rio de Janeiro, 2008. *Anais . . . ANPAD*: Rio de Janeiro, 2008.

IUDÍCIBUS, S. *Análise de Balanços*. 6 ed. São Paulo: Atlas, 1995.

- KANITZ, S. C. *Como prever falências*, São Paulo: Mc Graw-Hill do Brasil, 1978.
- KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, v. 97, Issues 1-2, dec., p. 273-324, 1997.
- LENNOX, C. Identifying failing companies: a reevaluation of the logit, probit and d.a. approaches, *Journal of Economics and Business*, v.51, p. 347-364, 1999.
- LIU, H.; MOTODA, H. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, Massachusetts, 1998.
- MANLY, B. F. J. *Métodos estatísticos multivariados: uma introdução*. 3 ed. Porto Alegre: Bookman, 2008.
- PEREIRA, José da Silva. Gestão e análise de risco de crédito. Dissertação (Mestrado em Administração) – Fundação Getúlio Vargas, São Paulo. 1982.
- PEREIRA, J. S. *Gestão e Análise de Risco de Crédito*. 5 ed. São Paulo: Atlas, 2006.
- PIRAMUTHU, S. On preprocessing data for financial credit risk evaluation. *Expert Systems with Applications*, v. 30, Issue 3, april, p. 489-497, 2006.
- REZENDE, S.O. (Org.). *Sistemas inteligentes: fundamentos e aplicações*. Barueri, SP: Manole, 2005.
- SANVICENTE, A. Z.; MINARDI, A. M. F. “Identificação de indicadores contábeis significativos para previsão de concordata de empresas”, 2000. Disponível: como http://www.risktech.br/artigos/artigos_técnicos/index.html. Acesso em 05/05/2003.
- SHIRATA, C. Y. Financial ratios as predictors of bankruptcy in Japan: an empirical research. 2001. Disponível: <http://www.shirata.net/eng/APIRA98.html>. Acesso em: 09/07/2004.
- SOMOL, P.; BAESENS, B.; PUDIL, P.; VANTHIENEN J. Filter versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems*. V.20, aug. p. 985 – 999, 2005.
- SUN, Z.; BEBIS, G.; MILLER, R. Object detection using feature subset selection. *Pattern Recognition*, V. 37, Issue 11, november, p.2165-2176, 2004.
- WINAKOR, A. H.; SMITH, R. F. Changes in the Financial Structure of Unsuccessful Industrial Corporations. Bulletin n. 51, *Bureau of Business Research*, University of Illinois, 1935.
- WITTEN, I. H., FRANK, E. Data Mining: practical machine learning tools and techniques. The Morgan Kaufmann Series in Data Management Systems, 2 ed. San Francisco: Elsevier, 2005.

WU, C. H.; FANG, W. C.; GOO, Y, J. Variable selection method affects SVM approach in bankruptcy prediction. 2006. Disponível em: http://www.atlantispress.com/php/download_paper.php?id=114. Acesso em 09/10/2008.

Anexo 1 – Índices econômicos financeiros e suas abreviaturas

Liquidez corrente - LC, Liquidez seca – LS, Liquidez Imediata – LI, Liquidez Geral – LG, Endividamento Oneroso sobre Patrimônio Líquido – EOPL, Endividamento Oneroso de Curto Prazo – EOCP, Endividamento Total sobre Patrimônio Líquido - ETPL, Endividamento Total sobre Ativo Total - ETAT, Margem Bruta – MB, Margem Operacional – MO, Margem Operacional sobre Resultado Financeiro – MORF, Margem Líquida – ML, Giro do Ativo – GA, Rentabilidade do Ativo Operacional Total – ROAT, Retorno dos Acionistas – ROE, Retorno do Investimento Total – ROI, Saldo de Ativo Total - STAT, Modelo Dupont Adaptado – RTA, Lucro antes dos juros, impostos - EBIT, Lucro antes dos juros, impostos, depreciações/exaustão e amortização - EBITDA.