TÉCNICAS DE DATA MINING NA SELEÇÃO DE ATRIBUTOS PARA PREVISÃO DE INSOLVÊNCIA: APLICAÇÃO E AVALIAÇÃO USANDO DADOS BRASILEIROS RECENTES

Autoria: Rui Américo Mathiasi Horta, Francisco José dos Santos Alves

RESUMO

A partir do final da década de 90, questões como o aparecimento de novas técnicas de modelagem, a expansão dos mercados de capitais, os impactos dos mercados imperfeitos e das informações assimétricas, a crescente importância da gerência do risco de crédito e as constantes mudanças no ambiente econômico das empresas trouxeram de volta o interesse pela previsão de insolvência de empresas. Nos dias atuais, para medir, gerir e prever a "saúde financeira" de empresas pode-se usar os chamados modelos de previsão de insolvência, construídos com apoio em técnicas de *data mining*, as quais são aplicadas para analisar índices econômico-financeiros selecionados a partir de demonstrativos contábeis. O objetivo principal deste artigo é comparar duas abordagens de avaliação de subconjuntos de atributos, filtro e *wrapper*, selecionados para elaboração de modelos de previsão de insolvência, fundamentando-os em técnicas de *data mining* utilizadas em uma aplicação empírica de empresas comerciais e industriais brasileiras, empregando dados para o período de 2004 a 2006. Neste trabalho se demonstrou, para a amostra utilizada, que a abordagem filtro é mais eficiente do que abordagem *wrapper,* a abordagem filtro obteve melhores resultados de classificação tanto na técnica de regressão logística (86,90%) como de redes neurais (92,26%).

Palavras chave: Índices econômico-financeiros; Previsão de insolvência; *Data mining*; Seleção de atributos.

INTRODUÇÃO

Prever com exatidão se 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. Entretanto, é possível identificar com auxílio de modelos de previsão de insolvência, aquelas organizações com maiores probabilidades de falir em futuro próximo, permitindo, assim, identificar e apontar medidas corretivas em tempo hábil.

Nos anos recentes, uma revolução de idéias e de novas tecnologias tem fervilhado relativamente à maneira pela qual o risco de crédito é medido e gerido (Saunders, A. Allen, L. DeLong, G. 2004). Algumas outras justificativas podem ser encontradas para este surto de interesse no assunto, conforme apresentado abaixo:

(i) Nas últimas décadas o ambiente econômico geral das empresas tem mudado com uma enorme velocidade e experimentado tendências para baixo (dificuldades financeiras); (ii) A desintermediação financeira está ocorrendo rapidamente devido à expansão dos mercados de capitais; (iii) As margens de juros, ou spreads, têm se tornado muito estreita – ou seja, a compensação de risco-retorno advinda de empréstimos piorou. (iv) Aumento nas incertezas financeiras nas 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; (v) Empresas insolventes acabam gerando envolvimentos em vários setores econômicos e grandes custos, inclusive os sociais para todos; (vi) Na linha de extensão das pesquisas acadêmicas dos impactos dos *mercados imperfeitos e informações assimétricas*, pesquisas de ranking de crédito e previsão de insolvência têm aumentado; por fim, (vii) Com a evolução da disponibilidade de dados e das técnicas computacionais há um aumento de possibilidades de pesquisas que concernem à previsão de insolvência.

Dentre essas novas técnicas computacionais, destaca-se *Data Mining* (DM) no que concerne ao gerenciamento, controle e administração dos riscos associados a avaliações erradas. Essa

técnica tornou-se proeminente no final da década de 90, e apresentou uma forte ênfase na combinação de conjuntos de dados para capturar padrões que são muito sutis ou complexos para serem detectados somente por analistas de dados (Kreuze, 2001).
O objetivo deste artigo é comparar as abordagens filtro e *wrapper* de avaliação de subconjuntos de atributos, selecionados para elaboração de modelos de previsão de insolvência, e ilustrá-lo através de uma aplicação empírica composta por indicadores econômico-financeiros obtidos em demonstrativos contábeis de empresas brasileiras no período de 2002 a 2006.
A principal vantagem do artigo é considerar explicitamente a fase de seleção das variáveis preditivas através de duas das mais importantes abordagens de avaliação. Alguns autores (Shirata, 2001, Wu, et al., 2006 e Piramuthu, 2006) têm chamado atenção para a importância do processo de seleção de atributos. Um fato que na maior parte dos estudos sobre risco de crédito, nem sempre é claramente discutido, o que dificulta o entendimento sobre como se chegou às variáveis utilizadas.

1. 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 possibilite obter uma classificação de sucesso quanto à "saúde financeira" das empresas. A sua aplicabilidade, visada inicialmente, são as operações de curto prazo, considerando que a insolvência está mais relacionada à perda pela empresa da capacidade de endividamento do que com o seu desempenho operacional.
Apesar de sua longa história na literatura especializada (Fitzpatrick 1932, Winakor & Smith, 1935), o estudo do "insucesso" de empresas com base em indicadores obtidos a partir dos demonstrativos contábeis tomou ímpeto nos anos 70 (Blum 1974, Deakin 1972, Edmister 1972, Kanitz 1978 e tantos outros), em seguida aos trabalhos pioneiros de Beaver (1967) e de Altman (1968).
No Brasil, a análise da insolvência de empresas com objetivos preditivos desenvolveu-se de modo significativo a partir dos anos 80 (Pereira 1982, Bragança & Bragança 1984, Kasznar 1986, Sanvicente & Minardi 2000, Horta 2001), seguindo o caminho aberto por Kanitz (1978).
No levantamento do referencial teórico desta pesquisa, constatou-se a prática de utilizar variáveis previamente relacionadas em pesquisas anteriores (Kanitz 1978, Pereira 1982, Bragança & Bragança 1984, Kasznar 1986, Sanvicente & Minardi 2000). Este procedimento pode desconsiderar fatores culturais e de legislações como a tributária, fiscal e societária de cada país.
A partir dos anos 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 & Varetto 1994, Back, B., Laitinen, T., Kaisa, S. 1996, Brockert et al. 1997, Eisenbeis 1997, Lennox 1999, Härdle, Moro & Schäfer, 2005, Abdelwahed & Amir, 2005 SUN, Wu Chih-Hung et al. 2006,Shenoy, P.P.2007 entre outros).
Apesar de novas técnicas de modelagem se tornar mais acessíveis, existem vários problemas relatados na literatura específica com a aplicação desses métodos no assunto de previsão de insolvência (Balcaen e Ooghe, 2006). Alguns desses problemas são categorizados nos seguintes tópicos: (i) a dicotomia da variável dependente; (ii) a seletividade da amostra; (iii) a não estacionariedade e dados instáveis; (iv) o uso anual de informações contábeis; (v) a seleção de variáveis independentes e (vi) a dimensão temporal.

1.2 Data mining

A definição aceita por diversos pesquisadores de DM foi elaborada por Fayyad, Piatetsky-Shapiro, & Smyth (1996) ao afirmarem que: "Extração de Conhecimento de Base de Dados é o processo de identificação de padrões validos, novos, potencialmente úteis e compreensíveis embutidos nos dados".
O processo de identificação de padrões em DM é dividido em três grandes etapas, conforme Rezende et all.(2005): pré-processamento, extração de padrões e pós-processamento.
a) pré-processamento: esta etapa se caracteriza pela adequação dos dados para a extração de conhecimento. Diversas adequações nos dados podem ser executadas na etapa de pré-processamento, entre elas: tratamento de valores desconhecidos, identificação e descrição de valores extremos, tratamento de conjunto de dados com classes desbalanceadas, e a seleção de atributos.
b) extração de padrões: compreende a escolha da tarefa de DM a ser empregada, a escolha do algoritmo e a extração propriamente dita. Essa escolha é feita de acordo com os objetivos desejáveis para a solução a ser encontrada e as tarefas possíveis de um algoritmo de extração de padrões podem se agrupadas em atividades de acordo com aquilo que ela pode fazer: (i) descrição e visualização; (ii) associação e *clusterização*; e (iii) classificação e estimação (predição) (CHYE K. H., CHIN. T.W., PENG G. C. 2004).
Descrição e visualização contribuem para a compreensão de uma série de dados, sobretudo quando a quantidade é bem grande, detectam padrões escondidos nos dados, especialmente em dados complicados que contêm interações complexas e não lineares. São geralmente executados antes da modelagem tentando representar e compreender melhor os dados.
Na associação, o objetivo é determinar a relação entre as variáveis. Em *clusterização*, a meta é agrupar objetos homogêneos de tal maneira que esses objetos pertençam ao mesmo conjunto e os objetos que pertencem aos conjuntos diferentes sejam heterogêneos.
A mais comum e importante aplicação em *data mining* provavelmente envolve predição, por vezes referido como modelagem. Classificação refere-se à previsão de um objetivo em que a variável é de natureza categórica. Para construção de modelos de previsão de insolvência, técnicas de modelagem preditiva são as mais relevantes.
Para modelagem preditiva, DM incluem técnicas estatísticas tradicionais, como a análise discriminante múltipla e regressão logística. Técnicas de DM incluem também métodos não tradicionais desenvolvidos nas áreas de inteligência artificial e aprendizado de máquina. Os dois mais importantes destes métodos são redes neurais e árvores de decisão (Chye., K.H., Chin., T.W., Peng., G.C. 2004). Neste trabalho serão utilizadas análise de regressão logística e redes neurais por serem amplamente utilizadas na literatura específica e serão mais bem definidos em itens adiante.
c) pós-processamento: a extração de padrões pode gerar uma quantidade enorme de padrões, muitos dos quais podem não ser relevantes o para o usuário. É importante desenvolver algumas técnicas de apoio no sentido de fornecer aos usuários apenas padrões mais interessantes através de medidas de desempenho e qualidade (Silberschatz & Tuzhilin, 1995).

1.2.1 Seleção de atributos

A seleção de atributos (SA) é uma etapa relevante para elaboração de modelos de previsão de insolvência. Apesar disso a grande maioria dos estudos sobre previsão de inadimplência parte de um conjunto inicial de variáveis, escolhida freqüentemente na base de sua popularidade na literatura específica e de seu sucesso preditivo nas pesquisas precedentes.

Ela desempenha uma tarefa essencial dentro desse processo, pois representa um problema de fundamental importância em DM, sendo freqüentemente realizada como uma etapa de pré-processamento. Os objetivos da seleção de atributos em modelos de previsão de insolvência são: (i) o desenvolvimento de modelos compactos, (ii) o uso e refinamento do modelo de classificação para avaliação e (iii) a identificação de índices financeiros relevantes (Piramuthu, 2006). Em problemas típicos de classificação, os valores de um conjunto de variáveis independentes de um conjunto de exemplos devem compor um modelo que tenha a capacidade de categorizar futuras classes corretamente (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 1 (Liu, H. e Motoda, H., 1998).

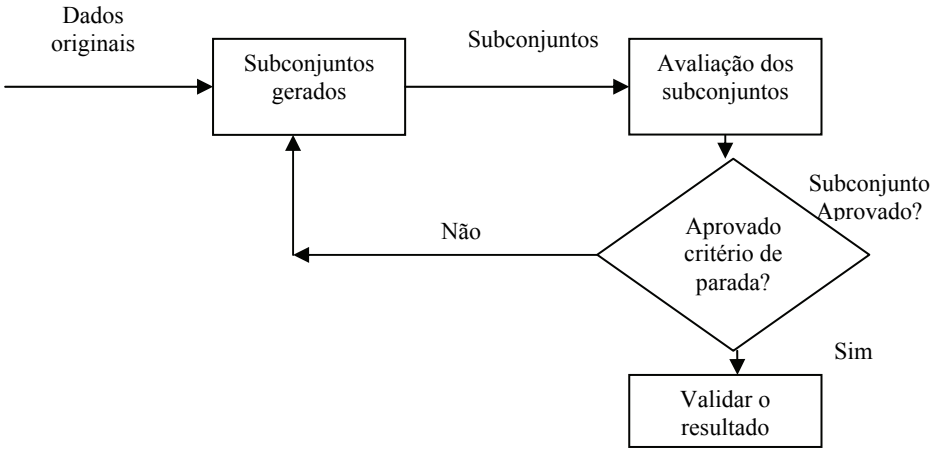

Figura 1. - Passos na Seleção de Atributos.

1.2.1.1 Busca do subconjunto de atributos

A partir de todas os atributos disponíveis, seleciona-se um subconjunto de variáveis relevantes com o apoio de um algoritmo de busca. Neste estudo foi utilizado, que emprega uma heurística para conduzir a busca. Este método de busca é feita no espaço de subconjuntos do atributo indo e voltando dentro do conjunto. Ajusta o número dos melhores atributos encontrados permitindo o controle da busca. Neste trabalho foi utilizada a seleção *forward* (vai adicionando um atributo por vez ao subconjunto até que não se consiga melhorar a qualidade do subconjunto de atributos) como ponto de partida na busca e foi a escolhida por ter sido a abordagem que apresentou os melhores resultados.

1.2.1.2 Avaliação do subconjunto de atributos

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, como ele interage com o algoritmo de aprendizado. Essa interação pode ser subdividida, basicamente, em duas abordagens principais: filtro e *wrapper*. (Kohavi & John, 1997).

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. 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 como exemplo o conjunto de exemplos descrito utilizado somente o subconjunto de atributos importantes selecionados pelo filtro. A meta é selecionar um subconjunto de atributos que preserva a informação pertinente no conjunto inteiro de atributos

(Freitas A. A., 1998, p. 66). A idéia é filtrar atributos irrelevantes segundo algum critério antes do aprendizado ocorrer (Jonh, G. H., Kohavi, R., Pfeger, K., 1994).

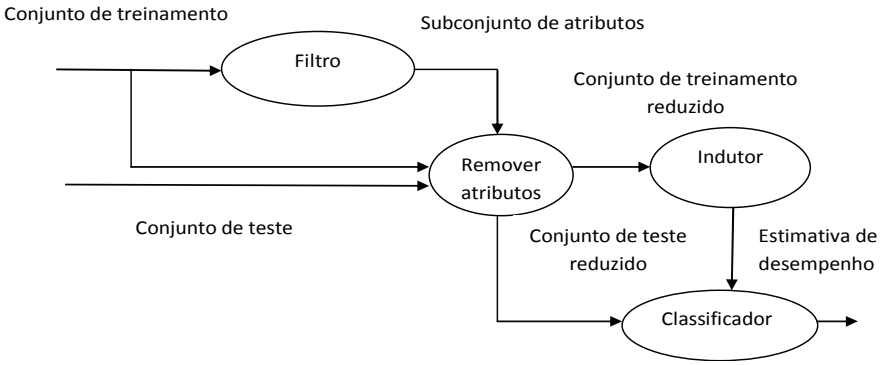

Figura 2 – Abordagem Filtro

A abordagem *wrapper* ocorre externamente ao algoritmo básico de aprendizagem, porém, utilizando tal algoritmo como uma caixa preta para analisar, a cada iteração, o subconjunto de atributos em questão – Figura 3. Em outras palavras, métodos *wrapper* geram um subconjunto candidato de atributos selecionado do conjunto de treinamento, e utilizam a precisão resultante do classificador induzido (neste trabalho foi utilizado regressão logística e redes neurais) 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. Esta abordagem avalia os atributos usando estimativas de precisão providas por algoritmos de aprendizado pré-determinados (Freitas A. A., 1998, p. 66).

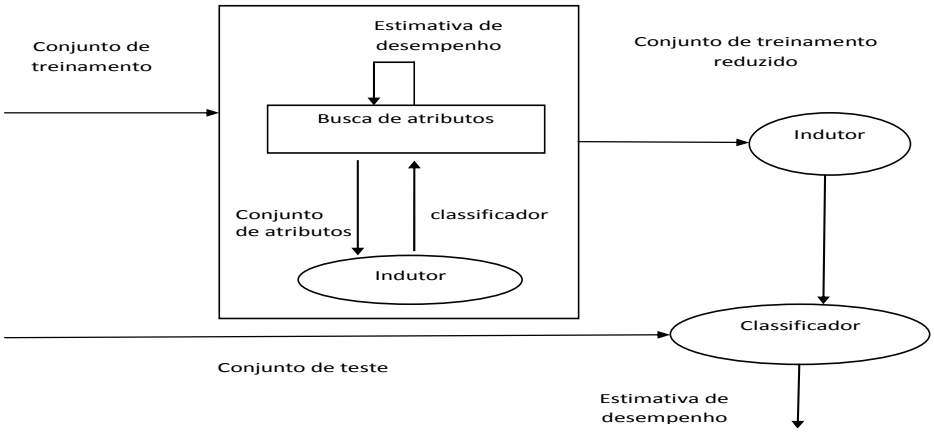

Figura 3 – Abordagem *Wrapper*

Os algoritmos de avaliação de atributos para seleção cobrem os principais métodos desenvolvidos para DM na última década.

O algoritmo de avaliação aplicado neste estudo foi o CFS. O algoritmo CFS (Seleção de atributos baseado em correlação), (Mark A. Hall and Geoffrey Holme, 2003) é um dos métodos que avaliam subconjuntos de atributos ao invés de avaliações individuais. O algoritmo avalia o subconjunto levando em conta a utilidade dos atributos individuais para predizer a classe junto com o nível de intercorrelação entre eles.

2. METODOLOGIA DA PESQUISA

Esta pesquisa de natureza empírica foi do tipo descritivo e quantitativo, compreendendo as empresas classificadas no SERASA e na BOVESPA como insolventes (concordatárias, em recuperação judicial e falidas), e compreendendo o período de 2004 a 2006.

2.1 A amostra em estudo

Para o objetivo deste estudo é necessário 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 problema é aqui entendido como sendo empresas classificadas no SERASA ou na BOVESPA como concordatária, em recuperação judicial ou falida. O segundo tipo diz respeito a 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 2004 a 2006.

Na amostra do segundo grupo, foram incluídas empresas "saudáveis", relacionadas ao primeiro da seguinte forma: para cada em presa do primeiro grupo, foram relacionadas duas empresas financeiramente saudável (empresas solventes), com tamanho do ativo, e pertencente ao mesmo setor de atividade econômica, tentando respeitar idêntica localização geográfica.

Com base nos critérios expostos, foram finalmente selecionadas 56 empresas insolventes e 112 empresas solventes totalizando 168 empresas.

2.2 Coleta dos dados

Neste estudo, tratando-se de uma pesquisa descritiva, o método de coleta de dados utilizado foi à pesquisa documental.

Segundo Gil (2002, p. 45), existe uma diferença entre pesquisa documental e bibliográfica, conforme descrito a seguir: "A diferença essencial entre ambas está na natureza das fontes. Enquanto a pesquisa bibliografia 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 2002 a 2006 das empresas em estudo junto ao SERASA e a BOVESPA.

O período escolhido foi o intervalo de tempo entre 2004 a 2006, de modo a dispor de informações que fossem, ao mesmo tempo, mais recentes e, supostamente, de pouca influência mais direta da nova lei de falências, que entrou em vigor em junho de 2005. Analisaram-se, então, os demonstrativos contábeis - Balanço Patrimonial e Demonstrativo de Resultado do Exercício - do ano do pedido de concordata ou falência e dos dois anos precedentes ao pedido.

A coleta de dados consistiu em levantar 20 (vinte) indicadores econômico-financeiros anuais das empresas selecionadas, no período de 2002-2006. Não se incluíram dados dos balanços consolidados, objetivando-se estudar as empresas singularmente. Os grupos de indicadores econômico-financeiros são os utilizados para análise das demonstrações contábeis: liquidez, endividamento, rentabilidade e lucratividade conforme Iudícibus (1998); Schrickel (1999); Matarazzo (2003); Pereira (2006).

2.3 Tratamento dos dados

Para realizar a análise dos dados deste estudo, buscou-se fazer análise das demonstrações contábeis dos anos de 2002 a 2006 das empresas que compõem o estudo. Os índices extraídos desta análise receberam tratamento de pré-processamento 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 *data mining* que foi aplicada é a de classificação, feita através de duas técnicas com metodologias distintas: *backpropagation* e regressão logística. No tratamento dos dados foi utilizado o software de *data mining* WEKA 3.5.6 (Witten, IH, Frank, 2005).

2.4 Avaliação dos resultados das abordagens de seleção dos atributos

Para avaliar os resultados das abordagens de seleção de atributos foram aplicados dois classificadores com metodologias distintas, regressão linear generalizada e *backpropagation,* ou seja, *r*egressão logística e redes neurais.
Regressão logística é uma técnica de regressão linear generalizada e que modela a probabilidade de alguns eventos ocorrerem como uma função linear em um grupo de variáveis preditoras. Esta regressão é bem adequada para grupo de dados que apresentem característica da distribuição de Poisson e são comumente modelados usando regressão de Poisson (Han & Kamber, 2006, p. 358). *Backpropagation* ou redes neurais artificiais são modelos matemáticos que se assemelham às estruturas neurais biológicas e que têm capacidade computacional adquirido por meio de aprendizado e generalização (Braga, Carvalho, & Ludemir, 2000 p, 45). Neste trabalho aplicaremos as redes neurais normalmente chamadas de *perceptrons de múltiplas camadas* (MLP, *multilayer perceptron*). Esta rede 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) (Haykin, 2001, p. 183).
Para avaliação da precisão dos modelos gerados pelos classificadores foram utilizadas matriz de confusão, validação cruzada e a medida F. A matriz de confusão é uma tabela onde são representados os T*p* (verdadeiros positivos), T*n* (verdadeiros negativos), F*p* (falsos positivos), F*n* (falso negativos). Validação cruzada permite que todos os dados da base de dados sejam utilizados para treinamento e teste, neste trabalho foi adotado 10 subconjuntos e a medida F é a média entre precisão e recall, medem a capacidade de reconhecer os exemplos negativos e positivos (Witten, IH, Frank, 2005).

3. 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.

3.1 Resultado da Seleção

3.1.1. Abordagem Filtro

Quinze foram as variáveis preditoras selecionadas pelo método de seleção de atributos baseado no algoritmo CFS, e avaliadas pela abordagem filtro conforme apresentadas no Quadro 1. Vale ressaltar de que os valores subscritos nas variáveis se referem ao período da variável. Por exemplo, LC_1 se refere a liquidez corrente de dois anos antes do ano da declaração da insolvência da entidade.

NOME DAS VARIÁVEIS	*ABREVIATURA*
1. Rentabilidade operacional sobre o ativo total	$ROAT_3$
2. Rentabilidade líquida sobre ativo total	$RLAT_3$
3. Saldo de tesouraria sobre ativo total	$STAT_3$
4. Giro do Ativo	GA_1
5. Margem bruta	MB_3
6. Liquidez Imediata	LI_3
7. Liquidez corrente	LC_2
8. Margem Operacional	MO_3
9. Liquidez corrente	LC_1
10. Endividamento Total sobre Pat. Liq.	$ETPL_3$
11. Margem Operacional Pós Receita Financeiras	$MORF_3$
12. Saldo de tesouraria sobre ativo total	$STAT_2$
13. Liquidez seca	LS_3
14. Rentabilidade líquida sobre o ativo total	$RLAT_2$
15. Margem líquida	ML_3

Quadro 1 – Variáveis selecionadas pela abordagem filtro.

3.1.2. Abordagem *wrapper*

Seis foram as variáveis preditoras selecionadas pelo método seleção de atributos *wrapper*, e apresentadas no Quadro 2:

NOME DAS VARIÁVEIS	ABREVIATURA
1. Saldo de tesouraria sobre ativo total	$STAT_1$
2. Endividamento Total sobre Pat. Liq.	$ETPL_3$
3. Margem Operacional Pós Receita Financeiras	$MORF_3$
4. Margem bruta	MB_3
5. Rentabilidade operacional sobre o ativo total	$ROAT_3$
6. Giro do Ativo	GA_1

Quadro 2 – Variáveis selecionadas pela abordagem *wrapper*.

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

A Tabela 1, abaixo, mostra a evolução de indicadores selecionados nos 3 últimos períodos anteriores à insolvência, para o caso do grupo das empresas insolventes.

Para o indicador referente a endividamento total sobre o patrimônio líquido (ETPL), verifica-se que suas médias vão-se deteriorando ao longo dos anos. Além disso, os desvios-padrões

reduzem acentuadamente para depois tomar algum fôlego, indicando uma relativa homogeneidade das características. A variação mais significativa se refere à passagem do período T-2 para o período T-1 (84,33%), mostrando o comprometimento da estrutura de capital do conjunto das empresas falidas. Esse resultado sugere a falta de resultados positivos nas ações para administrar as obrigações para com terceiros, valendo-se de prováveis trocas de ativos permanentes por obrigações, seja com terceiros seja com os proprietários e tentando, assim, uma sobrevida para poder sanar debilidades.

	Periodo T-2			Período T-1			Período T		
Varíavel	Média	DP	CV (%)	Média	DP	CV (%)	Média	DP	CV (%)
ETPL	7,46	30,58	409,91	2,02	4,79	237,12	1,33	7,28	547,36
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	201,42	-0,297	0,4638	126,79
ROAT	-0,11	0,3093	281,72	-0,255	0,5583	219,37	-0,291	0,4272	143,77
GA	5,69	24,45	429,70	1,39	3,69	265,46	1,93	4,00	207,25

Tabela 1 – Evolução dos indicadores no grupo das insolventes.
Nota - 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.

Comparando-se com a evolução do Endividamento Total sobre o Patrimônio Líquido (ETPL), pode-se supor que a maioria das empresas insolventes apresentou aumento relativo na conta lucros, relativo à margem, mas através de venda de bens. Talvez do ativo permanente, visando equilibrar a estrutura de capital, sobretudo aquelas de curto prazo para gerar uma relativa folga, mas comprometendo, provavelmente, a sua eficiência operacional. Com relação a este indicador, é interessante notar que o mercado, durante o período a que se referem os dados, passou por transformações relevantes no que se refere à facilitação de acesso ao crédito, mas com custos significativos. Provavelmente, as empresas insolventes não se adequaram, eficazmente, às novas facilidades ao crédito comprometendo demasiadamente os seus ativos operacionais.

A evolução da variável Saldo de Tesouraria sobre Ativo Total (STAT) é coerente com a situação de empresas em fase pré-falimentar, conforme Pereira, (2006, p. 236) 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 as 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 (ΔROAT = 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 uma acentuada queda para uma posterior melhora pouco relativa, pode-se supor de que houve uma desmobilização visando melhorar a estrutura de capital das empresas, mas pode-se supor também a perda da eficiência operacional se o índice for analisado em conjunto com o índice Margem Operacional após o Resultado Financeiro (MORF).

Pode-se supor que nas empresas insolventes estudadas ocorreram acentuados comprometimentos em sua estrutura de capital através do aumento do capital de terceiros,

talvez facilitado pela oferta acentuada de crédito no mercado durante o esse período estudado. Tais empresas tentando retomar uma melhor condição financeira acabaram levando a um comprometimento demasiado de sua capacidade operacional realizando desmobilizações, os índices estudados de rentabilidade, grau de imobilizado e de estrutura de capital facilitam essas deduções.

3.1.4 Análise descritiva - Evolução dos indicadores previsores nas empresas solventes selecionados pelas duas abordagens

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

	Periodo T-2			Período T-1			Período T		
Varíavel	Média	DP	CV (%)	Média	DP	CV (%)	Média	DP	CV (%)
ETPL	14,81	57,45	387,91	2,63	7,089	269,54	31,37	152,36	485,68
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	-0,0007	0,0967	1401,44	-0,012	0,1738	1448,33
ROAT	0,0015	0,277	18467	-0,0019	0,1598	8410,52	-0,006	0,087	1500,00
GA	10,28	35,32	343,57	2,30	5,26	43,20	20,95	101,40	484,00

Tabela 2 – Evolução dos indicadores no grupo das empresas solventes.

O inverso ocorre 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 estrutura de capital razoavelmente comprometida, a busca de rentabilidades razoáveis é abandonada, passando a ser uma estratégia secundária, resultando na aceleração do processo de insolvência.

3.2 Modelos de previsão

São apresentados nesta seção os resultados obtidos com a aplicação de técnicas de classificação, redes neurais e regressão logística, com o propósito de desenvolver modelos de insolvência, compostos pelas variáveis selecionados e pré-avaliados nas abordagens filtro e *wrapper*, e a seguir determinar os efeitos dessas abordagens nos classificadores estudados.

3.2.1 Modelo elaborado com o classificador regressão logística

O modelo elaborado com o classificador da regressão logística é composto pelos subconjuntos de variáveis selecionados e avaliados pela abordagem filtro, conforme abaixo:

$x = -3.3224 + 0.3303\ LC_1 + 1.4788\ GA_1 + 0.8859\ LS_3 + 0.4652\ ML_2 + 0.2371\ ML_3 + 0.3303\ LC_1 + 0.0147\ RLAT_3 + 0.0249\ LC_2 + 0.0147\ STAT_2 - 0.0103\ ETPL_3 + 0.0678\ MO_3 - 0.0774 RLAT_2.$

Grupo de Origem	*Insolventes*	*Solventes*	*Total*
Insolventes	42	14	56
Solventes	8	104	112
Medida F	0,65	0,79	
Classificação correta do grupo de origem (%)		86,90	

Tabela 3 - Classificação de resultados de previsão dos grupos composta pelos subconjuntos de atributos selecionados pela abordagem filtro aplicando o classificador regressão logistica.

Na Tabela 3 pode ser constatado que, nas empresas insolventes houve um índice de acertos de 75%. Já para as empresas solventes o índice de acertos foi de 92,8 % para as empresas solventes. A classificação correta do grupo de origem foi de 86,90%.

3.2.2 Modelo elaborado com o classificador redes neurais:

O modelo elaborado com o classificador redes neurais é composto pelos subconjuntos de variáveis selecionados e avaliados pela abordagem filtro, conforme abaixo:

$x = -1.307 + 1.344\ LI_1 + 0.1225\ ROAT_1 + 0.2691\ RLAT_1 - 0.5319\ LC_2 - 0.764\ GA_2 + 2.1108\ LI_3 + 1.0789\ LC_3 - 0.2078\ LS_3 - 0.3139\ EOAT_3 - 0.2161\ MB_3 - 0.3321\ GA_3 + 1.4838\ STAT_3$

Na Tabela 4 pode ser constatado que, nas empresas insolventes houve um índice de acertos de 89%. Já nas empresas solventes o índice de acertos foi de 93,75 % para as empresas solventes. A classificação correta do grupo de origem foi de 92,26%.

Grupo de Origem	*Insolventes*	*Solventes*	*Total*
Insolventes	50	6	56
Solventes	7	105	112
Medida F	0,6521	0,8849	
Classificação correta do grupo de origem (%)		92,26	

Tabela 4 - Classificação de resultados de previsão dos grupos composta pelas variáveis selecionadas pela abordagem filtro aplicando o classificador redes neurais.

Na Tabela 4 pode ser constatado que, nas empresas insolventes houve um índice de acertos de 89%. Já nas empresas solventes o índice de acertos foi de 93,75 % para as empresas solventes. A classificação correta do grupo de origem foi de 92,26%.

3.2.3 Modelo elaborado com o classificador regressão logística composto pelos subconjuntos de atributos selecionados e avaliados pela abordagem *wrapper*:

$x = -1.5395 + 0.1801\ GA_1 - 1.3201\ STAT_1 - 0.0015\ ETPL_3 - 0.1395\ MB_3 - 2.325\ MORF_3 - 4.7241\ ROAT_3$

Grupo de Origem	*Insolventes*	*Solventes*	*Total*
Insolventes	26	30	56
Solventes	7	105	112
Medida F	0,65	0,8	
Classificação correta do grupo de origem (%)		77,97	

Tabela 5 - Classificação de resultados de previsão dos grupos composta pelas variáveis selecionadas pela abordagem *wrapper* aplicando o classificador regressão logsitica.

Na Tabela 5 pode ser constatado que, nas empresas insolventes houve um índice de acertos de 46,42%. Já nas empresas solventes o índice de acertos foi de 93,75 % para as empresas solventes. A classificação correta do grupo de origem foi de 77,97%.

3.2.4 Modelo elaborado com o classificador redes neurais composto pelos subconjuntos de atributos selecionados e avaliados pela abordagem *wrapper*:

Grupo de Origem	*Insolventes*	*Solventes*	*Total*
Insolventes	25	31	56
Solventes	10	102	112
Medida F	0,55	0,64	
Classificação correta do grupo de origem (%)		75,59	

Tabela 6 - Classificação de resultados de previsão dos grupos composta pelas variáveis selecionadas pela abordagem *wrapper* aplicando o classificador redes neurais.

Na Tabela 6 pode ser constatado que, nas empresas insolventes houve um índice de acertos de 44,64%. Já nas empresas solventes o índice de acertos foi de 91,08 % para as empresas solventes. A classificação correta do grupo de origem foi de 75,59%.

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

Abordagem da avaliação de subconjuntos de atributos selecionados	Classificador	N° de acertos	N° de erros	Acertos %	Erros %
Filtro	Reg. logística	146	22	86,90	13,10
Filtro	Redes neurais	155	13	92,26	7,74
Wrapper	Reg. logística	131	37	77,97	22,03
Wrapper	Redes neurais	127	41	75,59	24,41

Tabela 7: Classificação de resultados de previsão dos modelos.

Pela Tabela 7 podemos comparar as resultados gerados pelos modelos compostos subconjuntos de atributos selecionados e avaliados pelas abordagens filtro e *wrapper*. Os modelos elaborados pelos subconjuntos avaliados pela abordagem filtro apresentaram melhores índices de classificação, tanto para regressão logística como pra redes neurais. Cabe ressaltar que a melhor performance foi com o classificador redes neurais (92,26%) com as melhores medidas Fs.

4. CONCLUSÃO

Nessa amostra de dados os resultados obtidos na classificação foram melhores quando aplicado a abordagem filtro para seleção de subconjuntos de atributos com a utilização dos classificadores tanto de redes neurais (92,26%) como de regressão logística (86,90%). As variáveis selecionadas foram capazes de realizar bem a tarefa de classificação. A redução da dimensionalidade conduz não somente a economia de custo computacional (e a aquisição de dados), mas também ao ganho de desempenho e de exatidão nas etapas de DM. A abordagem filtro selecionou um subconjunto de atributos mais eficientes na capacidade de predição do que a abordagem *wrapper,* permitindo que os classificadores obtivessem desempenhos bem melhores nos modelos de previsão. Não se pode afirmar que a abordagem filtro seja geralmente a mais preferível, a melhor recomendação a ser dada é que a redução da dimensionalidade deve ocorrer como parte da tarefa de classificação e na seleção de subconjuntos com os dados contábeis brasileiros, tendo a abordagem filtro se mostrado a mais eficaz neste estudo. Foi demonstrada também a importância da explicitação da etapa de

avaliação da seleção de atributos para a obtenção de melhores resultados na aplicação de técnicas de *data mining* para previsão de insolvência usando dados contábeis de empresas brasileiras. A conclusão obvia a respeito da usabilidade da abordagem filtro é que deve ser usada para avaliar subconjuntos de atributos que irão compor os modelos preditivos.

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. Esses modelos preditivos podem ser incorporados como procedimentos analíticos para avaliar a probabilidade de insolvência. São interessantes para Bancos, investidores, governos, auditores, gerentes, fornecedores, empregados e muitos outros poderem avaliar, com razoável antecedência, se há um problema de insolvência em andamento.

Das variáveis selecionadas pelas duas abordagens cinco foram "determinadas" pelos dois métodos estudados (ETPL, MORF, MB, ROAT, GA) podendo levar a conclusão de que são aquelas com melhor capacidade para caracterizar as classes de empresas solventes e insolventes. Portanto, devem compor modelo de previsão de insolvência.

Uma questão relativamente surpreendente nos modelos diz respeito à natureza das variáveis presentes nas equações finais. Em todas as equações estão presentes diferentes indicadores capazes de explicar a diferença entre empresas solventes e insolventes, tais como margem operacional após o resultado financeiro, rentabilidade e saldo de tesouraria sobre ativo total. Este resultado assume importância ainda maior ao considerar as modificações ocorridas no mercado, que se tornaram mais competitivas fortalecendo empresas que toleraram rentabilidades menores para manter a sua continuidade.

Apesar da qualidade dos dados contábeis serem ainda muitas vezes questionada em termos de utilização na construção de modelos de previsão de falência, os resultados obtidos foram bastante satisfatórios, da ordem de 85% ou mais de acerto, o que vem evidenciar o conteúdo de informação que esses dados proporcionam, ainda que seja para previsões.

É possível que, com o emprego de alguns desses dados, se possam prever a saúde financeira de uma empresa. Por exemplo, podem-se ponderar os diversos índices apresentados para obter um modelo que construa um índice de risco de crédito. A ponderação poderá ser feita com a técnica de *data mining*. Isso sugere que *data mining* apoiada em indicadores contábeis é uma ferramenta útil para prever concordatas de empresas, podendo ser utilizada para estabelecer *scores* associados a risco de crédito.

REFERÊNCIA BIBLIOGRÁFICA

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

ALTMAN, E. I., G. MARCO & F. VARETTO (1994), "Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience)", Journal of Banking and Finance, vol. 18, 505-529.

BACK, B., Laitinen, T., KAISA, S. (1996), Neural Networks and Genetic Algorithms for Bankruptcy Predictions. Expert Systems With Applications, Vol. 1 I, No. 4.

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, 38, 2006.

BEAVER, W. H. (1967), "Financial Ratios as Predictors of Failure", Journal of Accounting Research - Supplement, pp. 71-111.

BLUM, M. (1974), "Failing Company Discriminant Analysis", Journal of Accounting Research, Spring

BRAGA, A. P., A. P. L. F. Carvalho, & T. B. Ludemir (2000). Redes Neurais Artificiais: teoria e aplicações. Rio de Janeiro, RJ: Livros Técnicos e Científicos.
BRAGANÇA, L. A. de e S. L. de BRAGANÇA (1984), "Previsão de concordatas e falências no Brasil", Anais do VII Congresso ABAMEC.
BROCKETT, P. L., W. W. COOPER, L. GOLDEN & X. XIA, (1997), "A case study in applying neural networks to predicting insolvency for property and casualty insurers", Journal of the Operational Research Society, vol. 48, p. 1153-1162, abril.
CHYE K. H., CHIN. T.W., PENG G. C. (2004). Credit Scoring Using Data Mining Techniques Singapore Management review.
CLARK, P. & T. Niblett (1989). The CN2 induction algoritm. Machine Learning 3(4), 261-283.
DEAKIN, E. B.(1972), "A Discriminant Analysis of Prediction of Business Failure", Journal of Accouting Research, 167-179, Spring
EDMISTER, R. O. (1972), "An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction", Journal of Financial and Quantitative Analysis, vol. 7, 1477-1493, março.
EISENBEIS, R. A. (1997), "Pitfalls in the application of discriminant analysis in business, finance, and economics, The Journal the Finance, vol. 32, n. 3, p. 875-900, junho.
FAYYAD, U., G. Piatetsky-Shapiro, & P. Smyth (1996). From data mining to knowledge discovery: an overview. Em Advances in Knowledge Discovery & Data Mining, pp. 1-34.
FITZPATRICK, P. J. (1932), "A Comparison of ratios of successful industrial enterprises with those of failed firms", Certified Public Accountant, October pp. 598-605, November p. 656-62, and December p. 727-31.
FREITAS A. A. Data mining and knowlwdge discovery with evolutionary algorithms. Springer-Verlag Berlin Heidelberg New York, 1998.
GIL, Antônio Carlos. Como elaborar Projetos de Pesquisa. 4 ed. São Paulo: Atlas, 2002.
HALL M. A. Correlation-based Feature Subset Selection for Machine Learning, 1998.
Jonh, G. H., Kohavi, R., Pfeger, K. Irrelevant features and the subset selection problem. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 121-129, 1994.
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 (2001). Redes neurais: princípios e práticas. 2. Ed. Porto Alegre: Bookman.
HORTA, R. A. M. (2001), Utilização de indicadores contábeis na previsão de insolvência – análise empírica de uma amostra de empresas comerciais e industriais brasileiras, Dissertação de Mestrado inédita, Programa de Mestrado em Ciências Contábeis, UERJ, 2001.
IUDÍCIBUS, Sérgio de. Análise de Balanços. 6ª Ed. São Paulo: Atlas, 1995.
KANITZ, S. C. (1978), Como prever falências, São Paulo: Mc Graw-Hill do Brasil.
KASZNAR, I. K. (1986), "Falências e Concordatas de Empresas: modelos teóricos e estudos empíricos". Dissertação de Mestrado EPGE – FGV, Rio de Janeiro.
KOHAVI, R. and JOHN, G. H. (1997). Wrappers for feature subset selection. *Artif. Intell.,* 97(1-2):273-324. Citado nas paginas 17, 18, 22, 24, 26, 32, e 42.
LENNOX, C. (1999), "Identifying failing companies: a reevaluation of the logit, probit and d.a. aprproaches", Journal of Economics and Business, vol.51, p. 347-364.
LIU, H. and MOTODA, H. Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Massachusetts, 1998.
MARK A. Hall and GEOFFREY Holme, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining, 2003.

MATARAZZO, Dante C. Análise Financeira de Balanços. 6ª Ed. São Paulo: Atlas, 2003.
PEREIRA, José da Silva. Gestão e Análise de Risco de Crédito. 5ª Ed. São Paulo: Atlas, 2006.
PIRAMUTHU S. (2006). On preprocessing data for financial credit risk evaluation. Expert Systems with Applications 30 489-497.
REZENDE, Solange Oliveira (Org.). Sistemas inteligentes: fundamentos e aplicações. Barueri, SP: Manole, 2005.
SANVICENTE, A. Z. & A. M. A. F. MINARDI (2000), "Identificação de indicadores contábeis significativos para previsão de concordata de empresas", disponível: como http://www.risktech.br/artigos/artigos técnicos /index.html,
SAUNDERS, A. Allen, L. DeLong, G. , Issues in the credit risk modeling of retail markets. Journal of Bank & Finance 28, 2004.
SCHRICKEL, Wolfgang K. Demonstrações Financeiras. 2ª Ed. São Paulo: Atlas, 1999.
SHIRATA, Cindy Yoshiko. Financial ratios as predictors of bankruptcy in Japan:na empirical research. Fev. 14, 2001. Disponível Http://www.shirata.net/apira98.html.
SILBERSCHATZ, A. & A. Tuzhilin (1995). On subjective measures of intestingness in knowledge Discovery. Proceedings of the First International Conference on Knowledge Discovery and Data Mining 1, 275-281.
SOMOL, P., BAESENS, B., PUDIL, P., VANTHIENEN J. (2005) Filter versus wrapper-based feature selection for credit scoring. International Journal of Intelligent Systems. Volume 20, Issue 10 , Pages 985 – 999.
SUN, L.,Shenoy, P.P.(2007) Using Bayesian Networks for Bankruptcy Prediction: Some Methodological Issues. European Journal of Operational Research, 180(2), 2007, 738—753.
SUN, Z., BEBIS, G., & MILLER, R. (2004). Object detection using feature subset selection. Pattern Recognition, 27, 2165–2176.
WITTEN, I.H., Frank .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.