RAM. Revista de Administração Mackenzie
ISSN: 1518-6776
revista.adm@mackenzie.com.br
Universidade Presbiteriana Mackenzie
Brasil

MATHIASI HORTA, RUI AMÉRICO; DOS SANTOS ALVES, FRANCISCO JOSÉ; AZEVEDO DE CARVALHO, FREDERICO ANTÔNIO
SELEÇÃO DE ATRIBUTOS NA PREVISÃO DE INSOLVÊNCIA : APLICAÇÃO E AVALIAÇÃO USANDO DADOS BRASILEIROS RECENTES
RAM. Revista de Administração Mackenzie, vol. 15, núm. 1, enero-febrero, 2014, pp. 125-151
Universidade Presbiteriana Mackenzie
São Paulo, Brasil

Disponível em: http://www.redalyc.org/articulo.oa?id=195429840006

- Como citar este artigo
- Número completo
- Mais artigos
- Home da revista no Redalyc

Sistema de Informação Científica
Rede de Revistas Científicas da América Latina, Caribe , Espanha e Portugal
Projeto acadêmico sem fins lucrativos desenvolvido no âmbito da iniciativa Acesso Aberto

• RAM, REV. ADM. MACKENZIE, *15*(1) • SÃO PAULO, SP • JAN.-FEV. 2014 • ISSN 1518-6776 (impresso) • ISSN 1678-6971 (*on-line*) •
Submissão: 18 jan. 2012. **Aceitação:** 7 out. 2013. **Sistema de avaliação:** às cegas dupla (*double blind review*).
UNIVERSIDADE PRESBITERIANA MACKENZIE. Aureliano Angel Bressan (Ed. Seção), Walter Bataglia (Ed.), 125-151.

SELEÇÃO DE ATRIBUTOS NA PREVISÃO DE INSOLVÊNCIA: APLICAÇÃO E AVALIAÇÃO USANDO DADOS BRASILEIROS RECENTES

RUI AMÉRICO MATHIASI HORTA
Doutor em Engenharia Civil pelo Departamento de Engenharia da Universidade Federal do Rio de Janeiro (UFRJ).
Professor do Departamento de Finanças e Controladoria da Universidade Federal de Juiz de Fora (UFJF).
Rua José Lourenço Kelmer, s.n., campus *Universitário, São Pedro, Juiz de Fora – MG – Brasil – CEP 36036-900*
E-mail: *rui.horta@ufjf.edu.br*

FRANCISCO JOSÉ DOS SANTOS ALVES
Doutor em Ciências Contábeis pela Faculdade de Economia e Administração da Universidade de São Paulo (USP).
Professor do Departamento de Ciências Contábeis da Universidade Estadual do Rio de Janeiro (UERJ).
Rua São Francisco Xavier, 524, Maracanã, Rio de Janeiro – RJ – Brasil – CEP 20550-013
E-mail: *francisco.jose.alves@terra.com.br*

FREDERICO ANTÔNIO AZEVEDO DE CARVALHO
Doutor em Sciences Économiques pela Université Catholique de Louvain (Bélgica).
Professor do Departamento de Ciências Contábeis da Universidade Federal do Rio de Janeiro (UFRJ).
Avenida Pedro Calmon, 550, Cidade Universitária, Rio de Janeiro – RJ – Brasil – CEP 21941-901
E-mail: *fdecarv@gmail.comr*

Este artigo pode ser copiado, distribuído, exibido, transmitido ou adaptado desde que citados, de forma clara e explícita, o nome da revista, a edição, o ano e as páginas nas quais o artigo foi publicado originalmente, mas sem sugerir que a RAM endosse a reutilização do artigo. Esse termo de licenciamento deve ser explicitado para os casos de reutilização ou distribuição para terceiros. Não é permitido o uso para fins comerciais.

RESUMO

Previsão de falências pode ter grande utilidade para instituições financeiras e não financeiras no que se refere a tomar, antecipadamente, as melhores decisões possíveis quanto a empréstimos ou investimentos. Na literatura específica, muitos modelos de previsão de falência têm feito uso de técnicas de *data mining* (mineração de dados). O pré-processamento é passo importante para selecionar dados de boa qualidade para utilização em operações de mineração. Mesmo assim, apesar de a seleção de atributos poder ser muito benéfica para pré-selecionar dados representativos visando melhorar o desempenho da previsão final, não se sabe que método de seleção é o melhor. Este trabalho tem como objetivo principal comparar as duas abordagens mais utilizadas de avaliação de subconjuntos de atributos: Filtro e *Wrapper*. Apesar de serem fundamentadas em técnicas de mineração de dados e muito utilizadas na etapa de seleção de atributos em modelos de previsão de insolvência, essas técnicas são muito pouco utilizadas para tratar dados obtidos em demonstrativos contábeis de empresas brasileiras. Por isso, a base empírica deste estudo consiste em uma amostra de empresas comerciais e industriais brasileiras, coletando-se dados relativos ao período 2004-2011. Os resultados indicaram que, na amostra estudada, a abordagem Filtro foi a mais eficiente, fornecendo melhores resultados de classificação tanto para a técnica de regressão logística (91,80%), quanto para redes neurais (93,98%). Foi demonstrada, ainda, a importância da explicitação da etapa de avaliação da seleção de atributos para a obtenção de melhores resultados em aplicações de técnicas de mineração de dados na previsão de insolvência. Uma conclusão específica a respeito das vantagens da abordagem Filtro aponta que ela pode ser a preferida para avaliar os atributos que irão compor os modelos preditivos.

PALAVRAS-CHAVE

Índices econômico-financeiros. Previsão de insolvência. Mineração de dados. Seleção de atributos. Abordagens Filtro e *Wrapper*.

1 INTRODUÇÃO

A capacidade de gerar e coletar dados tem aumentado enormemente nos últimos anos, devido ao uso mais acessível das ferramentas de computação. Uma consequência desta tendência é a capacidade de discriminar a informação útil e relevante no processo de tomada de decisão. Na tomada de decisões no setor dos negócios, frequentemente esta situação se evidencia pela abundância de dados, disponíveis em várias formas e em um número crescente de fontes, entre as quais os dados contábeis. Em tais situações, muitos usuários podem ser incapazes de aproveitar a riqueza de dados à sua disposição, destacando, então, a utilidade de técnicas mais poderosas para processamento de informações, tais como a mineração de dados (MD).

Essas técnicas podem ser aplicadas para a extração de informações não triviais implícitas, previamente desconhecidas e potencialmente úteis a partir de dados empíricos, além de poderem reduzir os problemas gerados pela excessiva quantidade de dados. A resultante redução da dimensionalidade dos problemas de aplicação pode trazer diversos benefícios, seja para a compreensão do problema e de seus resultados, seja em relação aos custos envolvidos (por exemplo, os computacionais).

A mineração de dados tornou-se proeminente no final da década de 1990, apresentando forte ênfase na combinação de conjuntos de dados no intuito de capturar padrões muito sutis ou muito complexos para serem detectados somente por analistas de dados (Kreuze, 2001).

Relatórios da contabilidade são exemplo de um rico potencial para a extração de informações não triviais e busca de padrões. Com o rápido crescimento e importância de dados originados em relatórios contábeis, surgem oportunidades significativas para determinar a capacidade preditiva de tais informações (Hassan & Marston, 2010). Um tema de especial interesse para ambas as áreas – Contabilidade e Mineração de Dados – é a modelagem para previsão de insolvência de empresas.

Entre os vários estudos de revisão da literatura sobre aplicação de MD à solução de problemas de previsão de falência para instituições financeiras e não financeiras, pode-se destacar o de Verikas, Kalsyte, Bacauskiene e Gelzinis (2010), que acentua a importância de uma das etapas de pré-processamento de dados para a modelagem, a saber, a seleção de atributos. O objetivo deste artigo é precisamente comparar as abordagens Filtro e *Wrapper* para avaliação prévia de subconjuntos de atributos que venham a ser considerados (e eventualmente selecionados) em contextos de elaboração de modelos de previsão de insolvência. A ilustração empírica da análise comparativa está apoiada em indicadores econômico-financeiros obtidos em demonstrativos contábeis de empresas brasileiras, cobrindo o período de 2002 a 2011.

A principal contribuição do artigo é considerar explicitamente a fase de seleção das variáveis preditivas à luz de duas das mais importantes abordagens de avaliação. Alguns autores (Shirata, 2001; Wu, Fang, & Goo, 1996; Piramuthu, 2006; Tsai & Cheng, 2012) têm chamado atenção para a importância do processo de seleção de atributos 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. Outra contribuição reside na conclusão aqui obtida de que os dados contábeis brasileiros permitem construir modelos de previsão de falência com muito boa capacidade preditiva.

O trabalho está organizado em cinco seções, incluindo esta introdução. Na próxima seção apresentam-se os fundamentos temáticos do estudo, a partir de referências selecionadas na literatura específica. A terceira seção descreve os procedimentos metodológicos, enquanto que a quarta aborda os resultados obtidos. Os comentários conclusivos aparecem na quinta e última seção.

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 possibilita obter uma classificação confiável quanto ao futuro da "saúde financeira" das empresas. Em princípio, sua aplicabilidade está voltada às operações de curto prazo, dado que a insolvência está mais relacionada à perda da capacidade de endividamento do que ao desempenho operacional.

2.1 PREVISÃO DE INSOLVÊNCIA

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 1970 (Blum, 1974; Deakin, 1972; Edmister, 1972; Kanitz, 1978, entre tantos outros), em seguida aos trabalhos pioneiros de Beaver (1968) e de Altman (1968).

A partir dos anos 1990, 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 renovaram o interesse pela análise e pela previsão da insolvência de empresas, estimulando inúmeras pesquisas em diferentes mercados nacionais (Altman, Marco, & Varetto, 1994; Back, Laitinen, & Kaisa, 1996; Brockert, Cooper, Golden, & Xia, 1997; Eisenbeis, 1997; Lennox, 1999; Härdle, Moro, & Schäfer, 2005, entre outros).

No Brasil, a análise da insolvência de empresas com objetivos preditivos desenvolveu-se de modo significativo a partir dos anos 1980 (Pereira, 2006; Bragança & Bragança, 1984; Kasznar, 1986; Sanvicente & Minardi, 2000; Horta, 2001; Mario, 2005; Guimarães & Moreira, 2008), seguindo o caminho aberto por Kanitz (1978).

No levantamento deste referencial teórico, constatou-se a prática de utilizar variáveis previamente relacionadas em pesquisas anteriores (Kanitz, 1978; Bragança & Bragança, 1984; Sanvincente & Minardi, 2000, Guimarães & Moreira, 2008). Este procedimento pode desconsiderar especificidades, como os fatores culturais ou as legislações, por exemplo, tributária, fiscal ou societária de cada país.

A despeito da maior disponibilidade de novas técnicas de modelagem, parte da literatura específica aponta vários problemas relacionados à aplicação dessas técnicas na previsão de insolvência (Balcaen & Ooghe, 2006). Alguns desses problemas foram categorizados em tópicos, como: 1. a definição da variável dependente como ("apenas") dicotômica; 2. a seleção da amostra; 3. a não estacionariedade e outras instabilidades dos dados; 4. o uso de informações contábeis anuais; 5. a dimensão temporal; e 6. a seleção das variáveis independentes (ou preditivas). É precisamente este último problema que é abordado neste trabalho.

2.2 MINERAÇÃO DE DADOS (MD)

A definição hoje aceita pela maioria dos pesquisadores de MD foi elaborada por Fayyad, Piatetsky e Smith (1996, p. 41), ao afirmarem que: "Extração de Conhecimento em Bases de Dados é o processo de identificação de padrões embutidos nos dados que sejam válidos, novos, potencialmente úteis e compreensíveis". O processo de identificação de padrões em MD é dividido em três grandes etapas (Rezende, 2005): pré-processamento, extração de padrões e pós-processamento.

O *pré-processamento* caracteriza-se pela adequação dos dados para a extração de conhecimento. Diversas adequações nos dados podem ser executadas na etapa de pré-processamento, entre elas o tratamento de valores desconhecidos, a identificação e descrição de valores extremos, o tratamento de conjuntos de dados com classes desbalanceadas e a seleção de atributos, entre outras.

A *extração de padrões* compreende a escolha da tarefa de MD 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.

As tarefas possíveis para um algoritmo de extração de padrões podem ser agrupadas em três grupos de atividades: 1. descrição e visualização; 2. associação e clusterização (ou agrupamento); e 3. classificação e estimação (predição) (Chye, Chin, & Peng, 2004). A descrição e a visualização contribuem para a compreensão

de qualquer conjunto de dados, sobretudo quando a quantidade é bem grande, detectando-se padrões ocultos nos dados, especialmente naqueles em que, além de numerosos, são também "complicados" por conterem interações complexas e não lineares. Ambas as atividades são geralmente executadas antes da modelagem, tentando representar e compreender melhor os dados. Na associação, o objetivo é determinar a relação entre as variáveis. Na 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 "razoavelmente" distintos. Finalmente, a mais comum e importante aplicação em MD provavelmente envolve classificação e estimação com finalidade de predição, por vezes referidas como modelagem. Classificação refere-se à previsão de um objetivo em que a variável é de natureza categórica.

Para a construção de modelos de previsão de insolvência, as técnicas de modelagem preditiva são as mais relevantes. Neste caso da modelagem preditiva, a MD inclui técnicas estatísticas tradicionais, como análise discriminante múltipla e regressão logística, mas também pode incluir métodos não tradicionais, oriundos ou desenvolvidos nas áreas de inteligência artificial e aprendizado de máquina. Entre esses métodos não tradicionais, os dois mais importantes são redes neurais e árvores de decisão (Chye *et al.*, 2004). Neste trabalho serão utilizadas análise de regressão logística e redes neurais, que são amplamente utilizadas na literatura específica e serão mais bem definidas.

A extração pode gerar uma quantidade enorme de padrões, muitos dos quais podem não ser relevantes para o usuário.

No *pós-processamento* podem-se aplicar técnicas de apoio no sentido de fornecer aos usuários apenas os padrões "mais interessantes", por meio de medidas *ex post* relacionadas a desempenho e qualidade.

Realizadas as três etapas, na grande maioria dos estudos sobre modelagem de falência segue-se o processo de identificação de padrões eventualmente presentes nos dados. Como já foram comentados, esses padrões devem ser válidos, novos, potencialmente úteis e compreensíveis (Fayyad *et al.*, 1996).

2.3 SELEÇÃO DE ATRIBUTOS

A seleção de atributos (SA) é uma etapa muito relevante na elaboração de modelos de previsão de insolvência. A SA desempenha papel essencial nesses modelos, sendo frequentemente realizada como etapa de pré-processamento. Na grande maioria dos estudos, parte-se de um conjunto inicial de variáveis que, frequentemente, são escolhidas em razão de sua popularidade na literatura específica ou a seu sucesso preditivo em pesquisas precedentes.

Os objetivos da seleção de atributos em modelos de previsão de insolvência são para Piramuthu (2006):

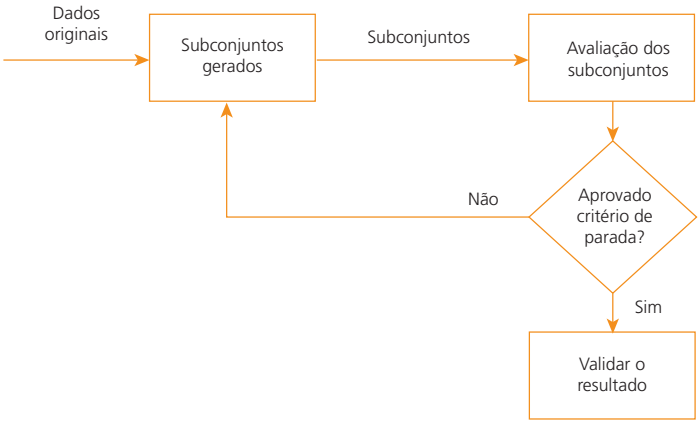
1. o desenvolvimento de modelos compactos;
2. o uso e refinamento do modelo de classificação para fins de avaliação; e
3. a identificação de índices financeiros relevantes.

Em problemas típicos de classificação, um conjunto de variáveis independentes deve compor um modelo que tenha a capacidade de categorizar o mais corretamente possível objetos de interesse em classes de pertinência futura (Piramuthu, 2006).

Os algoritmos usados para seleção de atributos podem ser distinguidos segundo a forma de perfazer duas atividades principais: a busca do subconjunto de atributos e a avaliação dos subconjuntos de atributos encontrados, como ilustra a Figura 1.

FIGURA 1

PASSOS NA SELEÇÃO DE ATRIBUTOS



Fonte: Liu e Motoda (1998, p. 38).

2.3.1 Busca do subconjunto de atributos

A partir de todos os atributos disponíveis, seleciona-se um subconjunto de variáveis relevantes com apoio de um (sub)algoritmo de busca. A concepção dessa busca é explorar o espaço de subconjuntos de atributos, percorrendo todo o conjunto e ajustando o número dos melhores atributos encontrados de modo a permitir o controle da busca.

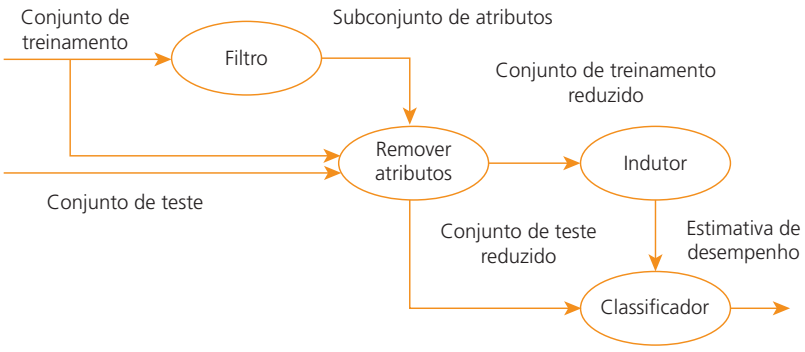
2.3.2 Avaliação do subconjunto de atributos

Avaliar o subconjunto de atributos selecionados é medir quão bom é determinado atributo segundo dado critério de avaliação (por exemplo, informação, distância, dependência, consistência, precisão, entre outros). Em outras palavras, avaliar como o atributo interage com o algoritmo de aprendizado. Essa interação pode ser subdividida, basicamente, em duas abordagens principais: Filtro e *Wrapper* (Kohavi & George, 1997).

A abordagem Filtro, utilizada para filtrar atributos durante o processo de pré-processamento, não depende do algoritmo de aprendizado que utilizará esse mesmo conjunto de atributos. A ideia é "filtrar" – isto é, deixar de lado – os atributos irrelevantes, segundo algum critério, antes de o aprendizado ocorrer. Essa etapa do pré-processamento considera características gerais do conjunto de dados para selecionar alguns atributos e excluir outros. É neste sentido que os métodos de Filtro são independentes do algoritmo de aprendizado, que simplesmente receberá como ponto de partida o conjunto de dados que foi construído utilizando somente o subconjunto de atributos "importantes" selecionados pelo Filtro. Em suma, a meta é selecionar um subconjunto de atributos que preservem a informação contida no conjunto completo dos atributos, ou seja, separar os atributos irrelevantes segundo algum critério, antes de o aprendizado ocorrer (John, Kohavi, & Pfeger, 1994).

Figura 2

ABORDAGEM FILTRO



Fonte: Freitas (1998, p. 67).

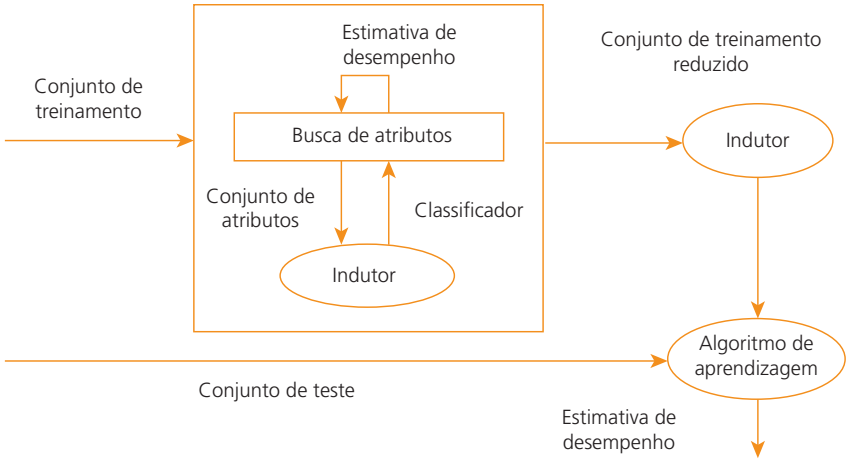
Os métodos *Wrapper* geram um subconjunto "candidato", contendo atributos selecionados no conjunto de treinamento, e utilizam a precisão resultante do classificador para avaliar o subconjunto de atributos "candidatos"; neste trabalho

foram utilizados dois classificadores, a Regressão Logística e as Redes Neurais. Esse processo é repetido para cada subconjunto de atributos até que um critério de parada, determinado pelo usuário, seja satisfeito. Esta abordagem avalia os atributos usando estimativas de precisão fornecidas por algoritmos de aprendizado predeterminados (Freitas, 1998, p. 66).

Os algoritmos de avaliação de atributos para seleção cobriram os principais métodos desenvolvidos para MD na última década. Neste estudo, foi aplicado o algoritmo CFS – seleção de atributos baseado em correlação (Hall, 1999; Hall & Holmes, 2003). Este algoritmo, que é um dos métodos que avaliam subconjuntos de atributos em vez de atributos individuais, avalia um subconjunto de atributos por meio da capacidade individual de previsão de cada recurso, juntamente com o grau de redundância entre eles.

FIGURA 3

ABORDAGEM *WRAPPER*



Fonte: Freitas (1998, p. 68).

3 METODOLOGIA

Esta pesquisa, do tipo descritivo e quantitativo, apoia-se empiricamente em uma amostra de empresas classificadas – tanto no Serviço de Proteção ao Crédito (Serasa), quanto na Bolsa de Valores de São Paulo (Bovespa) – como solventes ou insolventes (concordatárias, em recuperação judicial e falidas), relativamente ao período de 2004 a 2011.

3.1 AMOSTRA E COLETA DOS DADOS

Este estudo baseia-se em dois tipos de amostras. O primeiro tipo refere-se a um conjunto de empresas chamadas de "problemáticas", por apresentaram "problemas de insolvência" em determinado período. Tais "problemas" são aqui entendidos, segundo a classificação Serasa ou Bovespa, como "concordata", "recuperação judicial" ou "falência".

O segundo tipo de amostra diz respeito a empresas "saudáveis" por serem precisamente contrário ao do primeiro grupo, ou seja, empresas que não apresentaram "problemas de insolvência", no Serasa ou Bovespa, no período de tempo estudado. Nesta amostra do segundo grupo, foram intencionalmente incluídas empresas "saudáveis" usando o seguinte procedimento: para cada empresa do primeiro grupo, foram relacionadas duas empresas financeiramente saudáveis (empresas solventes) com tamanho do ativo "comparável" e pertencentes ao mesmo setor de atividade econômica, tentando, ainda, sempre que possível, replicar a localização geográfica.

Com base nos critérios expostos, foram finalmente selecionadas 61 empresas insolventes e 122 empresas solventes, totalizando 183 empresas.

O período escolhido foi o intervalo de tempo entre 2004 a 2011, de modo a dispor de informações que fossem, ao mesmo tempo, mais recentes e que, supostamente, refletissem influência direta da nova lei de falências, que entrou em vigor em junho de 2005. Considerando que os modelos utilizam alguns padrões de defasagens (*lags*) temporais, os dados coletados, de natureza quantitativa, têm origem em balanços e demonstrativos de resultado dos anos de 2002 a 2011 para as empresas amostradas conforme explicado antes. Foram efetivamente examinados, então, os demonstrativos contábeis – Balanço Patrimonial e Demonstrativo de Resultado do Exercício – do ano do pedido de concordata ou falência e dos dois anos anteriores ao pedido.

Assim, devido aos *lags* temporais, a coleta de dados consistiu em levantar vinte indicadores econômico-financeiros anuais das empresas selecionadas, no período de 2002-2011. Não se incluíram dados dos balanços consolidados, visto que o objetivo era estudar as empresas singularmente. Os grupos de indicadores econômico-financeiros são os tradicionalmente utilizados para análise das demonstrações contábeis: Liquidez, Endividamento, Rentabilidade e Lucratividade (Iudícibus, 1998; Schrickel, 1999; Matarazzo, 2003; Pereira, 2006).

Para o tratamento dos dados coletados, procedeu-se à análise das demonstrações contábeis dos anos de 2002 a 2011 das empresas amostradas. Os índices extraídos desta análise foram tratados, na etapa de pré-processamento, por meio de técnicas de seleção de atributos e de mineração de dados (MD). A seleção de atributos serve para identificar um subconjunto de atributos relevantes, com o

objetivo de caracterizar as empresas pré-classificadas como solventes ou insolventes. A técnica de MD aqui aplicada foi a de classificação, utilizando duas metodologias distintas, chamadas *backpropagation* e regressão logística. Nesse tratamento foi utilizado o *software* livre de *data mining* Weka, versão 3.5.6 (Witten & Frank, 2011).

3.2 AVALIAÇÃO DOS RESULTADOS DA SELEÇÃO DOS ATRIBUTOS

Para avaliar os resultados das abordagens de seleção de atributos foram aplicados dois classificadores: regressão logística e redes neurais artificiais (*backpropagation*). Regressão logística é uma técnica de regressão linear generalizada em que a variável dependente é categórica, postulando que a probabilidade de alguns eventos ocorrerem é função linear de um grupo de variáveis preditoras. *Backpropagation* ou redes neurais artificiais são modelos matemáticos que se assemelham às estruturas neurais biológicas, cuja capacidade preditiva baseia-se em aprendizado e generalização (Braga, Carvalho, & Ludemir, 2007, p. 45). Neste trabalho, aplicam-se as redes neurais normalmente chamadas de *perceptrons de múltiplas camadas* (MLP, *multilayer perceptron*). Essas redes consistem em um conjunto de "unidades sensoriais": os nós iniciais, que constituem a camada de entrada; uma ou mais camadas ocultas de nós computacionais; e, finalmente, uma camada de saída, também composta de nós computacionais. O sinal de entrada propaga-se para frente pela rede, camada por camada (Haykin, 2001, p. 183).

Para avaliação da precisão dos modelos gerados pelos classificadores foram utilizadas Matriz de Confusão, Validação Cruzada, Medida F e Área ROC. A Matriz de Confusão é uma tabela em que são representados os T*p* (verdadeiros positivos), T*n* (verdadeiros negativos), F*p* (falsos positivos), F*n* (falso negativos), e que permite calcular as percentagens de classificações corretas e incorretas. A Validação Cruzada requer que os dados originais na base de dados sejam utilizados para treinamento e teste; neste trabalho, foram adotados 10 subconjuntos para aprendizagem. A Medida F é a razão média entre precisão e *recall*, e mede a capacidade de reconhecer os exemplos negativos e positivos (Witten & Frank, 2011, p. 175). Por definição, uma curva ROC é um gráfico bidimensional em que o eixo horizontal representa a taxa de erro da classe negativa (1-*Spec*) e, no eixo vertical, os valores de sensitividade. O desempenho de um classificador é medido pela área sob a curva ROC (Han, Kamber, & Pei, 2011, p. 372).

4 RESULTADOS

Nesta seção, são apresentados os resultados da primeira etapa da modelagem, a saber, a seleção das variáveis 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. Além disso, expõe-se o principal resultado da segunda etapa, ou seja, o poder preditivo do modelo. Para ajudar a visualizar e a comparar a evolução de alguns indicadores selecionados, são também apresentados alguns resultados de natureza descritiva, para cada grupo de empresas (solventes ou insolventes). São também comparados resultados de alguns estudos feitos com dados contábeis de empresas brasileiras.

4.1 RESULTADO DA SELEÇÃO E ANÁLISE DESCRITIVA

Doze foram as variáveis preditoras selecionadas pelo método de seleção de atributos com base no algoritmo CFS (seleção de atributos baseado em correlação) e avaliadas pela abordagem Filtro, conforme apresentadas na Tabela 1. Vale ressaltar que os valores da coluna Ano referem-se ao período de tempo do valor daquela variável. Por exemplo, Ano 3 refere-se ao período de tempo no qual a entidade foi declarada insolvente, Ano 2 um período de tempo antes do ano da declaração da insolvência da entidade.

Quanto à aplicação do método *Wrapper* para seleção de atributos, foram oito as variáveis preditoras selecionadas, tal como apresentadas na Tabela 2.

Tabela 1

VARIÁVEIS SELECIONADAS PELA ABORDAGEM FILTRO

NOME DAS VARIÁVEIS	ABREVIATURAS	ANO
1. Rentabilidade Operacional sobre Ativo Total	ROAT	3
2. Recursos Provenientes de Operações Sociais sobre o Patrimônio Líquido	RPOS	3
3. Resultado sobre Patrimônio Líquido	RPL	3
4. Margem Líquida	ML	3
5. Margem Operacional	MO	3
6. Margem Operacional Pós-Receita Financeira	MORF	3
7. Endividamento Total sobre Patrimônio Líquido	ETPL	3
8. Liquidez Seca	LS	3
9. Saldo de Tesouraria sobre o Ativo Total	STAT	2

(continua)

Tabela 1 (Continuação)

VARIÁVEIS SELECIONADAS PELA ABORDAGEM FILTRO

NOME DAS VARIÁVEIS	ABREVIATURAS	ANO
10. Recursos Provenientes de Operações Sociais sobre o Patrimônio Líquido	RPOS	2
11. Giro do Ativo	GA	1
12. Recursos Provenientes de Operações Sociais sobre o Patrimônio Líquido	RPOS	1

Fonte: Elaborada pelos autores.

Tabela 2

VARIÁVEIS SELECIONADAS PELA ABORDAGEM *WRAPPER*

NOME DAS VARIÁVEIS	ABREVIATURAS	ANO
1. Liquidez Geral	LG	1
2. Recursos Provenientes de Operações Sociais sobre o Patrimônio Líquido	RPOS	1
3. Endividamento Oneroso sobre o Ativo Total	EOAT	2
4. Saldo de Tesouraria sobre Ativo Total	STAT	2
5. Rentabilidade Líquida sobre Ativo Total	RLAT	2
6. Margem Operacional Pós-Receita Financeira	MORF	3
7. Margem Líquida	ML	3
8. Rentabilidade Operacional sobre Ativo Total	ROAT	3

Fonte: Elaborada pelos autores.

Cinco variáveis ($RPOS_1$, $Stat_2$, $Morf_3$, ML_3 e $Roat_3$) foram selecionadas em ambas as abordagens estudadas. Aceitando esta interseção como indício de robustez, essas variáveis podem, então, ser consideradas como as mais capazes para caracterizar futuras prováveis empresas insolventes a partir dos dados disponíveis.

Na Tabela 3 mostra-se a evolução de indicadores selecionados nos três últimos períodos anteriores à insolvência, para o caso do grupo das empresas insolventes.

Para o indicador referente aos Recursos Provenientes de Operações Sociais sobre o Patrimônio Líquido (RPOS), verifica-se que as médias vão-se deteriorando ao longo dos anos. Já os indicadores referentes aos desvios-padrões em cada período ganham fôlego, evidenciando o crescente descontrole dessas entidades.

A relação RPOS/PL é um indicador que traduz a influência direta de eventos contábeis e monetários nos demonstrativos da entidade e que não estão diretamente relacionados com a atividade do negócio dela, mas que também influenciam na determinação da capacidade de solvência da empresa atuando, em algumas

ocasiões, como favoráveis e, em outras, como desfavoráveis. Para esta amostra estudada, esta variável apareceu como bem influente na insolvência das empresas.

Tabela 3

EVOLUÇÃO DOS INDICADORES NO GRUPO DAS INSOLVENTES

VARIÁVEL	ANO 1			ANO 2			ANO 3		
	MÉDIA	DP	CV(%)	MÉDIA	DP	CV(%)	MÉDIA	DP	CV(%)
RPOS	-0,89	7,65	-8,56	-0,98	10,42	-10,69	-1,79	14,89	-8,33
STAT	-0,26	11,23	-43,87	-0,40	19,54	-49,10	-0,68	27,65	-40,96
MORF	-0,31	6,78	-21,87	-0,42	37,89	-89,60	-0,57	45,78	-80,74
ML	-0,39	8,96	-23,10	-0,59	49,89	-84,70	-0,68	56,78	-84,12
ROAT	-0,17	17,78	-105,96	-0,27	19,45	-72,63	-0,30	17,81	-59,59

Média = Média aritmética dos índices; DP = Desvio-padrão; CV = Coeficiente de variação.

Fonte: Elaborada pelos autores.

O Saldo de Tesouraria sobre o Ativo Total (Stat) é apurado pela diferença entre o capital circulante líquido (CCL) e a necessidade de capital de giro (NCG). Esse saldo funciona como uma reserva financeira da empresa para fazer frente às eventuais expansões da necessidade de investimento operacional em giro, principalmente aquelas de natureza sazonal. Pela Tabela 3, pode ser verificado que tal variável vem cada vez mais se deteriorando comprometendo a capacidade de pagamento dessas entidades no transcorrer dos anos. Pelos índices dos desvios-padrões fica também evidente um crescente e permanente descontrole sobre tal variável.

No caso da variável Margem Operacional Pós-Receita Financeira (Morf), os índices apresentam comportamento comprometedor para essas entidades; tanto em termos das médias, quanto dos desvios-padrões, há um movimento de deterioração do índice comprometendo a continuidade dessas entidades.

A evolução da variável Margem Líquida é coerente com a situação de empresas em fase pré-falimentar, uma vez que, considerando os sinais, suas médias decrescem ligeiramente ao longo dos períodos precedentes à insolvência. Uma empresa nessa situação apresenta excessiva dificuldade para gerenciar sua margem, talvez devido à enxurrada de múltiplos compromissos, levando ao sacrifício das atividades operacionais na tentativa de atender às obrigações de curtíssimo prazo.

Para a variável indicativa de Rentabilidade Operacional sobre Ativo Total (Roat), o percurso das médias manteve-se coerente com a situação de insolvência,

mostrando diminuição dos valores no decorrer do período. O que se observa na Tabela 3 é a perda gradativa de rentabilidade das empresas insolventes aqui amostradas. Há maior deterioração durante a transição do ano 1 para o ano 2, sugerindo aumento de ineficiência operacional e, consequentemente, financeira, conduzindo a empresa ao estado de insolvência mais agudo.

Pode-se supor que, nas empresas insolventes aqui selecionadas, ocorreram acentuados comprometimentos nas rentabilidades, talvez negligenciando o controle da eficiência da empresa com seus ativos e gerência de suas operações. Tais empresas, tentando retomar uma melhor condição financeira, acabaram sendo levadas a um comprometimento demasiado em sua capacidade operacional, provavelmente realizando desmobilizações e descontroles gerenciais. Os índices calculados de rentabilidades e de margens estimulam essas inferências. Resumindo, com as rentabilidades e margens comprometidas, talvez a busca de liquidez (mesmo que apenas) seja priorizada, passando a ser o desempenho operacional e o controle de seus ativos estratégias secundárias, resultando na aceleração do processo de insolvência.

Na Tabela 4, em que constam os resultados descritivos para as empresas solventes, não ocorrem variações tão relevantes. O que chama a atenção são alguns valores encontrados para o coeficiente de variação. A comparação desses valores com os das empresas insolventes mostra que são muito superiores àqueles, indicando que a dispersão entre os índices das empresas solventes é muito maior do que entre os das empresas insolventes. Pode-se especular, então, se haveria alguma convergência no perfil das insolventes nos anos próximos à insolvência, ao menos para algumas das características tabuladas.

Tabela 4

EVOLUÇÃO DOS INDICADORES NO GRUPO DAS EMPRESAS SOLVENTES

VARIÁVEL	ANO 1			ANO 2			ANO 3		
	MÉDIA	DP	CV(%)	MÉDIA	DP	CV(%)	MÉDIA	DP	CV(%)
RPOS	0,12	3,65	30,42	0,16	4,76	29,75	0,17	3,98	23,41
STAT	0,06	6,71	111,83	0,05	5,67	113,40	0,07	6,89	98,43
MORF	0,02	3,45	172,50	0,03	3,89	129,67	0,03	3,78	126,00
ML	0,01	2,45	245,00	0,02	2,03	101,50	0,02	2,32	116,00
ROAT	0,02	4,78	239,00	0,02	3,98	199,00	0,01	4,09	409,00

Fonte: Elaborada pelos autores.

4.2 MODELOS DE PREVISÃO

São apresentados nesta seção, primeiro, os resultados obtidos com a aplicação de duas técnicas de classificação – Redes Neurais e Regressão Logística –, com o propósito de desenvolver modelos de previsão de insolvência compostos pelas variáveis selecionadas e pré-avaliados nas abordagens Filtro e *Wrapper*. Em segundo lugar, exibem-se os efeitos dessas abordagens sobre os classificadores estudados.

4.2.1 Resultados da classificação para a abordagem Filtro

O modelo elaborado com o classificador da Regressão Logística é composto pelos subconjuntos de variáveis selecionados e avaliados pela abordagem Filtro. O modelo de pontuação X, especificado a seguir, é um modelo classificatório para tomadores corporativos, combinados em uma função de classificação conforme a equação a seguir:

$$X = -0{,}8751 + 0{,}0054\ RPOS_1 + 0{,}1396\ GA_1 + 0{,}0345\ RPOS_2 + 0{,}1126\ STAT_2 - 0{,}8228\ LS_3 - 0{,}0059\ ETPL_3 - 5{,}2198\ MORF_3 + 0{,}0733\ MO_3 + 0{,}0603\ ML_3 - 0{,}0445\ RPL_3 - 0{,}022\ RPOS_3 - 3{,}282\ ROAT_3.$$

Na Tabela 5 pode ser constatado que, para as empresas insolventes, o índice de acertos atingiu 96,72% (59 em 61), enquanto para as empresas solventes os acertos alcançaram 89,34% (109 em 122). A classificação correta total do grupo de origem foi de 91,8%.

Tabela 5

RESULTADOS DO CLASSIFICADOR REGRESSÃO LOGÍSTICA PARA SUBCONJUNTOS DE ATRIBUTOS SELECIONADOS PELA ABORDAGEM FILTRO

GRUPO DE ORIGEM	GRUPO DE CLASSIFICAÇÃO		
	INSOLVENTES	SOLVENTES	TOTAL
Insolventes	59	2	61
Solventes	13	109	122
Medida F	0,88	0,93	
Área ROC	0,92	0,92	
Classificação correta do grupo de origem			**91,80%**

Fonte: Elaborada pelos autores.

Quando se considera o modelo que utiliza como classificador as Redes Neurais, aplicadas a subconjuntos de variáveis selecionadas pela abordagem Filtro, a equação que combina as variáveis é a seguinte:

$$X = -3{,}4261 - 1.0525\ RPOS_1 - 4{,}2915\ GA_1 - 4{,}6332\ RPOS_2 - 1{,}1663\ STAT_2 + 5{,}4912\ LS_3 - 6{,}3522\ ETPL_3 + 2{,}6939\ MORF_3 - 2{,}595\ MO_3 - 4{,}6848\ ML_3 + 17{,}1965\ RPL_3 + 20{,}9121\ RPOS_3 + 10{,}6644\ ROAT_3.$$

Na Tabela 6 pode ser constatado que, nas empresas insolventes, houve um índice de acertos de 95% (58 em 61). Já nas empresas solventes o índice de acertos foi de 93,8 % (105 em 112). A classificação correta do grupo de origem foi de 92,4%.

Tabela 6

RESULTADOS DO CLASSIFICADOR REDES NEURAIS PARA SUBCONJUNTOS DE ATRIBUTOS SELECIONADOS PELA ABORDAGEM FILTRO

GRUPO DE ORIGEM	GRUPO DE CLASSIFICAÇÃO		
	INSOLVENTES	SOLVENTES	TOTAL
Insolventes	58	3	61
Solventes	8	114	122
Medida F	0,913	0,954	
Área ROC	0,939	0,939	
Classificação correta			**93,98%**

Fonte: Elaborada pelos autores.

4.2.2 Resultados para a abordagem *Wrapper*

O modelo elaborado com o classificador da Regressão Logística é composto pelos subconjuntos de variáveis selecionados e avaliados pela abordagem *Wrapper*, combinados na seguinte função de classificação:

$$X = -2{,}5401 + 0{,}1748\ LG_1 + 0{,}0095\ RPOS_1 + 2{,}6245\ EOAT_2 - 2{,}6387\ RLAT_2 + 0{,}2369\ STAT_2 - 9{,}0368\ MORF_3 + 0{,}6548\ ML_3 - 4{,}2423\ ROAT_3.$$

Na Tabela 7 constata-se que, nas empresas insolventes, houve um índice de acertos de 65,57% (40 em 61). Já para as empresas solventes os acertos ficaram próximos de 90,16% (110 em 122), enquanto a classificação correta total do grupo de origem foi de 81,96%.

Tabela 7

RESULTADOS DO CLASSIFICADOR REGRESSÃO LOGÍSTICA PARA SUBCONJUNTOS DE ATRIBUTOS SELECIONADOS PELA ABORDAGEM *WRAPPER*

GRUPO DE ORIGEM	GRUPO DE CLASSIFICAÇÃO		
	INSOLVENTES	SOLVENTES	TOTAL
Insolventes	40	21	61
Solventes	12	110	122
Medida F	0,708	0,87	
Área ROC	0,892	0,892	
Classificação correta			**81,96%**

Fonte: Elaborada pelos autores.

Para a classificação com Redes Neurais, a função de classificação foi estimada como:

$$X = -1{,}328 + 1.1466\ LG_1 - 2{,}7321\ RPOS_1 - 1{,}1776\ EOAT_2 + 0.4137\ RLAT_2 + 1{,}271\ STAT_2 + 2{,}8047\ MORF3 - 0{,}9377\ ML_3 + 2{,}8188\ ROAT_3.$$

A Tabela 8 mostra que, para as empresas insolventes, o índice de acertos atingiu 45,90%, ao passo que para as empresas solventes o índice de acertos foi de 93,44%. Em termos totais, a classificação correta do grupo de origem foi de 77,59%.

Tabela 8

RESULTADOS DO CLASSIFICADOR REDES NEURAIS PARA SUBCONJUNTOS DE ATRIBUTOS SELECIONADOS PELA ABORDAGEM *WRAPPER*

GRUPO DE ORIGEM	GRUPO DE CLASSIFICAÇÃO		
	INSOLVENTES	SOLVENTES	TOTAL
Insolventes	28	33	61
Solventes	8	114	122
Medida F	0,577	0,848	
Área ROC	0,808	0,808	
Classificação correta			**77,59%**

Fonte: Elaborada pelos autores.

Com auxílio da Tabela 9, podemos comparar as resultados gerados pelos quatro modelos, em que os subconjuntos de atributos foram selecionados e avaliados pelas abordagens Filtro e *Wrapper* e as empresas foram classificadas como "solventes" ou "insolventes" por meio dos dois classificadores – Regressão Logística e Redes Neurais. Os modelos apoiados em subconjuntos de variáveis avaliados pela abordagem Filtro apresentaram melhores índices de classificação, independentemente do classificador, ou seja, tanto para regressão logística, quanto para redes neurais. Cabe ressaltar que a melhor *performance* – significando melhores medidas F e área ROC – foi obtida com o classificador redes neurais (93,98%).

Tabela 9

RESUMO COMPARATIVO DOS RESULTADOS

AVALIAÇÃO DE SUBCONJUNTOS DE ATRIBUTOS	CLASSIFICADOR	Nº DE ACERTOS	Nº DE ERROS	ACERTOS %
Filtro	Regressão Logística	168	15	91,80
Filtro	Redes Neurais	172	11	93,98
Wrapper	Regressão Logística	150	33	81,96
Wrapper	Redes Neurais	142	41	77,59

Fonte: Elaborada pelos autores.

Das variáveis escolhidas pelas duas abordagens, cinco foram selecionadas simultaneamente por ambos os métodos estudados – RPOS, Stat, Morf, Roat, ML –, o que permite sugerir que estas variáveis apresentam a melhor capacidade para caracterizar as classes de empresas solventes e insolventes, dada a informação disponível na amostra utilizada. Neste sentido, a despeito dessa evidência limitada presente nos dados amostrados, ganham destaque para compor modelos de previsão de insolvência.

Uma questão importante na forma final dos 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 margem líquida. Este resultado assume importância ainda maior ao considerar as modificações ocorridas no mercado, tornando-o mais competitivo e fortalecendo empresas que assumiram rentabilidades menores para manter sua sobrevivência.

4.2.3 Comparativo de alguns resultados de estudos sobre previsão de insolvência utilizando dados contábeis

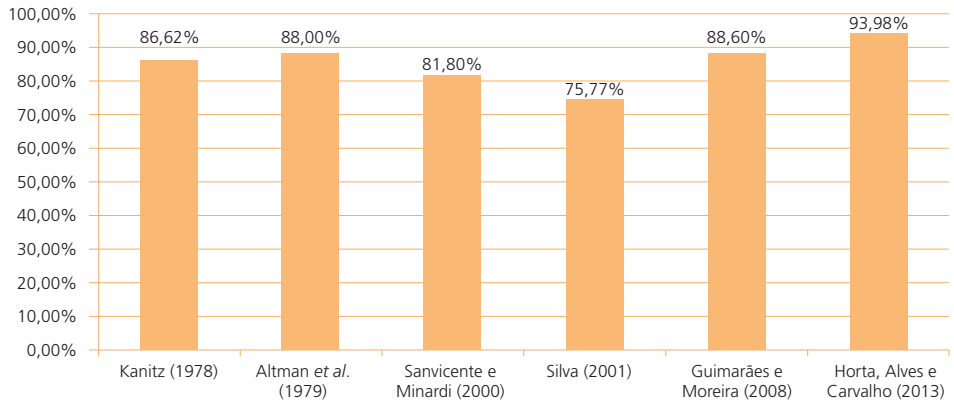
Neste item, foram feitos comparativos de alguns estudos de previsão de insolvência com dados contábeis de empresas brasileiras que ganharam relevância com o estudo do tema aqui estudado (Kanitz, 1978; Altman, Baidya, & Dias, 1979; Sanvicente & Minardi, 2000; Silva, 2001; Guimarães & Moreira, 2008), sendo apresentado um gráfico com os resultados obtidos.

Foi o trabalho desenvolvido por Kanitz em 1978 que gerou notoriedade no Brasil sobre o estudo de previsão de insolvência. Ele utilizou análise discriminante em cinco variáveis contábeis de 74 empresas brasileiras, sendo 49 solventes e 25 insolventes. O grau de acerto foi de 86,62%.

O modelo de Altman, Baidya e Dias em 1979 foi baseado no modelo Zeta™ desenvolvido nos Estados Unidos em 1968. No Brasil, foram utilizados dados contábeis de 58 empresas, sendo 23 consideradas insolventes e 35 solventes. O modelo obteve 88% de acertos.

Gráfico 1

COMPARAÇÃO DE RESULTADOS DE ESTUDOS SOBRE PREVISÃO DE INSOLVÊNCIA UTILIZANDO DADOS CONTÁBEIS DE EMPRESAS BRASILEIRAS



Fonte: Elaborado pelos autores.

O modelo de Sanvicente e Minardi (2000) foi baseado nos modelos de Kanitz (1978) e Altman *et al.* (1979). O índice de acerto foi de 81,8%.

O modelo de Silva (2001) utilizou análise discriminante em dados contábeis de 20 empresas solventes e 33 insolventes. O percentual de acerto do modelo foi de 75,77%.

O modelo de Guimarães e Moreira obteve um índice de acerto de 88,6%, tendo como amostra 116 empresas de capital aberto. Essa amostra foi dividida em duas subamostras: uma de desenvolvimento, sendo 35 solventes e 35 insolventes, e outra amostra de validação, contendo 23 empresas solventes e igual número de empresas insolventes.

Já o presente estudo obteve como melhor resultado um índice de 93,98%, utilizando como classificador redes neurais e filtro como avaliador dos subconjuntos de atributos. Compuseram a amostra 183 empresas, 61 insolventes e 122 solventes.

5 CONSIDERAÇÕES FINAIS

Com os dados coletados sobre as empresas amostradas nesta pesquisa, os resultados obtidos na classificação de empresas segundo sua "insolvência" (Tabela 9) foram melhores quando se aplicou a abordagem Filtro para seleção de subconjuntos de atributos, independentemente do classificador utilizado, a saber, Redes Neurais (93,98%) ou Regressão Logística (91,80%). Em outras palavras, as variáveis selecionadas foram capazes de realizar com benefícios a tarefa de classificação. Por exemplo, a redução da dimensionalidade devida à seleção de variáveis conduziu não somente à economia de custo computacional (e de aquisição de dados), mas também ao ganho de desempenho e de exatidão nas etapas de MD propriamente ditas.

Em termos de capacidade de predição, a abordagem Filtro selecionou um subconjunto de atributos mais eficiente do que a abordagem *Wrapper,* permitindo que, nos modelos de previsão, ambos os classificadores obtivessem desempenhos bem melhores.

Não há evidência suficiente para afirmar que a abordagem Filtro deva ser geralmente a preferida. A melhor recomendação a ser dada é que a redução da dimensionalidade deve ocorrer como etapa (preliminar) da tarefa de classificação. No caso da seleção de subconjuntos com dados contábeis brasileiros, tal como foi aqui realizada, foi mostrado que a abordagem Filtro foi robustamente mais eficaz.

Além disso, no caso da previsão de insolvência usando dados contábeis de empresas brasileiras, desenvolvida neste estudo, foi também demonstrada 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*. Assim, no que se refere à utilidade da abordagem Filtro, destaca-se a conclusão de que essa

abordagem pode ser usada com vantagem na avaliação dos subconjuntos de atributos que irão compor as funções de classificação, que são a efetiva base dos modelos preditivos.

Já em relação à precisão da classificação, é importante evidenciar que os índices encontrados pelos avaliadores de desempenho neste estudo (medida F e área ROC) foram bem significativos, acima de 0,9 quando aplicadas redes neurais, além disso foi utilizada validação cruzada (com 10 partições), evidenciando assim a precisão dos resultados quando da utilização deste classificador para a base de dados estudada.

Na comparação dos resultados com estudos já realizados sobre o tema, foi detectado que os achados foram consistentes e bem adequados com as magnitudes esperadas, apresentados no Gráfico 1 e na Tabela 9, evidenciando a importância da seleção de atributos para melhorar os resultados, além da aplicação de técnicas de *data mining* em dados contábeis de empresas brasileiras.

A utilização de modelos preditivos de insolvência, construídos com apoio em *data mining*, é uma entre várias formas de avaliar o risco de insolvência para uma instituição, sem depender apenas da avaliação subjetiva do analista. Esses modelos preditivos podem ser incorporados como procedimento analítico para avaliar a probabilidade de insolvência. São interessantes para que bancos, investidores, governos, auditores, gerentes, fornecedores, empregados e muitos outros agentes econômicos possam avaliar, com razoável antecedência, se há "problemas de insolvência" em andamento.

Apesar de a qualidade dos dados contábeis ser, ainda, muitas vezes questionada em relação a utilização na construção de modelos de previsão de falência, os resultados aqui obtidos quanto ao acerto na classificação foram bastante satisfatórios, da ordem de 80% ou mais de acerto, o que vem evidenciar o conteúdo de informação que os dados contábeis brasileiros proporcionam, pelo menos para realizar previsões. Em outras palavras, é adequado esperar que, com o emprego de dados contábeis brasileiros, se possa prever a saúde financeira de uma empresa operando no Brasil. Como exemplo específico, podem-se ponderar os diversos indicadores contábeis aqui apresentados para obter um modelo que forneça um índice de risco de crédito. A ponderação pode ser obtida pela aplicação de técnicas de *data mining*. Isso permite esperar que, apoiada em indicadores contábeis, *data mining* será ferramenta útil tanto para prever concordatas de empresas, quanto para estabelecer escores associados a risco de crédito.

ATTRIBUTE SELECTION IN BANKRUPTCY PREDICTION: APPLICATION AND EVALUATION USING RECENT BRAZILIAN DATA

ABSTRACT

Bankruptcy prediction may have great utility to financial and nonfinancial institutions with regard to take in advance the best possible decisions regarding loans or investments. In specific literature, many bankruptcy prediction models have made use of *data mining*. The preprocessing step is important to select good quality data for use in mining operations. Still, although the selection of attributes can be very beneficial to pre-select representative data to improve the forecast performance end, it is not known which method is the best selection. This work has as main objective to compare two approaches for evaluating subsets of attributes: Filter and Wrapper. Despite being based on data mining techniques and widely used in the step of feature selection in bankruptcy prediction models, these techniques are rarely used to treat data from financial statements of Brazilian companies. Therefore the empirical basis of this study consists of a sample of Brazilian industrial and commercial enterprises, collecting data for the period 2004 to 2011. The results indicated that, in this sample, the filter approach was more efficient, providing better classification results both for logistic regression (91,80%) and for neural networks (93,98%). It was shown also the importance of making explicit the evaluation stage of the selection of attributes for achieving better results in applications of data mining techniques to predict insolvency. A specific conclusion about the advantages of the filter approach shows that it may be preferred to assess the attributes that will make predictive models.

KEYWORDS

Financial indicators. Forecast insolvency. Data mining. Attribute selection. Filter and wrapper.

SELECCIÓN DE ATRIBUTOS EN PREVISIÓN DE INSOLVENCIA: APLICACIÓN Y EVALUACIÓN UTILIZANDO DATOS RECIENTES DE BRASIL

RESUMEN

Predicción de bancarrota puede tener gran utilidad para las instituciones financieras y no financieras con respecto a tomar de antemano las mejores decisiones posibles con respecto a los préstamos o inversiones. En la literatura específica, muchos de los modelos de predicción de bancarrota han hecho uso de minería de datos (*data mining*). En la etapa de pre-procesamiento es importante seleccionar datos de buena calidad para su uso en las operaciones mineras. Sin embargo, aunque la selección de atributos puede ser muy beneficioso para los datos representativos pre-selección para mejorar el rendimiento final previsto, no se sabe qué método es la mejor selección. Este trabajo tiene como principal objetivo comparar dos enfoques para la evaluación de los subconjuntos de atributos: con filtro y envoltura. A pesar de que se basan en técnicas de minería de datos y ampliamente utilizados en la selección de características en los modelos de predicción de la etapa de la insolvencia, estas técnicas son raramente utilizados para tratar datos de los estados financieros de las empresas brasileñas. Así que la base empírica de este estudio consiste en una muestra de empresas comerciales e industriales de Brasil, mediante la recopilación de datos para el período 2004-2011. Los resultados indicaron que, en este ejemplo, el enfoque de filtro fue más eficiente, proporcionando mejores resultados de la clasificación tanto para la regresión logística (91,80%) y de redes neuronales (93,98%). También se demostró la importancia de hacer explícita la etapa de evaluación de la selección de atributos para lograr mejores resultados en la aplicación de técnicas de minería de datos para predecir la insolvencia. Una conclusión específica acerca de las ventajas del enfoque de filtro de muestra que puede ser preferible para evaluar los atributos que harán modelos predictivos.

PALABRAS CLAVE

Indicadores financieros. Insolvencia. Minería de datos. Selección de atributos. Filtro y envoltura enfoques.

REFERÊNCIAS

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance, 23*, 589-609.

Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance, 18*(3), 505-529.

Altman, E. I., Baidya, T. K. N., & Dias, L. M. R. (1979). Previsão de problemas financeiros em empresas. *Revista de Administração de Empresas, 19*, 17-28.

Back, B., Laitinen, T., & Kaisa, S. (1996). Neural networks and genetic algorithms for bankruptcy prediction. *Expert Systems with Applications, 11*(4), 407-413.

Balcaen, S., & Ooghe, H. (2006). Five years of studies on business failure: on overview of the classical statistical methodologies and their related problems. *The British Accounting Review, 38*(1), 63-93.

Beaver, W. (1968). Alternative accounting measures as predictors of failure. *The Accounting Review, 43*(1), 112-122.

Blum, M. (1974). Failing company discriminant analysis. *Journal of Accounting Research, 12*(1), 1-25.

Braga, A. P., Carvalho, A. C. P. L. F., & Ludemir, T. B. (2007). *Redes neurais artificiais: teoria e aplicações* (2a ed.). Rio de Janeiro: Livros Técnicos e Científicos.

Bragança, L. A., & Bragança, S. L. (1984). Previsão de concordatas e falências no Brasil. *Anais Do Congresso Abamec*, Rio de Janeiro, RJ, Brasil.

Brockett, P. L., Cooper, W. W., Golden, L. L., & Xia, X. (1997). A case study in applying neural networks to predicting insolvency for property and casualty insurers. *Journal of the Operational Research Society, 48*(12), 1153-1162.

Chye, K. H., Chin, T. W., & Peng, G. C. (2004). Credit scoring using data mining techniques. *Singapore Management Review*, (1), 25-47.

Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research, 10*(1), 167-179.

Edmister, R. O. (1972). An empirical test of financial ratio analysis for small business failure prediction. *Journal of Financial and Quantitative Analysis, 7*(2), 1477-1493.

Eisenbeis, R. A. (1997). Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal the Finance, 32*(3), 878-900.

Fayyad, U., Piatetsky, S. G., & Smith, P. (1996). From data mining to knowledge discovery: an overview. *Advances in Knowledge Discovery & Data Mining, 7*(3), 37-54.

Fitzpatrick, P. J. (1932). A comparison of ratios of successful industrial enterprises with those of failed companies. *Certified Public Accountant*, 598-605.

Freitas, A. A. (1998). *Data mining and knowlwdge discovery with evolutionary algorithms*. New York: Springer-Verlag.

Guimarães, A., & Moreira, T. B. S. (2008). Previsão de insolvência: um modelo baseado em índices contábeis com utilização de análise discriminante. *Revista de Economia Contemporânea, 12*(1), 151-178.

Hall, M. A. (1999). *Correlation-based feature subset selection for machine learning*. Tese de doutorado, University of Waikato, Nova Zelândia.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, *15*(6), 1437-1448.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3a ed.). Waltham: Morgan Kaufmann.

Härdle, W., Moro, R. A., & Schäfer, D. Predicting bankruptcy with support vector machines. SFB 649. *Discussion Paper* 2005-009. Recuperado em 10 março, 2006, de http://sfb649.wiwi.huberlin.de.

Hassan, O., & Marston, C. L. Disclosure measurement in the empirical accounting literature – a review article. 15 July 2010. Recuperado em 12 julho, 2010, de http://ssrn.com/abstract=1640598.

Haykin, S. (2001). *Redes neurais: princípios e práticas* (2a ed.). Porto Alegre: Bookman.

Horta, R. A. (2001). *Utilização de indicadores contábeis na previsão de insolvência: análise empírica de uma amostra de empresas comerciais e industriais brasileiras*. Dissertação de mestrado, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, RJ, Brasil.

Iudícibus, S. de. (1998). *Análise de balanços* (6a ed.). São Paulo: Atlas.

Jonh, G. H., Kohavi, R., & Pfeger, K. (1994). Irrelevant features and the subset selection problem. *Proceedings of the 11th International Conference on Machine Learning*, Boca Raton, FL, USA.

Kanitz, S. C. (1978). *Como prever falências*. São Paulo: McGraw-Hill do Brasil.

Kasznar, I. K. (1986). *Falências e concordatas de empresas: modelos teóricos e estudos empíricos*. Dissertação de mestrado, Fundação Getulio Vargas, Rio de Janeiro, RJ, Brasil.

Kohavi, R., & George, J. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273-324.

Kreuze, D. (2001). Debugging hospitals. *Technology Review*, *32*.

Lennox, C. (1999). Identifying failing companies: a reevaluation of the logit, probit and MDA approa ches. *Journal of Economics and Business*, *5*(4), 347-364.

Liu, H., & Motoda, H. (1998). *Feature selection for knowledge discovery and data mining*. Massachusetts: Kluwer Academic Publishers, 101Philip Driver, Assinippi Park, Norwell.

Mario, P. C. (2005). *O fenômeno da falência: análise das causas*. Tese de doutorado, Universidade de São Paulo, São Paulo, SP, Brasil.

Matarazzo, D. C. (2003). *Análise financeira de balanços* (6a ed.). São Paulo: Atlas.

Pereira, J. S. (2006). *Gestão e análise de risco de crédito* (5a ed.). São Paulo: Atlas.

Piramuthu, S. (2006). On preprocessing data for financial credit risk evaluation. *Expert Systems with Application*, *30*(3), 489-497.

Rezende, S. (Org.). (2005). *Sistemas inteligentes: fundamentos e aplicações*. Barueri: Manole.

Sanvicente, A. Z., & Minardi, A. M. A. F. (2000). Identificação de indicadores contábeis significativos para previsão de concordata de empresas. Recuperado em 5 fevereiro, 2005, de http://www.risktech.br/artigos/artigostécnicos /index.html.

Schrickel, W. K. (1999). *Demonstrações financeiras* (2a ed.). São Paulo: Atlas.

Shirata, C. Y. (2001). *Financial ratios as predictors of bankruptcy in Japan: an empirical research*. Recuperado em 9 agosto, 2006, de http://www.shirata.net/apira98.html.

Tsai, C.-F., & Cheng, K.-C. (2012). Simple instance selection for bankruptcy prediction. *Knowledge--Based Systems*, *27*, 333-342.

Verikas, A., Kalsyte, Z., Bacauskiene, M., & Gelzinis, A. (2010). Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. *Soft Computing – A Fusion of Foundations, Methodologies and Applications*, *14*(9), 995-1010.

Winakor, A., & Smith, R. (1935). Changes in the financial structure of unsuccessful industrial corporations. *Bulletin, Bureau of Business Research, 32*(51).
Witten, I. H., & Frank, E. (2011). *Data mining: practical machine learning tools and techniques* (3a ed.). Burlington, VT: Morgan Kaufmann.
Wu, C.-H., Fang, W.-C., & Goo, Y.-J. (1996). Variable selection method affects SVM-based models in bankruptcy prediction. *Advances in Intelligent Systems Research*, 1-34.