



Revista Base (Administração e
Contabilidade) da UNISINOS

E-ISSN: 1984-8196

cd@unisinos.br

Universidade do Vale do Rio dos Sinos
Brasil

MATHIASI HORTA, RUI AMÉRICO; DOS SANTOS ALVES, FRANCISCO JOSÉ; DE CARVALHO,
FREDERICO A.; JORGE, MARCELINO JOSÉ

SELEÇÃO DE ATRIBUTOS CONTÁBEIS NA PREVISÃO DE INSOLVÊNCIA DE EMPRESAS
BRASILEIRAS – UMA COMPARAÇÃO DE ABORDAGENS

Revista Base (Administração e Contabilidade) da UNISINOS, vol. 12, núm. 1, enero-marzo, 2015, pp.
27-39

Universidade do Vale do Rio dos Sinos
São Leopoldo, Brasil

Disponível em: <http://www.redalyc.org/articulo.oa?id=337238452004>

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

redalyc.org

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

SELEÇÃO DE ATRIBUTOS CONTÁBEIS NA PREVISÃO DE INSOLVÊNCIA DE EMPRESAS BRASILEIRAS – UMA COMPARAÇÃO DE ABORDAGENS

SELECTION OF ACCOUNTING FEATURES IN THE PREDICTION OF BANKRUPTCY IN BRAZILIAN COMPANIES – A COMPARISON OF APPROACHES

RUI AMÉRICO MATHIASI HORTA

rui.horta@ufjf.edu.br

FRANCISCO JOSÉ DOS SANTOS ALVES

fjalves@globo.com

FREDERICO A. DE CARVALHO

fdecarv@gmail.com

MARCELINO JOSÉ JORGE

marcelino.jorge@ipecc.fiocruz.br

RESUMO

Previsão de insolvência tem sido um tema de estudo que tem ganho muita atenção em análise de negócios devido à importância de informações precisas e oportunas nas decisões estratégicas de negócios. Isto porque a incorreta tomada de decisão nas instituições pode gerar dificuldades financeiras, além de causar grandes custos sociais que afetam os proprietários ou acionistas, gestores, trabalhadores, credores, fornecedores, clientes, comunidade, governo, etc. Como resultado, a previsão de falência tem sido uma das tarefas mais desafiadoras e um tópico de pesquisa importante na contabilidade, finanças, computação, e as técnicas de mineração de dados têm sido aplicadas para resolver problemas de previsão de falências. Seleção de atributos é uma etapa importante para selecionar dados mais representativos de um conjunto de índices contábeis obtidos a partir de demonstrativos financeiros de empresas brasileiras; esta etapa visa melhorar o desempenho da previsão final. O objetivo principal deste artigo é comparar três abordagens de seleção de atributos, filtro, *wrapper* e análise de componentes principais, em dados selecionados para elaboração de modelos de previsão de insolvência. Esta pesquisa é de natureza empírica, descritiva e quantitativa, compreendendo as empresas classificadas no SERASA e na BOVESPA como insolventes no período de 2005 a 2007. Neste trabalho, demonstrou-se, para a amostra utilizada, que a abordagem *wrapper* é a mais eficiente; ela obteve os melhores resultados de classificação nas técnicas de regressão logística (89,88%), árvore de decisão (93,45%) e máquina de vetor suporte (97,02%).

Palavras-chave: seleção de atributos, previsão de insolvência, índices contábeis, mineração de dados.

ABSTRACT

Insolvency prediction has been a topic of study that has gained much attention in business analysis because of the importance of accurate and timely information on strategic business decisions. This is because the incorrect decision-making in institutions can generate financial difficulties besides causing huge social costs that affect the owners or shareholders, managers, employees, creditors, suppliers, customers, community, government, etc. As a result, bankruptcy prediction has been one of the most challenging tasks and an important research topic in accounting, finance and computer science and data mining techniques have been applied to

solve problems in bankruptcy prediction. The selection of attributes is important to select the most representative data from a set of accounting ratios derived from financial statements of Brazilian companies; this step aims to improve the performance of the final prediction step. The main objective of this paper is to compare three approaches to feature selection, viz. Filter, wrapper and principal component analysis, in data selected for the development of insolvency prediction models. This research is of an empirical, descriptive and quantitative nature, comprising companies classified at SERASA and BOVESPA as insolvent in the period of 2005–2007. This work demonstrated, for the sample used, that the wrapper approach is the most effective one; it obtained the best classification results in the techniques of logistic regression (89,88%), decision tree (93,45%) and support vector machine (97,02%).

Keywords: selection of attributes, insolvency prediction, accounting ratios, data mining.

INTRODUÇÃO

No Brasil, as comunidades empresariais e acadêmicas têm dado cada vez mais atenção ao tema previsão de insolvência devido ao aumento do acesso ao crédito nos últimos anos no mercado brasileiro. Uma incorreta tomada de decisão por parte das organizações pode gerar dificuldades financeiras, ocasionando custos sociais que venham a afetar os proprietários, acionistas, gestores, o governo, etc.

Como resultado, a previsão de falência tem sido uma das tarefas mais desafiadoras e um dos tópicos de pesquisa mais relevantes na área de contabilidade e finanças. O avanço da tecnologia da informação tem permitido obter uma grande variedade de informações sobre a situação de risco de uma empresa.

No processo de avaliação de uma grande quantidade de informação, muitas pessoas geralmente dependem do julgamento de um analista. No entanto, alguns fatores podem influenciar o resultado da análise. Técnicas de mineração de dados são métodos que têm sido usados para identificar fatores importantes para previsão de falências. Essas técnicas podem ser usadas como "sistemas de alerta precoce". Estes sistemas permitem que sejam tomadas medidas para evitar a falência de empresas. Essas medidas incluem a decisão sobre a fusão, liquidação ou reorganização da empresa.

Ademais, esses sistemas podem ajudar os tomadores de decisão de instituições financeiras quanto à seleção e avaliação de empresas com risco acentuado de descontinuidade. Quando a informação irrelevante é superabundante, é improvável que se venha a interpretar e absorver informações úteis com facilidade. Portanto, para filtrar e condensar grande quantidade de dados é muito importante selecionar os atributos (variáveis), especialmente para casos de previsão de insolvência.

O objetivo deste estudo é comparar técnicas de seleção de atributos aplicados em previsão de insolvência de empresas brasileiras não financeiras de capital aberto. Foram aplicadas três abordagens de seleção de atributos, sendo duas com origem na aprendizagem de máquina – filtro e *wrapper* – e uma de estatística multivariada – análise de componentes principais. Posteriormente é feita uma avaliação dos atributos

selecionados por essas abordagens através de três algoritmos de classificação – regressão logística, árvore de decisão e máquina de vetor suporte.

A relevância deste estudo se justifica a partir da visão de alguns autores (Shirata, 2001; Piramuthu, 2006; Tsai, 2009), os quais têm chamado a atenção para a importância do processo de seleção de atributos. Este é um fato a que, na maior parte dos estudos sobre previsão de insolvência, nem sempre é dada a devida atenção, o que dificulta o entendimento sobre como se chegou às variáveis utilizadas. Este estudo considera explicitamente a fase de seleção das variáveis preditivas como forma de evidenciar a sua importância para a modelagem de previsão de falência.

REFERENCIAL TEÓRICO

Os modelos de previsão de insolvência oferecem aos analistas e aos gestores de crédito uma ferramenta avançada, isenta de influências subjetivas e que lhes possibilita classificar a "saúde financeira" das empresas. Esses modelos consideram, sobretudo, as operações de curto prazo, sabendo que a insolvência está mais relacionada à perda da capacidade de endividamento pela empresa do que com o seu desempenho operacional.

Apesar de sua longa história na literatura especializada (Fitzpatrick, 1932; Winakor e Smith, 1935), o estudo do "insucesso" de empresas com base em indicadores obtidos a partir dos demonstrativos contábeis tomou ímpeto nos anos 70 (Deakin, 1972; Edmister, 1972; Kanitz, 1978, e tantos outros), em seguida aos trabalhos pioneiros de Beaver (1968) 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 in Silva, 2006; Kasznar, 1986; Sanvicente e Minardi, 2000; Horta, 2001), seguindo o caminho aberto por Kanitz (1978).

No decorrer do levantamento do referencial teórico desta pesquisa, constatou-se a prática de utilizar variáveis previamente relacionadas em pesquisas anteriores (Kanitz,

1978; Pereira, 1982 in Silva, 2006; Kasznar, 1986; Sanvicente e Minardi, 2000). Este procedimento pode desconsiderar outras variáveis que melhor representem fatores, como cultura e legislação.

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 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 *et al.*, 1994; Back *et al.*, 1996; Eisenbeis, 1997; Lennox, 1999; Härdle *et al.*, 2005, entre outros).

Apesar de novas técnicas de modelagem se tornarem mais acessíveis, existem vários problemas relatados na literatura com a aplicação desses métodos na previsão de insolvência (Balcaen e Ooghe, 2006). Alguns desses problemas podem ser 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.

DATA MINING

A definição aceita por diversos pesquisadores de *data mining* (DM) foi elaborada por Fayyad, Piatetsky-Shapiro e Smyth (1996) ao afirmarem que "Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados".

O processo de identificação de padrões em DM é dividido em três grandes etapas, conforme Rezende (2005, p. 311): "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, e entre elas está 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 ser 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).

A mais comum e importante aplicação em *data mining* envolve classificação, por vezes referida como modelagem. Classificação refere-se à tarefa de aprender uma função-alvo que mapeie cada conjunto de atributos x para um dos rótulos de classes y predeterminados. A função-alvo também é conhecida informalmente como modelo de classificação. Para a construção de modelos de previsão de insolvência, técnicas de modelagem preditiva são as mais relevantes.

(c) Pós-processamento: a extração de padrões pode gerar uma quantidade enorme de padrões, muitos dos quais podem não ser relevantes para o usuário.

SELEÇÃO DE ATRIBUTOS

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

Os algoritmos usados para a seleção de atributos 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 e Motoda, 1998).

Na atividade de busca de subconjuntos de atributos, seleciona-se um subconjunto de variáveis relevantes com o apoio de um algoritmo de busca. Neste trabalho, foi utilizada a seleção *forward* (adiciona 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.

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 e John, 1997).

ABORDAGEM FILTRO

A abordagem filtro introduz um processo separado de seleção, o qual ocorre antes da aplicação do algoritmo de aprendizagem propriamente dito. A ideia é filtrar os atributos irrelevantes, segundo algum critério, antes do aprendizado ocorrer (Figura 2).

Essa etapa do pré-processamento considera as características gerais do conjunto de dados para selecionar alguns atributos e excluir outros; sendo assim, os métodos de filtro são independentes do algoritmo de aprendizado que, simplesmente, receberá o subconjunto de atributos selecionados pelo filtro.

Na abordagem filtro, a meta é selecionar um subconjunto de atributos que preserve a informação pertinente no conjunto inteiro de atributos (Freitas, 1998, p. 67). Os algoritmos de busca do melhor subconjunto de atributos utilizados na abordagem filtro foram: *Genetic Selection* (GS) e outro *GreedyStepwise* (GSP).

O *Genetic Selection* é realizado através de algoritmos genéticos (AGs), que são algoritmos de otimização global, baseados nos mecanismos de seleção natural e da genética.

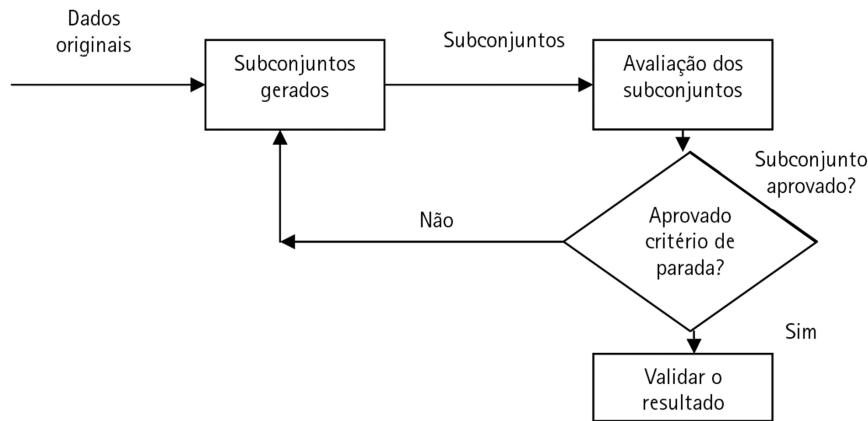


Figura 1. Passos na seleção de atributos.
Figure 1. Steps in selecting attributes.

Fonte: Liu e Motoda (1998, p. 23).

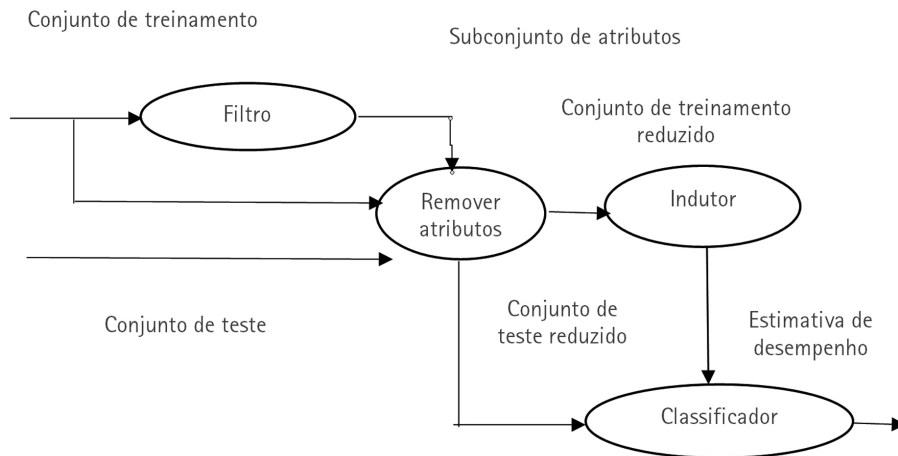


Figura 2. Abordagem filtro.
Figure 2. Filter approach.

Fonte: Lee (2005, p. 25).

Eles empregam uma estratégia de busca paralela e estruturada, embora aleatória, direcionada à busca de pontos de "alta aptidão", ou seja, pontos nos quais a função a ser minimizada ou maximizada tem valores relativamente baixos ou altos. Apesar de aleatórios, os AGs não são buscas aleatórias não direcionadas, pois exploram informações históricas para encontrar novos pontos de busca onde são esperados melhores desempenhos. Embora possam parecer simplistas do ponto de vista biológico, esses algoritmos são suficientemente complexos para fornecer mecanismos poderosos e robustos de busca adaptativa (Rezende, 2005, p. 227).

O método *GreedyStepwise* seleciona as variáveis passo a passo. GSP admite que uma variável selecionada em uma etapa possa ser eliminada posteriormente. Outra vantagem deste método é que o processo de seleção pode começar a

partir da equação construída com um subconjunto de variáveis independentes, ou com todas elas (Aranaz, 1996, p. 213).

O algoritmo de avaliação de atributos na seleção utilizado neste artigo foi o CFS (Seleção de atributos baseado em correlação). O CFS (Hall e Holmes, 2003) utiliza a seguinte métrica para fazer a avaliação:

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

Depois de calcular uma matriz de correlação, CFS aplica uma estratégia heurística de busca para encontrar um bom subconjunto de atributos de acordo com (2).

$$\text{Merit}_s = \frac{\overline{kr_{cf}}}{\sqrt{k + k(k-1)r_{ff}}} \quad (2)$$

onde $H(X)$ e $H(Y)$ de (2) são os atributos, e Merit_s heurística é o "mérito" de um subconjunto S_k contendo atributos, r_{cf} é a média da correlação entre classes, e r_{ff} a média da correlação entre os atributos.

ABORDAGEM WRAPPER

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 de algoritmos de aprendizado (neste trabalho, foram utilizadas regressão linear, regressão robusta 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 predeterminados (Freitas, 1998, p. 66).

No estudo, foram testados três algoritmos de aprendizagem, um com origem na econometria, um com origem em regras de indução e o terceiro no aprendizado de máquina: (i) regressão logística (RL); (ii) árvore de decisão (AD); e (iii) máquina de vetor suporte (MVS). Os três algoritmos de aprendizado serão explicados nos itens posteriores.

ANÁLISE DE COMPONENTES PRINCIPAIS (ACP)

Análise de componentes principais é um dos métodos estatísticos multivariados (Manly, 2008, p. 89) também muito utilizado como selecionador de atributos. O objetivo da análise é

tomar p variáveis X_1, X_2, \dots, X_p e encontrar combinações destas para produzir índices Z_1, Z_2, \dots, Z_p que sejam não correlacionados na ordem de sua importância e que descrevam a variação nos dados. A ideia básica do método é de transformar p variáveis tipicamente correlacionadas em $k < p$ combinações lineares não correlacionadas.

METODOLOGIA DA PESQUISA

Esta pesquisa é de natureza empírica, descritiva e quantitativa, compreendendo as empresas classificadas no SERASA e na BOVESPA como insolventes (concordatárias, em recuperação judicial e falidas) e abrangendo o período de 2005 a 2007.

DADOS BASEADOS NO MERCADO VERSUS DADOS BASEADOS NOS DEMONSTRATIVOS CONTÁBEIS

Em finanças corporativas, os estudos podem ter a sua origem nos dados contábeis, nos dados de mercado, ou podem utilizar os dois conjuntamente; a partir destas opções para estudo, Agarwal e Taffler (2008) compararam as alternativas de se utilizar os dados baseados no mercado com os dados baseados nos livros contábeis e chegaram ao resultado de que modelos de previsão de insolvência baseados em dados contábeis produzem significativamente maiores benefícios econômicos do que aqueles baseados no mercado. Já Blöchlinger e Leippold (2006) observaram que as diferenças nos erros são economicamente significantes a favor dos dados contábeis. Como alternativa para este trabalho, adotaram-se somente dados baseados nos demonstrativos contábeis.

A AMOSTRA EM ESTUDO

Para o objetivo deste estudo, foram necessários dois tipos de amostras. O primeiro tipo refere-se a um conjunto de empresas chamadas de "problemáticas" no período de 2005 a

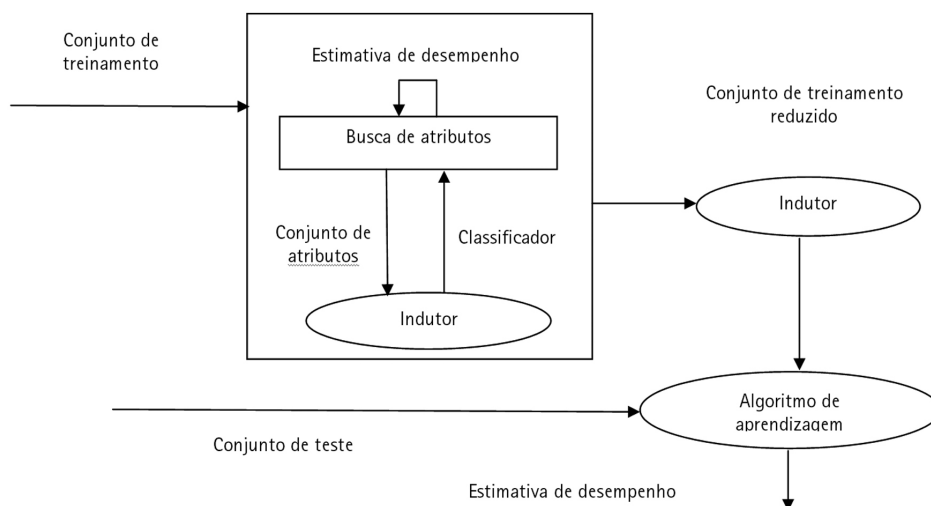


Figura 3. Abordagem wrapper.

Figura 3. Wrapper approach.

Fonte: Lee (2005, p. 26).

2007, isto é, que apresentaram problemas de insolvência em um determinado período. Tal problema é aqui entendido como referente a empresas classificadas no SERASA ou na BOVESPA como concordatárias, em recuperação judicial ou falidas.

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. Nesse segundo grupo, foram incluídas empresas "saudáveis", relacionadas ao primeiro grupo, da seguinte forma: para cada empresa do primeiro grupo, foram relacionadas duas empresas financeiramente saudáveis (empresas solventes), com tamanho do ativo, e pertencentes 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.

COLETA DOS DADOS

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

Segundo Gil (2010, 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 bibliográfica se utiliza fundamentalmente das contribuições dos diversos autores sobre determinado assunto, a pesquisa documental vale-se de materiais que não receberam ainda tratamento analítico, ou que ainda podem ser elaborados de acordo com os objetivos da pesquisa".

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

O período escolhido das empresas "problemáticas" foi o intervalo de tempo entre 2005 e 2007, 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 (Anexo 1) anuais das empresas selecionadas, no período de 2003–2007. 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 Schrickel (1999) e Pereira (2006).

TRATAMENTO DOS DADOS

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. Foram aplicadas três técnicas de seleção de atributos, abordagem filtro, abordagem *wrapper* e análise de componentes principais (ACP).

Já a técnica de *data mining* que foi aplicada é a de classificação, feita através de três técnicas com metodologias distintas: regressão logística, árvore de decisão e máquina de vetor suporte. No tratamento dos dados, foi utilizado o software de *data mining* WEKA 3.5.6 (Witten *et al.*, 2011).

AVALIAÇÕES DOS RESULTADOS DAS ABORDAGENS DE SELEÇÃO DOS ATRIBUTOS

Para avaliar os resultados das abordagens de seleção de atributos, foram aplicados três classificadores com metodologias distintas: regressão linear generalizada ou regressão logística, árvore de decisão e máquina de vetor suporte.

Regressão logística é uma técnica de regressão linear generalizada que se destina a aferir a probabilidade de ocorrência de um evento e a identificar características dos elementos pertencentes a cada grupo determinado pela variável categórica. Esta regressão não pressupõe a existência de homogeneidade de variância e normalidade dos resíduos (Fávero *et al.*, 2009, p. 441).

Já a árvore de decisão toma como entrada um objeto ou situação descrito por um conjunto de atributos e retorna uma "decisão" – o valor de saída previsto, de acordo com a entrada (Russell e Norvig, 2004, p. 633).

Por fim, a máquina de vetor suporte são funções não lineares de *kernel* que mapeiam os vetores de entrada em um espaço de mais alta dimensão (espaço de características), onde um hiperplano de separação é obtido para a resolução de problemas de classificação. As funções de *kernel* são geralmente executadas por funções não lineares como polinômios, funções radiais e *perceptron* (Rezende *et al.*, 2005, p. 162).

Para a avaliação da precisão dos modelos gerados pelos classificadores, foram utilizadas matriz de confusão, validação cruzada, medida F e área ROC (Receiver Operating Characteristic). A matriz de confusão é uma tabela onde são representados os Tp (verdadeiros positivos), Tn (verdadeiros negativos), Fp (falsos positivos) e Fn (falso negativos). A validação cruzada permite que todos os dados da base de dados sejam utilizados para treinamento e neste teste foram adotados 10 subconjuntos, e a medida F é a média entre precisão e *recall* eles medem a capacidade de reconhecer os exemplos negativos e positivos (Witten *et al.*, 2011, p. 175).

A área ROC é um gráfico bidimensional em que o eixo horizontal representa a taxa de erro da classe negativa ($1 - S_{pec}$) e, no eixo vertical, estão representados os valores de sensibilidade. O desempenho do classificador é medido pela área sob a curva ROC (Witten *et al.*, 2011, p. 172).

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.

RESULTADO DA SELEÇÃO

No Quadro 1, são apresentadas as variáveis preditoras selecionadas pelos respectivos métodos; vale ressaltar que os valores subscritos nas variáveis se referem ao período temporal da variável. Por exemplo, *Liquidez imediata₁*, se refere a dois anos antes do ano da declaração da insolvência da entidade.

ANÁLISE DESCRITIVA – INDICADORES NAS EMPRESAS INSOLVENTES SELECIONADOS

A Tabela 1 mostra a evolução de indicadores mais bem selecionados pelas abordagens estudadas nos dois últimos períodos anteriores (T-1 e T-2) e no período da insolvência (T), para o caso do grupo das empresas insolventes.

Para o indicador referente à liquidez seca (LS), considerado como um índice bem relacionado com a capacidade de liquidez de curtíssimo prazo de uma empresa, verifica-se que suas médias vão se deteriorando ao longo dos períodos. Os desvios-padrão mostram uma relativa homogeneidade nos seus valores, não apresentando variações relevantes. A variação mais significativa se refere à passagem do período T-1 para o período T (81,84%), no qual ocorre o "colapso final". Essa variação associada aos índices de coeficientes de variação permite supor a ocorrência de um rápido comprometimento na capacidade das empresas de quitar suas obrigações de curtíssimo prazo, valendo-se de seus ativos mais líquidos.

No caso da variável Margem Operacional após o Resultado Financeiro (MORF), as médias se deterioram ano a ano; elas mostram um decréscimo na capacidade operacional das empresas, podendo se supor que ganham menos em seus negócios. Já os valores dos desvios-padrão não apresentam variações significativas. Com os coeficientes de variação as dispersões entre as médias são significativas, não apresentando nem tendências nem regularidades representativas durante os períodos estudados.

A evolução da variável Saldo de Tesouraria sobre Ativo Total (STAT) é coerente com a situação de empresas em fase pré-falimentar, uma vez que suas médias vão se deteriorando 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 para empresas pré-falimentares, com diminuição dos valores no decorrer dos períodos. O que se observa neste estudo é a perda gradativa de rentabilidade das empresas insolventes na amostra aqui utilizada. Há maior deterioração durante a transição do período T-2 para T-1 ($\Delta ROAT = 144,36\%$), sugerindo aumento de ineficiência operacional e, consequentemente, financeira, conduzindo a empresa ao estado de insolvência mais agudo e comprometedor.

Associando as quatro variáveis mais bem selecionadas, pode-se deduzir que houve perdas na capacidade operacional das empresas estudadas, comprometendo sua saúde financeira e, por conseguinte, levando essas entidades à insolvência.

ANÁLISE DESCRITIVA - INDICADORES NAS EMPRESAS SOLVENTES SELECIONADOS

A evolução dos indicadores contábeis, nas empresas solventes, se comportou com certa normalidade, não apresentando grandes saltos ou dispersões entre os valores, o que não deixa de ser novidade para empresas que são sucesso de continuidade.

Quando se comparam os índices das empresas solventes com os das empresas insolventes, pode-se verificar que os coeficientes de variação, excetuando os de LS, apresentam distorções bem acentuadas nas empresas solventes, bem superiores aos das empresas insolventes. Já para as outras variáveis, média e DP, apresentam distorções, mas não tão acentuadas como os de CV. Talvez seja interessante um estudo mais detalhado sobre tais distorções, o que não é o objetivo desta pesquisa.

RESULTADOS DAS TÉCNICAS APLICADAS COM OS RESPECTIVOS CLASSIFICADORES

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

Na abordagem *wrapper*, somente serão apresentados os resultados encontrados quando o algoritmo indutor for o mesmo do classificador. Os resultados encontrados quando do uso do algoritmo indutor diferente do classificador foram pouco significativos.

RESULTADOS DOS SUBCONJUNTOS DE ATRIBUTOS APLICANDO COMO CLASSIFICADOR REGRESSÃO LOGÍSTICA

Na Tabela 3, pode ser constatado que a abordagem *wrapper* (WR RL) obteve os melhores resultados tanto na matriz de confusão como na medida F, na área ROC e na classificação correta do grupo de origem em que seu percentual foi de 89,88%.

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

Chart 1. Summary table of selected attributes with their techniques.

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

Nota: Wr/RL/AD/MPL = Abordagem de seleção de atributos wrapper tendo como algoritmos de aprendizagem regressão logística (RL), árvore de decisão (AD) e máquina de vetor suporte (MVS); Fil/SG/SP = Abordagem de seleção de atributos filtro tendo como algoritmo de busca *search genetic* (SG) e como algoritmo de busca *greedy stepwise* (SP); ACP = Análise de componentes principais.

Tabela 1. *Evolução dos indicadores no grupo das insolventes.*Table 1. *Evolution of indicators in the group of insolvent.*

Variável	Período T-2			Período T-1			Período T		
	Média	DP	CV (%)	Média	DP	CV (%)	Média	DP	CV (%)
LS	0,6203	0,6138	98,52	0,5225	0,661	126,52	0,435	0,541	124,36
MORF	0,3	0,51	170	-0,51	0,257	-50,39	-0,57	1,35	-236,84
STAT	-0,193	0,381	200,27	-0,211	0,425	201,421	-0,297	0,4638	126,76
ROAT	-0,1098	3098	281,72	-0,2545	0,5583	219,37	-0,291	0,4272	143,77

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

Tabela 2. *Evolução dos indicadores no grupo das empresas solventes.*Table 2. *Evolution of indicators in the group of solvent companies.*

Variável	Período T-2			Período T-1			Período T		
	Média	DP	CV (%)	Média	DP	CV (%)	Média	DP	CV (%)
LS	0,85	0,7334	86,28	0,862	0,6521	75,64	0,86	0,6981	81,17%
MORF	0,21	1,79	873,13	0,89	6,42	721,34	0,73	5,41	741,10%
STAT	0,0018	0,227	12611,1	-0,0007	0,0976	1401,44	0,0021	0,169	8047,62%
ROAT	0,0015	0,277	18466,7	-0,0019	0,1598	8410,52	0,001	0,2431	24310,00%

Tabela 3. *Resultados dos subconjuntos de atributos selecionados pelas abordagens wrapper, filtro e ACP aplicando como classificador regressão logística.*Table 3. *Results of selected subsets of the wrapper approaches, filter and ACP applying as logistic regression classifier.*

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

RESULTADOS DOS SUBCONJUNTOS DE ATRIBUTOS APLICANDO COMO CLASSIFICADOR ÁRVORE DE DECISÃO

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

RESULTADOS DOS SUBCONJUNTOS APLICANDO COMO CLASSIFICADOR MÁQUINA DE VETOR SUPORTE

Na Tabela 5, pode ser constatado que a abordagem *wrapper* foi a que obteve os melhores resultados. A melhor classificação correta do grupo de origem foi de 97,02%.

QUADRO COMPARATIVO DOS RESULTADOS

Pela Tabela 6, podemos comparar os resultados gerados pelos modelos compostos de atributos selecionados e avaliados pelas abordagens filtro, *wrapper* e análise de componentes principais. Os modelos elaborados pelos subconjuntos avaliados pela abordagem *wrapper* apresentaram melhores índices de classificação, em todos os classificadores estudados. Cabe ressaltar que a melhor *performance* foi com o classificador máquina de vetor suporte com 97,02% de classificação correta no grupo de origem, com as melhores medidas *Fs* e a área ROC.

Tabela 4. Resultados dos subconjuntos de atributos selecionados pelas abordagens *wrapper*, filtro e ACP aplicando como classificador árvore de decisão.

Table 4. Results of selected subsets of the wrapper approaches, filter and ACP applying as decision tree classifier.

ALGORITMO DE APRENDIZAGEM – AD								
Técnica de seleção	WR AD		GS		SP		ACP	
Grupo de origem	I	S	I	S	I	S	I	S
Insolventes – I	51	5	50	6	51	5	50	6
Solventes – S	6	106	8	104	8	104	11	101
Medida F	0,903	0,951	0,877	0,937	0,887	0,941	0,855	0,922
Área Roc	0,942	0,942	0,953	0,953	0,958	0,958	0,931	0,931
Classificação correta no grupo de origem	93,45%		91,86%		92,26%		89,88%	

Tabela 5. Resultados dos subconjuntos de atributos selecionados pelas abordagens *wrapper*, filtro e ACP aplicando como classificador máquina de vetor suporte.

Table 5. Results of selected subsets of the wrapper approaches, filter and ACP applying as support vector machine classifier.

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

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

Table 6. Ranking of the best results of forecasting models for each selection technique.

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

ANÁLISE DE RESULTADOS

O objetivo deste estudo foi de comparar três abordagens de seleção de atributos em banco de dados elaborado com dados contábeis de empresas brasileiras, e os resultados obtidos mostram que a abordagem *wrapper* obteve os melhores resultados na seleção de subconjuntos de atributos nos classificadores aplicados. Dentre os algoritmos utilizados pela abordagem filtro, o *greedystepwise (Sp)* foi o que melhores resultados obteve. Dentre os classificadores, o que alcançou o melhor resultado foi máquina de vetor suporte com 97,02% de acertos na classificação correta no grupo de origem. Em todos os classificadores de avaliação, a abordagem *wrapper* obteve os melhores resultados entre os testados; nesta abordagem, as variáveis selecionadas foram capazes de realizar melhor a tarefa de classificação. Tais afirmações foram evidenciadas na Tabela 6. Não se pode afirmar que a abordagem *wrapper* seja geralmente a mais eficaz; 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, a abordagem *wrapper*, para a amostra estudada, mostrou a maior eficácia. 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 óbvia a respeito da usabilidade da abordagem *wrapper* é que deve ser usada para avaliar subconjuntos de atributos que irão compor os modelos preditivos.

CONCLUSÕES E FUTUROS ESTUDOS

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 abordagens, quatro foram "determinadas" pelos métodos estudados (LS, MORF, STAT, ROAT), podendo levar à 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.

Este resultado assume importância ainda maior ao se considerar as modificações ocorridas no mercado, que se tornou mais competitivo, fortalecendo empresas que toleraram rentabilidades menores para manter a sua continuidade.

Cabe ressaltar que as variáveis STAT e MORF representativas do período mais próximo da insolvência foram selecionadas por todos os métodos estudados, o que evidencia o seu poder de representar aquelas empresas com alta probabilidade de vir a se tornarem insolventes.

Outra importância apresentada pelo estudo é a presença da variável MORF entre as mais representativas para discriminar empresas solventes das insolventes. Normalmente entende-se que a insolvência está mais relacionada à perda da capacidade de endividamento pela empresa, com a presença da variável MORF, entre aquelas bem representativas das empresas insolventes, a margem operacional é importante para entender a insolvência da empresa, cabendo, inclusive, estudos mais detalhados sobre esta variável nas próximas pesquisas sobre o tema.

Apesar da qualidade dos dados contábeis ser 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 90% 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 possa 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.

Sugere-se, nos estudos futuros, a aplicação de outras técnicas de seleção de atributos, além do uso de outros classificadores com técnicas de DM, como combinação de múltiplos classificadores, balanceamento do banco de dados, obtenção de regras de classificação para empresas insolventes.

REFERÊNCIAS

- AGARWAL, V.; TAFFLER, R. 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of Banking & Finance*, 32(8):1541-1551. <http://dx.doi.org/10.1016/j.jbankfin.2007.07.014>
- ALTMAN, E.I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4):589-609. <http://dx.doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- 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. [http://dx.doi.org/10.1016/0378-4266\(94\)90007-8](http://dx.doi.org/10.1016/0378-4266(94)90007-8)
- ARANAZ, M.F. 1996. *SPSS para Windows – Programación y análisis estadístico*. Madrid, McGraw-Hill, 580 p.
- BACK, B.; LAITINEN, T.; KAISA, S. 1996. Neural networks and genetic algorithms for bankruptcy prediction. *Expert Systems with Applications*, 11(4):407-413. [http://dx.doi.org/10.1016/S0957-4174\(96\)00055-3](http://dx.doi.org/10.1016/S0957-4174(96)00055-3)
- BALCAEN, S.; OOGHE, H. 2006. Thirty-five years of studies on business failure: an overview of the classical statistical methodologies and their related problems. *The British Accounting Review*, 38(1):63-93. <http://dx.doi.org/10.1016/j.bar.2005.09.001>

- BEAVER, W. 1968. Alternative accounting measures as predictors of failure. *The Accounting Review*, 43(1):112-122.
- BLÖCHLINGER, A.; LEIPPOLD, M. 2006. Economic benefit of powerful credit scoring. *Journal of Banking & Finance*, 30(3):851-873. <http://dx.doi.org/10.1016/j.jbankfin.2005.07.014>
- DEAKIN, E.B. 1972. A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10(1):167-179. <http://dx.doi.org/10.2307/2490225>
- 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. <http://dx.doi.org/10.2307/2329929>
- EISENBEIS, R.A. 1997. Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal the Finance*, 32(3):875-900. <http://dx.doi.org/10.1111/j.1540-6261.1977.tb01995.x>
- FÁVERO, L.P.; BELFIORE, P.; DA SILVA, F.L.; CHAN, B.L. 2009. *Análise de dados: modelagem multivariada para tomada de decisões*. Rio de Janeiro, Elsevier, 672 p.
- FAYYAD, U.; PIATETSKY, S.G.; SMITH, P. 1996. From data mining to knowledge discovery: an overview. *Advances in Knowledge Discovery & Data Mining*, 7(3):1-34.
- FITZPATRICK, P.J. 1932. A comparison of ratios of successful industrial enterprises with those of failed firms. *Certified Public Accountant*, October, p. 598-605; November, p. 656-662; December, p. 727-731.
- FREITAS, A.A. 1998. *Data mining and knowledge discovery with evolutionary algorithms*. New York, Springer-Verlag, 264 p.
- GIL, A.C. 2010. *Como elaborar projetos de pesquisa*. 5ª ed., São Paulo, Atlas, 200 p.
- 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. <http://dx.doi.org/10.1109/TKDE.2003.1245283>
- HÄRDLE, W.; MORO, R.A.; SCHÄFER D. 2005. Predicting bankruptcy with support vector machines. SFB 649, Discussion Paper 2005-2009. Disponível em: <http://sfb649.wiwi.hu-berlin.de>. Acesso em: 10/03/2006.
- HAYKIN, S. 2001. *Redes neurais: princípios e práticas*. 2ª ed., Porto Alegre, Bookman, 900 p.
- HORTA, R.A.M. 2001. *Utilização de indicadores contábeis na previsão de insolvência – análise empírica de uma amostra de empresas comerciais e industriais brasileiras*. Rio de Janeiro, RJ. Dissertação de Mestrado. Universidade do Estado do Rio de Janeiro, 108 p.
- KANITZ, S.C. 1978. *Como prever falências*. São Paulo, Mc Graw-Hill do Brasil, 176 p.
- KASZNAR, I.K. 1986. *Falências e concordatas de empresas: modelos teóricos e estudos empíricos*. Rio de Janeiro, RJ. Dissertação de Mestrado. Escola Brasileira de Economia e Finanças (EPGE – FGV), 198 p.
- KOHAVALI, R.; JOHN, G.H. 1997. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273-324. [http://dx.doi.org/10.1016/S0004-3702\(97\)00043-X](http://dx.doi.org/10.1016/S0004-3702(97)00043-X)
- LEE, H.D. 2005. *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. São Carlos, SP. Tese de Doutorado. Universidade de São Paulo, 182 p.
- LENNOX, C. 1999. Identifying failing companies: a reevaluation of the Logit, Probit and MDA approaches. *Journal of Economics and Business*, 5(4):347-364. [http://dx.doi.org/10.1016/S0148-6195\(99\)00009-0](http://dx.doi.org/10.1016/S0148-6195(99)00009-0)
- LIU, H.; MOTODA, H. 1998. *Feature selection for knowledge discovery and data mining*. Norwell, Kluwer Academic Publishers, 440 p. <http://dx.doi.org/10.1007/978-1-4615-5689-3>
- MANLY, B.F.J. 2008. *Métodos estatísticos multivariados*. 3ª ed., Porto Alegre, Bookman, 230 p.
- PIRAMUTHU, S. 2006. On preprocessing data for financial credit risk evaluation. *Expert Systems with Application*, 30(3):489-497. <http://dx.doi.org/10.1016/j.eswa.2005.10.006>
- REZENDE, S.O. (org.). 2005. *Sistemas inteligentes: fundamentos e aplicações*. Barueri, Manole, 550 p.
- RUSSEL, S.; NORVIG, P. 2004. *Inteligência Artificial*. Rio de Janeiro, Elsevier, 1030 p.
- SANVICENTE, A.Z.; MINARDI, A.M.A.F. 2000. Identificação de indicadores contábeis significativos para previsão de concordata de empresas. Disponível em: http://www.risktech.br/artigos/artigos_técnicos/index.html. Acesso em: 05/02/2005.
- SCHRICHEL, W.K. 1999. *Demonstrações financeiras*. 2ª ed., São Paulo, Atlas, 370 p.
- SHIRATA, C.Y. 2001. Financial ratios as predictors of bankruptcy in Japan: an empirical research. Disponível em: <http://www.shirata.net/apira98.html>. Acesso em: 09/08/2006.
- SILVA, J.P. da. 2006. *Gestão e análise de risco de crédito*. 5ª ed., São Paulo, Atlas, 448 p.
- TSAL, C.F. 2009. Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22(2):120-127. <http://dx.doi.org/10.1016/j.knosys.2008.08.002>
- WINAKOR, A.; SMITH, R. 1935. *Changes in the financial structure of unsuccessful industrial corporations*. Bulletin n. 51, vol. 32. Urbana, Bureau of Business Research, University of Illinois.
- WITTEN, I.H.; FRANK, E.; HALL, M.A. 2011. *Data mining: practical machine learning tools and techniques*. 3ª ed., Burlington, The Morgan Kaufmann, 664 p.

Submitted on July 26, 2012

Accepted on December 22, 2014

ANEXO 1**VARIÁVEIS CONTÁBEIS COLETADAS**

Liquidez corrente – LC, Liquidez seca – LS, Liquidez Imediata – LI, Liquidez Geral – LG, Endividamento Oneroso sobre Patrimônio Líquido – EOPL, Endividamento Total sobre o Patrimônio Líquido – EOAT, Endividamento Oneroso de Curto Prazo sobre Ativo Total – EOCpOT, Grau de Alavancagem Financeira – GAF, Imobilizado dos Recursos Permanentes – IMCP, Margem Bruta – MB, Margem Operacional – MO, Margem Operacional após Resultado Financeiro – MORF, Margem Líquida – ML, Giro do Ativo – GA, Rentabilidade do Ativo Operacional – ROA, Retorno dos Acionistas – ROE, Retorno do Investimento Total – ROI, Saldo de Tesouraria sobre Ativo Total – STAT, Modelo Dupont Adaptado – RTA, Lucro antes dos juros, impostos, Lucro antes dos juros, impostos, depreciações/exaustão e amortização – EBITDA.

RUI AMÉRICO MATHIASI HORTA

Doutor em Engenharia Civil pela Universidade Federal do Rio de Janeiro

Professor Adjunto do Depto de Finanças e Controladoria da Universidade Federal de Juiz de Fora

Rua José Lourenço Kelmer, s/n, Campus Universitário, São Pedro, 36036-900, Juiz de Fora, MG, Brasil

FRANCISCO JOSÉ DOS SANTOS ALVES

Doutor em Controladoria e Contabilidade pela Universidade de São Paulo

Programa de Pós Graduação em Ciências Contábeis da Universidade Estadual do Rio de Janeiro

Rua São Francisco Xavier, 524, 9º andar, Bloco E, Maracanã, 20550-013, Rio de Janeiro, RJ, Brasil

FREDERICO A. DE CARVALHO

Doutor em Sciences Économiques pela Université Catholique de Louvain

Faculdade de Administração e Ciências Contábeis da Universidade Federal do Rio de Janeiro

Av. Pasteur, 250, sala 242, Praia Vermelha, 22290-240, Rio de Janeiro, RJ, Brasil

MARCELINO JOSÉ JORGE

Doutor em Engenharia da Produção pela Universidade Federal do Rio de Janeiro

Laboratório de Pesquisas em Economia das Organizações de Saúde do Instituto Nacional de Pesquisa Clínica Evandro Chagas da Fundação Oswaldo Cruz (LAPECOS/INI/Fiocruz)

Av. Brasil, 4365, Manguinhos, 21040-900, Rio de Janeiro, RJ, Brasil