DESCONTINUIDADE DE EMPRESAS BRASILEIRAS DO SETOR DE CONSUMO NÃO CÍCLICO: UM ESTUDO COM DADOS CONTÁBEIS UTILIZANDO TÉCNICAS DE *DATA MINING*

DISCONTINUANCE OF BRAZILIAN COMPANIES FROM NON-CYCLIC CONSUMER GOODS INDUSTRY: A STUDY WITH FINANCIAL DATA USING DATA MINING TECHNIQUES

***Rui Américo Mathiasi Horta*[1]**
***Francisco Jose dos Santos Alves* [2]**
***Marcelino José Jorge*[3]**

Resumo: Este estudo aplica técnicas de data mining para analisar a descontinuidade de empresas brasileiras do setor de consumo não cíclico, com o objetivo de identificar as empresas com maior probabilidade de sofrerem descontinuidade de suas atividades. A base de dados é originada de demonstrações contábeis do período 2000-2011, de empresas listadas na BOVESPA. É utilizada estratégia de *data mining* para selecionar os atributos. Os resultados obtidos e as validações realizadas evidenciam sucesso da estratégia adotada, podendo ser considerada bem competitiva com outras estratégias utilizadas em estudos similares.

Palavras-chave: Descontinuidade de empresas. Consumo não-cíclico. Contabilidade. *Data mining*. Balanceamento de base de dados.

Abstract: This research applies data mining techniques to analyze discontinuance of Brazilian companies from non-cyclic consumer goods industry, aiming to identify the ones with higher probability of discontinuity. The study comprises years 2000-2011 from financial reports of companies listed in the Brazilian Stock Exchange (BOVESPA). Data mining strategy was used to selects the attributes. The results and validations performed show success of the selected strategy, being considered competitive with other strategies used in similar studies.

Key-words: Discontinuity of companies. Brazilian non-cyclic consumer goods industry. Accounting. Data mining. Balancing database.

[1]Doutor em Engenharia Civil pela COPPE/UFRJ e Mestre em Ciências Contábeis pela UERJ, rui.horta@ufjf.edu.br.
[2]Doutor em Controladoria e Contabilidade pela FEA-USP, francisco.jose.alves@terra.com.br.
[3]Doutor em Engenharia da Produção pela COPPE/UFRJ, marcelino.jorge@ipec.fiocruz.br.

1 Introdução

Este estudo tem como objetivo propor uma estratégia de balanceamento da base de dados com a seleção de atributos originados de demonstrativos contábeis de empresas pertencentes ao setor econômico de consumo não cíclico, visando melhorar a capacidade de caracterizar as empresas deste setor que podem vir a se tornar descontínuas.

O desenvolvimento de estudos na área de modelagem em descontinuidade de empresas cada vez mais vem adquirindo importância nas áreas acadêmicas e empresarial. Ela permite que seja possível prever uma situação financeira difícil com certa antecedência, de forma que haja tempo hábil para serem adotadas medidas que reverta essa situação impedindo a geração de grandes custos sociais e financeiros.

Vários fatores tem concorrido para o aumento dos estudos nesta área, dentre eles podem ser citados, (i) em vários países a maioria das estatísticas de falências mostrou um significativo aumento de sua ocorrência em comparação à recessão anterior, (ii) nas últimas décadas o ambiente econômico geral das empresas, na grande maioria dos países, tem mudado com uma enorme velocidade e experimentado tendências para baixo (dificuldades financeiras) (iii) a implementação, em vários países, de normas internacionais de contabilidade e finanças (tais como Basiléia II e III, Solvência II, Sarbanes-Oxley e IFRS) e (iv) permissão legal, no ambiente brasileiro, para alimentação de bases de dados por parte de instituições financeiras e não financeiras de clientes para análise de concessão de crédito.

Apesar das inúmeras pesquisas na área, há ainda questões pouco exploradas (BALCAEN E OOGHE, 2006; TSAI E WU, 2008; NANNI E LUMINI, 2009; GESTEL *et al*., 2010; ZHOU L., 2013). Uma dessas questões envolve o desequilíbrio de classes. Em mercados com ambientes econômicos normais, o número de empresas insolventes é bem menor do que o de empresas solventes, gerando um problema de desequilíbrio de classes. Para Japkowicz e Stephen (2002, p. 431):

> (...) os sistemas de aprendizado (classificadores) normalmente encontram dificuldades em induzir o conceito relacionado à classe minoritária. Nessas condições, modelos de classificação (de descontínuas) que são otimizados em relação à sua precisão têm tendência de criar modelos triviais, predizendo bem a classe majoritária.

Com isso, gera-se uma supremacia na classificação das empresas contínuas sobre as descontínuas, distorcendo o objetivo principal desta modelagem que é o de caracterizar melhor as empresas descontínuas.

Neste contexto este estudo tem por objetivo propor uma estratégia de balanceamento da base de dados com a seleção de atributos (*data mining*) originados de demonstrativos contábeis de empresas classificadas na BOVESPA (Bolsa de Valores de São Paulo) no setor econômico de consumo não cíclico, visando melhorar a capacidade de caracterizar as empresas deste setor que podem vir a se tornar descontínuas.

> O pressuposto comum assumido na predicão de insolvência é de que os principais indicadores macro-econômicos (inflação, os juros, impostos, etc.), juntamente com as características da empresa (concorrência, gestão, capacidade produtiva, produto, etc.) estão devidamente

> refletidos nos demonstrativos contábeis, então a futura situação financeira da empresa pode ser prevista usando dados provenientes desses demonstrativos utilizando técnicas de modelagem avançadas (GESTEL *et al.*, 2010, p. 2956).

O artigo está organizado da seguinte forma: a seção 2 apresenta a revisão da literatura que dará suporte ao desenvolvimento da pesquisa; na seção 3 descrevem-se os procedimentos metodológicos realizados. Na seção 4, apresentam-se os resultados obtidos e, na seção 5 são feitas as conclusões da pesquisa e sugeridos futuros estudos.

2 Fundamentação téorica

A descontinuidade das empresas tornou-se um assunto mais pesquisado e difundido a partir na década de 60 através do modelo Escore-Z de Altman (1968). O mesmo Altman *et al.,* (1977) desenvolve um novo modelo de classificação de insolvência chamado Zeta™, uma atualização e aprimoramento do modelo Escore-Z original. Martin (1977) elaborou um modelo de previsão em que utilizou regressão logística (RL). Ohlson (1980) empregou modelo *logit* para previsão de falência de firmas. West (1985) utilizou análise fatorial para compor as variáveis.

No século XXI começaram a prevalecer, nas pesquisas de descontinuidade de empresas, aplicações de técnicas de aprendizagem de máquina. Min, Lee e Han (2006) propuseram métodos para melhorar o desempenho de SVM em dois aspectos: a seleção de atributos e otimização de parâmetros. Para Hua *et al.,* (2007) SVM foi aplicado no problema de previsão de falências, e provou ser superior aos métodos concorrentes, como a rede ANNs, as múltiplas abordagens de LDA (análise discriminante) e RL. Ding, *et. al.* (2008) desenvolveram um modelo de previsão de insolvência em SVM para um exemplo de empresas chinesas de alta tecnologia. Kim e Sohn (2009) com SVM elaboraram um modelo para prever insolvência em pequenas e médias empresas no setor de tecnologia sul coreanas.

Alguns autores além de classificarem também desenvolveram metodologias para os classificadores e/ou na base de dados visando obter melhores resultados nas classificações, esses estudos são citados a seguir.

Atiya (2001) desenvolveu um estudo sobre previsão de insolvência no qual é aplicado redes neurais no modelo em bancos de dados desbalanceados. West *et al*., (2005) investigaram três estratégias de comitê de classificadores (combinação de classificadores) em aplicações de decisões financeiras incluindo previsão de insolvência, visando obter modelos com maiores acurácia. Hung *et al*., (2009) aplicaram probabilidade híbrida baseada em comitê de classificadores para previsão de insolvência utilizando votação majoritária e votação ponderada.

Tsai e Wu (2008) estudaram o desempenho de um classificador simples de ANNs com os (diversificado) múltiplos classificadores baseados em ANNs. Nanni e Lumini, 2009 desenvolveram uma metodologia de mineração de dados para a previsão de insolvência de empresas italianas.

No Brasil uma das principais dificuldades ainda é a escassez de pesquisas desenvolvidas com o propósito de encontrar parâmetros para previsão de insolvência, além da escassez de dados adequados e confiáveis para a realização deste estudo. Essa situação começa a ser mudada, mas ainda se está bem longe de poder fazer esse tipo de trabalho com a facilidade de obtenção de

dados como ocorre em outros países. A seguir serão apresentados alguns trabalhos que acabaram conquistando destaque no estudo sobre o tema no Brasil.

Elizabetsky (1976), Kanitz (1978), Matias (1978) trabalharam em modelos de previsão de insolvência utilizando análise discriminante. A metodologia dos trabalhos seguintes de Altman, Baidya e Dias (1979), utilizaram também a ferramenta estatística de análise discriminante. Já Morozini *et al.,* (2006) utilizaram análise dos componentes principais para evidenciar os principais índices entre os selecionados para o estudo. Guimarães e Moreira (2008) utilizaram indicadores contábeis com o uso de análise discriminante para propor um modelo de previsão de insolvência. Silva Brito *et al.,* (2009) empregaram a regressão logística para examinar se eventos de default de companhias abertas no Brasil são previstos por um sistema de classificação de risco de crédito baseado em índices contábeis.

3 Metodologia

3.1 Base de dados e métricas de avaliação

Foram obtidos no BOVESPA 20 indicadores (Anexo) contábeis anuais das empresas do setor econômico de consumo não cíclico classificadas de acordo com grupos de índices contábeis-financeiros: liquidez, endividamento e rentabilidade. Para Iudícibus (1998, p.62) os índices contábeis-financeiros permitem detectar os problemas da entidade que merecem uma investigação maior e também podem facilitar nas decisões gerenciais quando analisados em conjunto.

Estas empresas foram classificadas como concordatária, em recuperação judicial ou falida na BOVESPA durante o período de 2000 a 2011. Para cada empresa classificada como descontínua, foi selecionada uma quantidade superior de empresas de capital aberto com controle privado nacional, financeiramente saudável (no sentido de que não há solicitação de concordata por parte da empresa no período considerado), com tamanho de seu ativo e pertencente ao mesmo setor de atividade. O estabelecimento de uma quantidade superior de empresas adimplente para cada inadimplente, por outro lado, baseia-se na hipótese de que quanto maior a quantidade de dados existentes, menor a probabilidade de erros objetivando, também ficar mais próximo da realidade econômica.

Neste grupo de empresas há 25 instâncias representando as insolventes (descontínuas) e 170 instâncias representando as empresas solventes (contínuas), também deste setor. A reduzida dimensão da amostra se deve principalmente a não obrigatoriedade de um grande número de empresas de publicar seus demonstrativos contábeis.

A base foi composta por dados referentes aos demonstrativos contábeis dos cinco anos anteriores ao ano em que a empresa foi declarada insolvente. De acordo com Altman *et al.,* (1994) e Hung e Chen, (2009) as empresas insolventes começam a apresentar características ou indícios de insolvência num período de cinco anos antes ao ano que ocorre efetivamente a falência.

Os dados das empresas solventes são de dez anos, facilitando assim uma melhor caracterização dessas empresas. Pretendeu-se também: (i) uma adequação ao ano (2005) no qual ocorreu a mudança na lei de falência, (ii) utilizar demonstrativos contábeis sem a influência da inflação e (iii) estudar um período de tempo em que as empresas pertencentes a este setor econômico

tiveram que adequar suas estratégias gerenciais e operacionais devido a inúmeras mudanças do ambiente econômico.

Vale aqui evidenciar a importância e os principais motivos da escolha de um conjunto de empresas pertencentes a um setor econômico. Para Iudícibus (2008, p. 91)

> (...) empresas de mesmo setor econômico apresentam em seus demonstrativos contábeis semelhanças devido a suas estruturas patrimoniais e econômicas. Indicadores como de liquidez, endividamento e rentabilidade, por exemplo, devem apresentar valores bem próximos na sua média setorial. Empresas que apresentam índices bem distintos ao da média setorial no qual pertencem, devem apresentar situações com certas anomalias econômicas ou financeiras principalmente, portanto, em condições normais os seus indicadores também devem apresentar comportamentos ajustados.

Neste setor econômico as empresas se caracterizam com valores proporcionalmente baixos em seus ativos imobilizados e valores mais substanciais em seus ativos circulantes e com baixos níveis de endividamento. O sector de consumo não-cíclico comporta as empresas que se dedicam à venda de bens alimentares, bebidas, tabaco e produtos de casa. Este sector caracteriza-se pela sua menor correlação com a evolução da economia. Em épocas de menor crescimento econômico, ou mesmo de recessão, o consumo de produtos deste sector tende a não ser particularmente afetado. Os bens essenciais de consumo têm uma menor elasticidade preço do que quaisquer outros produtos e serviços. Significa isto que quando os preços aumentam a procura destes produtos não se ressente tanto quanto a procura de outro gênero de produtos, como por exemplo, os tecnológicos.

Do ponto de vista do ciclo dos mercados, o sector de consumo não-cíclico enquadra-se nos momentos finais do *bull market*. Assim, enquanto que no início de um *bull market*, o mercado se vira para os setores mais pró-cíclicos e para as *growth stocks*, deixando esquecido este sector, nos momentos finais do mesmo, quando as perspectivas escurecem, os títulos deste setor surgem como ilhas de valor seguro, o mesmo valor seguro que antes fora encarado como valor aborrecido. O Beta histórico deste setor ronda os 0,80, ou seja, por cada subida do S&P de 1% o setor sobe apenas 0,80%.

Empresas deste setor, durante o período estudado, passaram por várias alterações em seus ambientes macro e micro econômicos influenciando em suas estratégias gerenciais e operacionais refletindo em seus demonstrativos contábeis. No período estudado, segundo a BOVESPA, empresas deste setor representavam em média 8,2% do total.

Das métricas de avaliação alternativas existentes para lidar com o problema do desequilíbrio de classes citadas por Käuck (2004), e Gary (2004) foram escolhidas: a matriz de confusão (MC), área sob a curva ROC (AUC) e medida F (F). Para a avaliação do classificador serão utilizadas validação cruzada com 10 sub-amostras e resubstituição (BRAGA-NETO, *et al*., 2004). Também será utilizada a técnica da votação majoritária (LI HUI E JIE SUN, 2009) para acurar os resultados.

3.2 Técnicas de tratamento de base de dados desbalanceados

A abordagem baseada em amostras é amplamente usada para resolver o problema de desequilíbrio de classe. A ideia da amostragem é modificar a distribuição de instâncias de forma que a classe minoritária seja bem representada no conjunto. Entretanto, o excesso de amostragem pode causar superajustamento (*overfitting*) porque alguns dos exemplos podem ser replicados muitas vezes.

Na determinação da proporção de 50% para a classe das solventes e insolventes utilizou-se o estudo de Weiss e Provost (2001) no qual afirma que "alocar 50% dos exemplos de aprendizagem para a classe minoritária frequentemente apresentam resultados superiores aos resultados obtidos com a distribuição natural das classes".

Uma maneira de solucionar o problema de classes desbalanceadas numa base de dados é balancear artificialmente a sua no conjunto de exemplos. Duas abordagens principais são utilizadas neste estudo, são elas:

a) Remoção de exemplos da classe majoritária - u*nder-sampling* e
b) Inclusão de exemplos da classe minoritária - *over-sampling*.

Alguns trabalhos recentes têm tentado superar as limitações existentes tanto nos métodos de *under-sampling*, quanto aos métodos de *over-sampling*. Por exemplo, Chawla *et al.*, (2002) combinam métodos de *under* e *over-sampling*, em seu trabalho (SMOTE: *Synthetic Minority Over-sampling Technique*) o método *over-sampling* não replica os exemplos da classe minoritária, mas cria novos exemplos dessa classe por meio da interpolação de diversos exemplos da classe minoritária que se encontram próximos. Dessa forma, é possível evitar o problema do superajustamento. O algoritmo SMOTE, que é uma técnica bem citada na literatura específica, servirá como comparativo com a metodologia aqui proposto.

3.3 Uma estratégia para a predição de empresas insolventes

Descreve-se, nesta subseção, um método – SEID - construído especificamente para a predição de insolvência em uma base de dados desbalanceada, composta por variáveis originadas de demonstrativos contábeis de empresas brasileiras.

Um dos principais modos para tratar uma base de dados desbalanceada baseia-se (a) em procedimentos randômicos de diminuição dos dados da classe majoritária (*under-sampling*), (b) no incremento dos dados da classe minoritária por meio da replicação randômica com reposição (*over- sampling*), e (c) na combinação dessas duas estratégias. Neste caso, não existe geração de novas instâncias: o balanceamento é feito com a simples manipulação da base de dados original.

Ao contrário, o SMOTE é exemplo de estratégia de inserção de novas instâncias, geradas artificialmente, na classe minoritária. A maior dificuldade neste caso é a falta de garantia de que as instâncias sintéticas venham a pertencer, de fato, à classe a que foram associadas. Ambas as classes de estratégia baseiam-se em um processo totalmente estocástico para a obtenção de bases balanceadas.

O modelo desenvolvido busca diminuir este componente estocástico, visando (i) a utilização dos dados da classe minoritária de forma mais intensa ou redundante, pois busca-se maior nível de

acerto nesta classe, tal como é intuitivamente desejável em problemas de previsão de insolvência; e (ii) a decomposição da classe majoritária de forma a torná-la de dimensão "aceitavelmente" mais próxima a classe minoritária. É importante ressaltar que a obediência a estes dois objetivos acarreta, como característica adicional, a diminuição da aleatoriedade na obtenção do balanceamento. Assim este modelo denomina-se *Semi-Determinístic Ensemble Strategy for Imbalanced Data* (SEID). A forma definida para levar em conta aqueles dois objetivos conjuntamente foi utilizar um conjunto (ou comitê) de classificadores (TSAI; WU, 2008; NANNI; LUMINI, 2009).

Um procedimento de comitê apresenta, naturalmente, uma facilidade de implementação dos objetivos para cada classe, tal como descrita acima. No caso da necessidade de redundância das unidades minoritárias, tem-se a facilidade de utilizá-las em cada base do comitê. No caso das unidades majoritárias, em que se pretende particionar ou decompor os subconjuntos, pode-se colocar algumas de suas unidades em bases diferentes para gerar os classificadores que formam o comitê. Desta forma, a partição não prejudica nem a representatividade dos dados da classe majoritária, que devem compor pelo menos uma base de dados do comitê, nem a dimensão do banco, pois uma estratégia de comitê lida bem com bancos menos completos, por não basear a decisão em somente um dos classificadores gerados. Além disto, os parâmetros para determinar tamanhos mínimos da base dos classificadores do comitê servem para evitar a utilização de bases com dimensões consideradas inadequadas. Esta estratégia para balanceamento baseada em comitê permite o uso de um procedimento de seleção de características de forma diferenciada, descrita mais adiante.

Vamos examinar, agora, o método para predição de insolvência em empresas. Considera-se, inicialmente, a composição do conjunto:

$$Str = Str_m \cup Str_M,$$

ou seja, o conjunto formado pela união das unidades da classe minoritária (Str_m) e da classe majoritária (Str_M), sendo #(Str_M) > #(Str_m), onde #(*) significa número de elementos do conjunto.

Os conjuntos gerados para a obtenção dos classificadores iniciais serão balanceados com n_{ic} instâncias de cada classe c. Para que se obtenham conjuntos de acordo com o modelo proposto, adota-se para o valor mínimo n_{ic} do número de unidades para a classe o seguinte:

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

onde n_{cb} é o número de classificadores iniciais usados no comitê de classificadores e o operador max (*) calcula o maior valor entre os avaliados. Quanto maior o valor de n_{ic}, mais próximo o algoritmo se torna do algoritmo de *bagging*, ou seja, é um algoritmo que cria amostras repetidamente a partir de um conjunto de dados de acordo com uma distribuição uniforme de distribuição.

O SEID, visando equilibrar a base de dados, divide a base de dados original em um número impar (neste estudo em três) de sub-bases para mais adiante ser possível fazer a votação majoritária. Destas subases divididas quando se tem a classe de insolventes minoritária o procedimento é se igualar essa subamostra com dados das empresas insolventes retirando os

dados excedentes das empresas solventes. Já quando a subamostra prevalecem os dados das empresas insolventes o SEID retira dados das empresas insolventes até igualar a subamostra. Com isso sempre prevalecerá os dados daquelas empresas nos quais se pretende caracterizar melhor, ou seja, as insolventes. A seguir são classificadas as sub-bases para depois, e através dos resultados aplicarem a votação majoritária. Com essa estratégia há diminuição da aleatoriedade na obtenção do balanceamento e as bases ficam mais próximos da realidade. O pseudo-código do SEID pode ser obtido por solicitação ao autor principal.

3.4 Seleção de atributos

A seleção de atributos representa um problema de fundamental importância em mineração de dados (DM), sendo frequentemente realizada como uma etapa de pré-processamento. Os objetivos principais da seleção de atributos para previsão de insolvência, segundo Piramuthu (2006), 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.

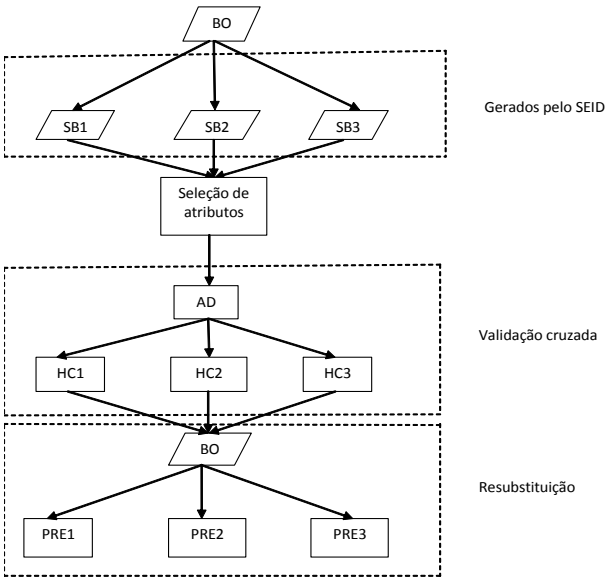
Neste trabalho foram utilizadas duas abordagens de direção de busca (Witten e Frank, 2011, p.293): (i) seleção *forward* e (ii) seleção aleatória. Estas abordagens foram escolhidas por serem bem citadas na literatura especifica. Avaliar o subconjunto de atributos selecionados é medir quão bom um determinado atributo é segundo um critério de avaliação (informação, distância, dependência, consistência, precisão). Em outras palavras, é como ele interage com o algoritmo de aprendizado. Essa interação pode ser subdividida, basicamente, em duas abordagens principais, filtro e *wrapper*. A abordagem *wrapper* é a utilizada neste artigo. Para Somol *et al*. (2005) a abordagem *wrapper* é preferida para ser utilizada quando se refere a estudos sobre insolvência de empresas.

3.5 Uma estratégia de predição de insolvência com seleção de atributos - SEIDwS

Apresenta-se, nesta seção, uma técnica para seleção de atributos a ser acoplada ao modelo de predição desenvolvido (SEID), completando a proposta deste trabalho. A ideia é considerar a aplicação dos métodos de seleção de forma individualizada nas bases que compõem o comitê, configurando o modelo proposto - *Semi-Determinístic Ensemble Strategy for Imbalanced Data* with attribute Selection (SEIDwS). O pseudo código do modelo para predição de insolvência com a inclusão do procedimento de seleção de atributos também pode ser obtido por solicitação ao autor principal.

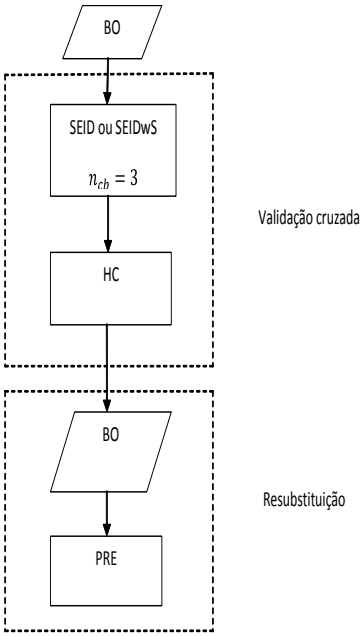
A seguir será apresentado os fluxogramas do SEIDwS, nas Figura 1 e 2.

Figura 1 - Fluxograma referente aos procedimentos para se chegar aos resultados após os balanceamentos e a seleção de atributos da base de dados original



Legenda das siglas na Figura 1:
BO: Base de dados original;
SB: Sub-bases geradas pelo SEID;
AD: Classificador árvore de decisão;
HC: Modelos gerados após a seleção de atributos e a aplicação do classificador;
PRE: Resultados das classificações encontrados após testar os modelos gerados na base de dados original.

Figura 2 – Procedimento de classificação com o SEID ou SEIDwS.



OBSERVAÇÃO: As siglas na Figura 2 são as mesmas da Figura 1,
entretanto aqui é sintetizado o processamento da estratégia apresentada.

As Figuras 1 e 2 oferecem uma visualização dos procedimentos do SEIDwS. A base de dados original (BO) aplicando o SEIDwS é subdividida em três sub-bases por meio da técnica de seleção de atributos. Nessas sub-bases é feita a classificação, gerando três modelos (por validação cruzada) para serem testados na base de dados original (resubstituição). A seguir aplica-se a técnica da votação majoritária.
A validação do algoritmo proposto será realizada em duas etapas, visando atender dois objetivos - (i) testar os algoritmos aqui propostos em bases de dados diferentes daquelas aqui estudadas; (ii) comparar os resultados gerados pelo SEIDwS com outras pesquisas realizadas nesse tema.

O cumprimento da primeira etapa (i) consistiu em testar o algoritmo SEIDwS em três bases de dados originadas do Repositório UCI para Aprendizado de Máquina - *Japanese Credit Screening, Australian Credit Approval, German Credit Data*. Essas bases são normalmente utilizadas para estudos sobre modelagem de previsão de insolvência (ver http://archive.ics.uci.edu/ml).

3.5.1 Validação do SEIDwS nas bases do Repositório UCI para Aprendizado de Máquina

Nesta subseção são apresentados os resultados da validação do SEIDwS através das três bases do UCI, o procedimento é o mesmo apresentado na Figura 1. O classificador AD (árvore de decisão) foi o utilizado. A Tabela 1 apresenta os resultados dos testes do SEIDwS e do algoritmo SMOTE, neste algoritmo o "*k*" (o número de vizinhos mais próximos) usado foi igual a 5. As bases de dados utilizadas do UCI foram as que normalmente são utilizadas para testes na

previsão de insolvência, nelas as classes são discriminadas em empresas insolventes (INS) e solventes (SOL).

Os softwares utilizados foram o WEKA 3.5.6 (Witten e Frank, 2011) e o Matlab 7.1. Em todas as análises apresentadas nas etapas de classificação e de seleção de atributos foram aplicadas 10 partições na validação cruzada.

A abordagem seleção aleatória utilizou o algoritmo de busca *Genetic Selection* (GS). Em GS foram usados os valores do tamanho da população e do número das gerações iguais a 20, e a probabilidade de *crossover* e mutação igual a 0,6 e 0,033, respectivamente. Na abordagem filtro, as técnicas de avaliação de atributos (Witten e Frank, 2011, p.422) foram medidas de consistência (*consistency*) e o CFS (Seleção de atributos baseado em correlação). Na abordagem *wrapper* os algoritmos indutores foram RL, ANNs, SVM e AD. Os resultados apresentados nas tabelas seguintes, que obtiveram os melhores resultados, utilizaram GS, *wrapper* e AD.

Tabela 1 – Resultados dos testes do algoritmo SEIDwS com as bases de dados sobre insolvência do UCI.

Bases de dados do UCI	Nº de atributos	Classe	Instâncias	Base original		SEIDwS		SMOTE	
				F	AUC	F	AUC	F	AUC
Japanese Credit Screening	6	INS	383	0,38	0,57	0,73	0,88	0,77	0,91
		SOL	307	0,77	0,57	0,90	0,88	0,90	0,91
Australian Credit Approval	5	INS	383	0,00	0,47	0,57	0,88	0,79	0,93
		SOL	307	0,96	0,47	0,97	0,88	0,98	0,93
German Credit Data	7	INS	300	0,55	0,73	0,88	0,94	0,81	0,93
		SOL	700	0,81	0,73	0,93	0,94	0,93	0,93

A Tabela 1 apresenta os resultados dos testes feitos com algoritmos de balanceamentos (SEIDwS e SMOTE). Pelos resultados apresentadas na Tabela 1 o SEIDwS apresentou resultados bem promissores e competitivos com o SMOTE como estratégia de balanceamento quando testado nas bases de dados do UCI.

A Tabela 2 apresenta as comparações com estudos publicados sobre o tema na literatura específica utilizando como parâmetros acurácia (classificações corretas), Erro Tipo I e Erro Tipo II (classifica instância falidas no grupo das não falidas). As comparações foram feitas através dos melhores resultados publicados pelos autores. Os estudos utilizados para comparação são de Tsai e Wu (2008), Tsai (2009) e Nanni e Lumini (2009). Estes autores utilizaram as bases de dados do UCI, as mesmas utilizadas pelo algoritmo proposto neste artigo, o SEIDwS.

Tabela 2 – Comparação dos resultados do algoritmo SEIDwS com as bases de dados do UCI com estudos publicados

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

Na Tabela 2, os resultados mostram a eficácia do algoritmo SEIDwS. A comparação mostra que SEIDwS obteve melhores resultados na acurácia, nos Erros Tipo I e II, e que em todos esses parâmetros há um ganho do SEIDwS sobre os outros estudos. No Erro Tipo II o SEIDwS obteve melhores resultados sobre os outros testes em dois das três bases de dados. Na base *German Credit Data*, a mais desbalanceada, os resultados foram os melhores.

Vale aqui ressaltar que os resultados apresentados na Tabelas 1 e 2 foram gerados através de bases de dados (*Japanese Credit Screening, Australian Credit Approval, German Credit Data*) com variáveis distintas às variáveis contábeis aplicando duas diferentes estratégias de balanceamento. O propósito é de validar a estratégia de balanceamento proposta neste artigo, ou seja, a estratégia SEIDwS apresenta resultados competitivos a outras estratégias divulgadas pela literatura específica.

4 Análise e Interpretação dos dados

Nesta seção serão apresentados os resultados das aplicações na base de dados referente as variáveis obtidos em demonstrativos contábeis de empresas do setor de consumo cíclico dos classificadores sem a estratégia SEIDwS (sub-seção 4.1) e a aplicação dos classificadores após o uso da estratégia SEIDwS (sub-seção 4.2), na sub-seção seguinte os resultados aplicando votação majoritária e na sub-seção 4.4 são comparados os resultados encontrados nas sub-seções anteriores com os resultados gerados pelo algoritmo SMOTE.

4.1 Aplicações de classificadores na base de dados sem aplicar estratégia do balaceamento

As técnicas aplicadas para a classificação das empresas foram: Regressão Logística (RL), Máquina de Vetor Suporte (SVM), *Multilayerperceptron* (MLP), e Árvore de Decisão (AD). Estes classificadores foram escolhidos por serem considerados eficientes bem como por serem largamente utilizados na determinação de insolvência de empresas.

Foram feitos ajustes paramétricos iniciais para cada classificador utilizado, visando obter uma parametrização adequada para esta base. Os resultados apresentados foram obtidos por meio de validação cruzada com 10 partes. Para que haja um melhor entendimento do desempenho de

cada classificador apresentam-se os resultados de cada classificador da matriz de confusão, medida F e AUC. A Tabela 3 apresenta os resultados dos classificadores da base de dados original.

Tabela 3- Resultados dos classificadores da base de dados original

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	19	6	0,745	0,927	20	5	0,889	0,9	20	5	0,877	0,923	22	3	0,88	0,942
S	7	163	0,962	0,927	0	170	0,986	0,9	2	168	0,86	0,923	3	167	0,982	0,942

Para a base de dados do setor econômico de empresas de consumo não cíclico, em relação a matriz de confusão, a medida F e área AUC apresentaram pouca diferença entre os classificadores. O método AD obteve um resultado superior, portanto sendo utilizado como algoritmo indutor no processo de seleção de atributos. As variáveis selecionadas empregando-se as técnicas estudadas (seção 3.4) foram: EOPL, ETPL, GAF, ROI, EOCpOT.

4.2 Aplicações de classificadores na base de dados após a aplicação da estratégia SEIDwS

Quando a seleção de atributos foi aplicada antes do balanceamento, os resultados encontrados não foram compatíveis para um nível mínimo aceitável em uma previsão de insolvência de empresas (valores de F e AUC próximos a 0,65). Diante disso, a etapa de seleção de atributos foi executada após a realização do balanceamento das bases de dados (Figura 1). A estratégia foi testada e os resultados são apresentados nos itens seguintes.

4.2.1 Resultados para a abordagem de seleção de atributos *wrapper* no SEIDwS

Nas aplicações com a abordagem *wrapper* (conforme Figura 1) foram testados para o método de busca GS (GeneticSearch) e GD (GreedyStepwise). GS obteve os melhores resultados. O classificador utilizado foi o AD por ter obtido os melhores resultados para essa base de dados foram apresentados na Tabela 2.

Tabela 4 – Resultado para as sub-bases utilizando seleção de atributos abordagem *wrapper*.

	SB1				SB2				SB3			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	25	0	0,862	0,947	25	0	0,877	0,95	25	0	0,877	0,95
S	8	162	0,976	0,947	7	163	0,978	0,95	7	163	0,978	0,95

Os resultados evidenciam a influencia do balanceamento seguido da seleção de atributos na abordagem *wrapper* de um ganho de desempenho em relação a classificação sem o balanceamento e sem a aplicação de técnicas de seleção de atributos (Tabela 3). Das 18 variáveis totais seis foram selecionadas pela abordagem *wrapper*. As variáveis selecionadas foram: EOPL, ETPL, GAF, ROI, EOCpOT, EOAT.

Com a aplicação da estratégia SEIDwS foram selecionadas mais uma variável EOAT. Como a estratégia SEIDwS melhora a caracterização das empresas potencialmente insolventes pode-se deduzir que nas empresas pertecentes ao setor econômico de consumo não cíclico as variáveis que representam os desempenhos de estrutura de capital dessas empresas são importantes para a discriminação das solventes e insolventes.

4.3 Balanceamento e seleção de características para base de dados com votação majoritária - SEIDwS

Nesta seção é aplicada a estratégia SEIDwS desenvolvida para a predição de insolvências em empresas. Na pratica, a aplicação completa do SEIDwS é obtida com o uso da votação majoritária (LI HUI e JIE SUN, 2009) em relação aos resultados dos modelos das sub-bases obtidas na definição da instância que está sendo avaliada. Desta forma, as sub-bases passam a representar um comitê de classificadores conforme descrito anteriormente.

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

Os atributos selecionados continuam os mesmos para cada sub-base. Porém, a votação majoritária deve aumentar a robustez na predição obtida para as instâncias avaliadas. O procedimento foi exposto nas Figuras 3.1 e 3.2. No caso do SEIDwS, são avaliados os melhores algoritmos de seleção determinados para as estratégias filtro e *wrapper*. Os resultados obtidos são mostrados na Tabela 5.

Tabela 5 - Resultados referentes à base de dados balanceadas aplicando SEIDwS

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

Fonte: Autores.

4.2.4 Comparação dos resultados encontrados

Na Tabela 6 é feita uma comparação da base original com os melhores resultados encontrados pelo SEIDwS utilizando modelo *wrapper e o SMOTE*.

Tabela 6 – Comparação dos resultados

	BASE ORIGINAL				SEIDwS				SMOTE			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	22	3	0,88	0,942	25	0	0,963	0,992	24	1	0,948	0,993
S	3	167	0,982	0,942	6	164	0,97	0,992	4	166	0,983	0,993

Pela Tabela 6 pode ser concluído que o balanceamento com a seleção de atributos e um comitê de classificadores (SEIDwS) melhoraram bem a capacidade de caracterização das empresas classificadas como insolventes, os resultados da MC, F e AUC evidenciam esses ganhos (BASE ORIGINAL x SEIDwS).

No mesmo quadro a comparação do SEIDwS com o SMOTE pode ser evidenciado a melhor capacidade do SEIDwS de caracterizar aquelas empresas classificadas como insolventes (I) em relação ao SMOTE – MC e F . Somente na classificação geral (AUC) o SEIDwS obteve um resultado inferior em relação ao SMOTE (0,992 x 0,993). Diante desses resultados pode ser concluído da capacidade bem competitiva existente entre a estratégia aqui apresentada (SEIDwS) e o SMOTE, técnica bem referenciada na literatura específica.

5 Conclusões

Esta pesquisa apresentou e testou uma estratégia para solucionar um problema pouco estudado em modelagens para prever descontinuidade de empresas - o desequilíbrio entre as classes de empresas classificadas como solventes e as empresas classificadas como insolventes. Na grande maioria das pesquisas existentes a amostra estudada é uma *paired sample*, ou seja, composta com número igual de empresas solventes e insolventes. Esta paridade entre as classes de empresas representa mal a realidade do ambiente econômico, distorcendo a utilidade da amostra e, comprovadamente, priorizando a classificação correta somente para as empresas solventes. Nesta pesquisa buscou-se, através do procedimento proposto, adequar a base de dados ao ambiente econômico das empresas.

Primeiramente, foi elaborada uma base de dados secundários, inédita, contendo índices calculados a partir de demonstrativos contábeis de empresas do setor econômico de consumo cíclico, previamente classificadas como solventes e insolventes pela BOVESPA no período de 2000 a 2011. A essa base de dados original foi aplicada a estratégia aqui apresentada, denominada SEIDwS, que, após gerar três sub-bases, selecionou atributos priorizando os índices daquelas empresas classificadas como insolventes e não descartando os índices das empresas solventes no conjunto das sub-bases.

Para cada sub-base se obteve um modelo de classificação, posteriormente testado na base de dados original, gerando, assim, três resultados – um para cada uma das classificações. Na etapa seguinte foi realizada a votação majoritária dos resultados encontrados das três classificações na base original (resubstituição), obtendo, então, o resultado final das classificações.

Foram feitas validações utilizando dados disponíveis em bases de dados públicas e muito utilizadas na literatura especifica, no propósito de testar novas técnicas de modelagem. As bases utilizadas foram três: *japanese credit, australian credit* e *german credit*. Também efetuaram-se comparações com o chamado SMOTE e os resultados foram apresentados na Tabela 1. Na outra validação foram comparados os resultados obtidos pela estratégia SEIDwS com outros estudos contemporâneos, todos publicados na literatura específica (Tabela 2).

Possíveis extensões ao presente estudo deveriam contemplar a inclusão de novas técnicas de comitês de classificações e a inclusão de variáveis qualitativas na base de dados. Em ambos os casos deve resultar melhor capacidade preditiva dos modelos de previsão.

5.1 Implicações metodológicas

Diante das validações realizadas, os resultados obtidos revelaram que a estratégia apresentada – o SEIDwS – é bem competitiva em relação ao SMOTE, tendo boa capacidade de classificar aquelas empresas pertencentes à classe das insolventes e com resultados ainda melhores naquela base de dados em que o desequilíbrio entre as classes é mais crítico e mais acentuado, como é o caso do *German Credit*, apresentados na Tabela 1.

Na comparação feita entre os resultados encontrados pela aplicação do SEIDwS e outras estratégias publicadas na literatura específica, os resultados foram bastante animadores. Na Tabela 2 está evidenciada a capacidade do SEIDwS em relação a outros trabalhos contemporâneos. Na grande maioria dos resultados obtidos o SEIDwS foi mais eficiente do que

as outras estratégias, havendo ganhos na capacidade de classificar aquelas empresas pertencentes à classe das insolventes, o que era exatamente a situação que se queria melhorar.

Quando da aplicação da estratégia SEIDwS e do SMOTE à base de dados de empresas brasileiras, os resultados são ainda mais convincentes no que diz respeito à diferença na capacidade de classificar as empresas insolventes. O SEIDwS mostrou melhor capacidade para classificar aquelas empresas pertencentes à classe das insolventes do que o SMOTE (Tabela 6). Nesta mesma tabela é evidenciada a importância do tratamento da base dados para equacionar o problema do desequilíbrio das classes e para melhorar a capacidade do modelo de previsão na classificação das empresas pertencentes à classe das insolventes.

Pode-se argumentar, então, que o presente estudo ilustra bem a importância do desenvolvimento de estratégias para resolver o problema existente na maioria dos modelos de previsão de insolvência de empresas, a saber, o desequilíbrio ou desbalanceamento de classes. Além disso, o estudo apresenta uma estratégia que tem plenas condições de competir com a conhecida estratégia SMOTE.

Cabe ressaltar que o SEIDwS minora o efeito estocástico do modelo em relação ao SMOTE. Diante disso pode ser visto como uma contribuição metodológica às pesquisas sobre previsão de insolvência de empresas em bases desbalanceadas, tema ainda pouco explorado na literatura contábil no Brasil.

Dada a natureza do exercício empírico, fica evidente que a vantagem do novo procedimento não depende do setor estudado e nem da área de aplicação.

5.2 Implicações para a análise contábil

Em relação às variáveis contábeis selecionadas, prevaleceram às originadas do Balanço Patrimonial, sobretudo aquelas que aferem a composição (estrutura) das fontes passivas de recursos das empresas (EOPL, ETPL, GAF, EOCpOT). Depois da aplicação da estratégia SEIDwS somam-se àquelas a variável EOAT, também representativa da capacidade de endividamento das empresas.

O conjunto das variáveis selecionadas pode-se inferir que, para a amostra estudada, a descontinuidade das empresas do setor de consumo não cíclico está bem relacionada a aspectos relativos à incapacidade de se endividar. Para essas empresas quanto maior cada um desses índices, maior a probabilidade delas serem incapazes de cumprir suas obrigações e virem a se tornarem descontínuas. Em outras palavras, na amostra estudada empresas tornam-se insolventes porque perdem a capacidade (financeira) de se endividar independente de serem (EOPL, EOAT e EOCpOT) ou não obrigações onerosas (GAF e ETPL). Apesar de potencialmente esperada antes do exercício, esta conclusão pode ser considerada outra contribuição específica deste trabalho às pesquisas sobre previsão de insolvência no Brasil.

Uma terceira contribuição deste trabalho às pesquisas sobre previsão de insolvência foi a utilização de dados secundários obtidos exclusivamente em demonstrativos contábeis de empresas brasileiras. Durante algum tempo os dados contábeis, no Brasil, foram tratados com desconfiança, de modo que pesquisas e conclusões como as presentes servem para reiterar a

conveniência daquela utilização para a análise da evolução econômica de empresas em nosso país.

A quarta contribuição que pode ser considerada deste trabalho diz respeito ao uso exclusivo de dados originados de demonstrativos contábeis de empresas pertencentes a somente um setor econômico. Com isso ficam bem mais relacionado os motivos da insolvência daquelas empresas, não havendo influências de variáveis atribuíveis a diferenças setoriais.

Referências

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

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

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

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

ATIYA, Amir. F. Bankruptcy prediction for credit risk using neural network: a survey and new results. *IEEE transactions on neural networks*, v. 12 nº 4, july 2001, p. 929-935.

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

BRAGA-NETO, U. ; HASHIMOTO, Ronaldo; DOUGHERTY, Edward R.; NGUYEN, Danh V.; CARROLL, Raymond J. Is cross-validation better than resubstitution for ranking genes? Vol. 20 nº 2, 2004, p. 253-258. *DOI: 10.1093/bioinformatics/btg399*.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, 2002, p. 321-357.

DING, Yongsheng; SONG, Xinping, ZEN,Yueming. Forecasting financial condition of Chinese listed companies based on support vector machine**.** *Expert Systems with Applications*, v. 34, Issue 4, may 2008, Pages 3081-3089.

ELIZABETSKY, Roberto. *Um modelo matemático para decisão no banco comercial.* (Trabalho apresentado ao Departamento de Engenharia de Produção da Escola Politécnica da USP). São Paulo: USP, 1976.

GARY M. Weiss; KATE Mccarthy; BIBI, Zabar. Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?, In: *Proceedings of the 2007 International Conference on Data Mining*, CSREA Press, 35-41.

GESTEL, Tony Van; BAESENS, Bart; MARTENS, David. From linear to non-linear kernel based classifiers for bankruptcy prediction. *Neurocomputing, v.* 73, 2010, p. 2955–2970.

GUIMARÃES, A.; MOREIRA, T.B.S. Previsão de insolvência: um modelo baseado em índices contábeis com utilização de análise discriminante. *Revista de Economia Contenporânea*, Rio e Janeiro, v. 12, n.1, 2008, p. 151-178.

HUA, Zhongsheng; WANG, Yu; XU, Xiannoyan; ZHANG, Bin; LIANG, Liang. Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications*, v. 33, Issue 2, 2007, p. 434-440.

HUNG, Chihli; CHEN, Jing-Hong. A selective ensemble based on expected probabilities for bankruptcy prediction. *Expert systems with applications,* 2009, v. 36, Issue 3, p. 3297-5309.

IUDÍCIBUS, Sérgio de. *Análise de Balanços*. 9ª Ed. São Paulo: Atlas, 2008. 58 p.

IUDÍCIBUS, Sérgio de. *Contabilidade Gerencial*. 6ª Ed. São Paulo: Atlas, 1998. 336 p.

JAPKOWICZ N.; STEPHEN, S. The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis,* v. 6, Issue 5, 2002, p. 429-450.

KANITZ, Stephen Charles. *Como prever falências.* São Paulo: Mc Graw-Hill do Brasil, 1978. 174 p.

KÄUCK, H. Bayesian formulations of multiple instance learning with applications to general object recognition. Master's thesis, University of British Columbia, Vancouver, BC, Canada, 2004.

KIM, Hong Sik; SOHN, So Young. Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, v. *201, Issue 3*, 2010, p. 838-846.

LI HUI, JIE SUN. Majority voting combination of multiple case-based reasoning for financial distress prediction. *Expert Systems with Applications,* v.36, apr. 2009, p. 4363-4373.

MARTIN, D. Early warning of bank failure: A logit regression approach, *Journal of Banking and Finance,* v.1, 1977, p. 249–276.

MATIAS, Alberto Borges. *Contribuição às técnicas de análise financeira: um modelo de concessão de crédito.* (Trabalho apresentado ao Departamento de Administração da Faculdade de Economia e Administração da USP.) São Paulo: [s.n.], 1978, p. 82, 83, 90.

MIN,Sung-Hwan.; LEE, Jumin,;HAN. Ingoo. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. *Expert Systems with Applications,* v. 31, Issue 3, oct. 2006, p. 652-660.

MOROZINI, João Francisco; OLINQUEVITCH, José Leônidas; HEIN, Nelson. Seleção de índices na análise de balanços: uma aplicação da técnica estatística 'ACP'. REVISTA CONTABILIDADE & FINANÇAS USP. São Paulo, v. 2 Número 41, p. 87-99, Maio/Agosto 2006.

NANNI, Loris,; LUMINI, Alessandra. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications,* v. 36, Issue 2, Part 2, mar. 2009, p. 3028-3033.

OHLSON, J.A. Financial ratios and the probabilisitic prediction of bankruptcy. *Journal of Accounting Research*, v. 18, 1980, p.109-131.

PIRAMUTHU S. On preprocessing data for financial credit risk evaluation. *Expert Systems with Applications***,** v. 30, 2006, p.489-497.

SILVA BRITO, Giovani Antônio; ASSAF NETO, Alexandre; CORRAR, Luiz João. Sistemas de classificação de risco de crédito: uma aplicação a companhias abertas no Brasil. *Revista Contabilidade & Finanças USP*. São Paulo, v. 20 Número 51, p. 28-43, Setembro/Dezembro, 2009.

SOMOL P.; BAESENS B.; PUDIL P.; VANTHIENEN J., Filter-versus *Wrapper*-based Feature Selection for Credit Scoring, International Journal of Intelligent Systems, v. 20, Number 10, 2005, p. 985-999.

TSAI., C. F.; WU J. W. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with applications*, v. 34, Issue 4, may. 2008, p. 2639-2649.

VERGARA, Sylvia Constant. *Projetos e relatórios de pesquisa em administração*. 6 ed. São Paulo: Atlas, 2005. 96 p.

WEISS, G. M.; PROVOST, F. The Effect of Class Distribution on Classifier Learning: An Empirical Study. Technical Report ML-TR-44, Rutgers University, Department of Computer Science, 2001.

WEST, David; DELLANA, Scott; QIAN, Jingxia. Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*, v. 32, Issue 10, october 2005, p. 2543-2559.

WEST, R. C, A factor analytic approach to bank condition, *Journal of Banking and Finance*, v. 9, jun.1985, p. 253–266.

WITTEN, Ian .H.; FRANK, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques.* The Morgan Kaufmann Series in Data Management Systems, 3ª ed. 2011. 630 p.

ZHOU, Ligang. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods**.** *Knowledge-Based Systems*, *January 2013.*

Rui Américo Mathiasi Horta é Mestre em Ciências Contábeis pela Universidade Estadual do Rio de Janeiro e Doutor em Engenharia Civil pela Universidade Federal do Rio de Janeiro, é professor do departamento de finanças e controladoria da UFJF rui.horta@ufjf.edu.br Rua José Lourenço Kelmer, s/n - Campus Universitário Bairro São Pedro, Juiz de Fora - MG, 36036-900.	**Francisco Jose dos Santos Alves** é Doutor e Professor do Programa de Pós-Graduação em Ciências Contábeis da Universidade Estadual do Rio de Janeiro, francisco.jose.alves@terra.com.br Rua São Francisco Xavier, 524 – Maracanã – 9º Andar – Rio de Janeiro - RJ, 20550-013.	**Marcelino José Jorge** é Doutor e Professor do Instituto de Pesquisa Clínica Evandro Chagas – IPEC/FIOCRUZ, marcelino.jorge@ipec.fiocruz.br Av. Brasil 4.365 – Manguinhos, Rio de Janeiro – RJ, 21.040-360.

ANEXO

SIGLAS DAS VARIÁVEIS UTILIZADAS

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 Líquida – ML
Giro do Ativo – GA
Rentabilidade do Ativo Operacional – ROA
Retorno dos Acionistas – ROE
Retorno do Investimento Total – ROI
Termômetro Financeiro – TERFIN
Modelo Dupont Adaptado – RTA
Lucro antes dos juros impostos - EBIT
Lucro antes dos juros impostos depreciações/exaustão e amortização - EBITDA.