

PREVISÃO DE INSOLVÊNCIA: UMA ESTRATÉGIA PARA BALANCEAMENTO DA BASE DE DADOS UTILIZANDO VARIÁVEIS CONTÁBEIS DE EMPRESAS BRASILEIRAS – SEID

RESUMO: O tema previsão de insolvência vem cada vez mais sendo objeto de novos estudos e pesquisas porque 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. Este estudo tem adquirido mais importância devido às mudanças ocorridas nos ambientes de negócios, o aumento das bases de dados e o desenvolvimento de novas tecnologias de sistemas computacionais. No Brasil os estudos neste tema ainda sofrem o efeito por se encontrar bases de dados com dimensões reduzidas devido à qualidade dos dados disponíveis, felizmente essa situação vem se alterando. Utilizando dados originados de demonstrativos contábeis de empresas brasileiras listadas na BOVESPA, é apresentada uma estratégia de *data mining* que ataca o problema do desequilíbrio de classes, problema pouco estudado e existente neste tema porque nos ambientes econômicos normais o número de empresas classificadas como solventes são bem maiores do que aquelas classificadas como insolventes. Tal estratégia visa melhor caracterizar aquelas empresas que apresentam maiores potenciais de virem a se tornar insolventes atacando o problema do desequilíbrio entre as classes. Os resultados obtidos e as validações realizadas da estratégia apresentada evidenciam o seu sucesso podendo ser considerada bem competitiva com outras estratégias apresentadas na literatura específica.

Palavras-chave: Previsão de insolvência; Variáveis contábeis; *Data mining*; Comitê de classificadores; Balanceamento de base de dados.

1 INTRODUÇÃO

O desenvolvimento de estudos na área de modelagem em previsão de insolvência é relevante, pois permite prever com certa antecedência uma situação financeira difícil, de forma que haja tempo hábil para serem adotadas medidas que reverta essa situação impedindo a geração de grandes custos sociais e financeiros.

Nos anos recentes, uma revolução vem ocorrendo em relação à maneira pela qual a previsão de insolvência é medida e gerida (Saunders *et al.*, 2004). Alguns fatores concorrem para que isso ocorra: (i) em vários países a maioria das estatísticas de falências mostrou um significativo aumento de sua ocorrência em comparação à recessão anterior, (Saunders e Cornett, 2008); (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) como os mercados de capitais vêm se expandindo e se tornando acessíveis a um maior número de empresas, a desintermediação está ocorrendo rapidamente; (iv) quase que paradoxalmente, apesar de um declínio na qualidade média dos empréstimos (devido ao motivo anterior), as margens de juros ou *spreads*, especialmente em mercados de empréstimos por atacado, têm-se tornado muito estreitas, ou seja, a compensação de risco-retorno advinda de empréstimos vem piorando, (Saunders e Cornett, 2008); (v) concomitantemente com a recente crise financeira mundial, crises bancárias e de seguradoras em mercados de países desenvolvidos tem mostrado que quanto mais fracos e incertos forem os valores das garantias reais, mais arriscada se torna a avaliação da capacidade financeira de uma empresa; (vi) com os avanços das tecnologias computacionais, tais como os sistemas de computadores relacionados à tecnologia da informação junto com a evolução da disponibilidade e desenvolvimento de base de dados, há um aumento de oportunidades de testar técnicas cada vez mais sofisticadas de modelagem à previsão de insolvência (Chye e

Chin, 2004); (vii) o crescimento da exposição do crédito, ou risco de contrapartida, devido à fenomenal expansão de mercados de derivativos, estendeu a necessidade de análises mais acuradas sobre previsão de insolvência utilizando registros contábeis e extra contábeis de empréstimos. (Saunders, 2000); (viii) outro grande incentivo para instituições financeiras desenvolverem novos modelos de previsão de insolvência é a imposição pós-1992 de exigências de reservas de capital para empréstimos, pelo BIS¹ e por bancos centrais (Cornett, 2007); (ix) na linha de extensão das pesquisas acadêmicas dos impactos dos *mercados imperfeitos e informações assimétricas*, trabalhos de previsão de insolvência têm aumentado; (x) a implementação, em vários países, das normas internacionais de contabilidade e finanças como o IFRS.

Apesar das inúmeras pesquisas na área, há ainda questões pouco exploradas (Balcaen e Ooghe, 2006; Ravi *et al.*, 2008; Tsai e Wu, 2008; Nanni e Lumini, 2009; Verikas *et al.*, 2010; Chaudhuri e Kajal De, 2010; Gestel *et al.*, 2010).

Uma dessas questões é o problema do 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. Diante disso, os sistemas de aprendizado (classificadores) normalmente encontram dificuldades em induzir o conceito relacionado à classe minoritária (Japkowicz e Stephen, 2002). Nessas condições, modelos de classificação (de insolvência) que são otimizados em relação à sua precisão têm tendência de criar modelos triviais, que quase sempre predizem bem a classe majoritária, gerando uma supremacia na classificação das empresas solventes sobre as insolventes, distorcendo o objetivo principal desta modelagem que é o de melhor caracterizar as empresas insolventes.

Este estudo tem por objetivo propor uma estratégia de balanceamento da base de dados originada de demonstrativos contábeis, em conjunto com a utilização de comitês de classificadores (agregação de previsões de múltiplos classificadores), que melhore a capacidade de caracterizar as empresas que podem vir a se tornar insolventes.

O pressuposto comum assumido na predição 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).

O artigo está organizado da seguinte forma: a seção 2 apresenta a revisão bibliográfica 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 REVISÃO BIBLIOGRÁFICA

A previsão de insolvência 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. Ohlson (1980) empregou modelo *logit* para previsão de falência de firmas. West (1985) utilizou análise fatorial para compor as

¹ BIS – Banco para Compensações Internacionais.

variáveis. Jones e Hensher (2004) apresentaram um modelo de previsão de falência baseado em *logit* misto. Canbas *et al.*, (2005) propuseram a integração do sistema de alarmes (IEWS) combinando análise discriminante, regressão logística, *probit* e análise de componentes principais.

Odom e Sharda (1990) compararam a habilidade de predição de redes neurais artificiais (ANNs) e análise de discriminante (LDA) no risco de falência. Roy e Cosset (1990) também compararam ANNs com regressão logística para avaliar risco de insolvência em países usando dados econômicos e políticos. Várias pesquisas em previsão de insolvência incluindo Lacherr *et al.*, (1995), Sharda e Wilson (1996), Tam e Kiang (1992) e Wilson e Sharda (1994) Zhang *et al.*, (1999) relataram que ANNs produzem significativas melhoras na acurácia dos modelos de predição comparados com aqueles elaborados com técnicas estatísticas.

Lee e Cheng (2005) fizeram um estudo com o propósito de avaliar um previsor de crédito modelando com redes neurais artificiais e *multivariate adaptive regression splines* (MARS). Chen e Du (2009) utilizaram redes neurais e técnicas de mineração de dados para elaborar modelos de previsão de insolvência. Youn, Hyewon *et al.*, (2009) desenvolveram modelos de previsão de insolvência de empresas baseados em redes neurais artificiais e regressão logística com variáveis financeiras.

Sinh *et al.*, (2005) investigaram a eficácia da aplicação de SVM (Máquina de Vetor Suporte) para o problema de previsão de falências, eles mostraram que o classificador SVM supera BPN (*Backpropagation*) para problemas de previsão de falências de empresas. Min e Lee (2005) aplicaram SVM para o problema de previsão de falência. Min, Lee e Han (2006) propuseram métodos para melhorar o desempenho da SVM em dois aspectos: a seleção de atributos e otimização de parâmetros. Chen e Shih (2006) propuseram um modelo de classificação automática para as classificações de crédito através da aplicação de SVM. Para Hua *et al.*, (2007) SVM foi aplicado no problema de previsão de falências, e provou ser superior aos métodos concorrentes, como a rede neural, as múltiplas abordagens discriminante linear e regressão logística. 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.

Dimitras, *et al.*, (1999) usa regras de indução para fornecer um conjunto de regras capazes de prever insucesso empresarial. Tay e Shen (2002) demonstraram que os modelos com regras de indução são aplicáveis a problemas práticos relacionados com a previsão econômica e financeira. Park e Han (2002) através do CBR (classificador baseado em instâncias) desenvolveram um modelo de previsão de insolvência. Chen *et al.*, (2009) em seu artigo procuram oferecer uma alternativa para a modelagem de previsão de falência utilizando neuro *fuzzy*, uma abordagem híbrida que combina a funcionalidade de lógica *fuzzy* e da capacidade de aprendizagem das redes neurais. Nwogugu (2006) apresenta vários modelos dinâmicos de falência. Premachandra *et al.*, (2009) propuseram análise envoltória de dados (DEA) como ferramenta para avaliar a falência da empresa. Tseng e Li (2005) propuseram um modelo *logit* quadrático com base em uma abordagem de programação quadrática para processar variáveis de resposta binária.

Alguns autores além de classificarem também desenvolveram pesquisas específicas para seleção de atributos e 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 de redes neurais no modelo com bancos de dados desbalanceados. West *et al.*, (2005)

investigaram três estratégias de comitê de classificadores em aplicações de decisões financeiras incluindo previsão de insolvência, visando obter modelos com maiores acurácia: validação cruzada (*cross validation*), *bagging* e *boosting*. Hung *et al.*, (2007) aplicaram probabilidade híbrida baseada em comitê de classificadores para previsão de insolvência utilizando votação majoritária (*majority voting*) e votação ponderada (*weighted voting*). Huang *et al.*, 2007 investigaram três estratégias para construção de modelos de híbridos baseado no SVM para *credit scoring* e compararam suas performances com redes neurais (ANNs), algoritmo genético (GA) e árvore de decisão. Tsai e Wu (2008) estudaram o desempenho de um classificador simples de redes neurais com os (diversificado) múltiplos classificadores baseados em redes neurais. Yu L. *et al.*, (2008) utilizam ANNs para avaliar o risco de crédito com a aplicação de comitês de classificadores. Ravi V. *et al.*, (2008) elaboram e testam comitê de classificadores para previsão de insolvência. Nanni e Lumini, 2009 desenvolveram uma metodologia de mineração de dados para a previsão de insolvência de empresas.

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 com 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), Pereira (1982, apud. Silva, 2006, p.266) utilizaram também a ferramenta estatística de análise discriminante. Bragança e Bragança (1984) combinaram métodos estatísticos (análise discriminante) com dados obtidos no DOAR (Demonstrativo de Origens e Aplicações de Recursos) das empresas, principal contribuição do estudo. Kasznar (1986) aplicou análise discriminante para desenvolver um modelo linear. Carmo (1987) utilizou modelos de funções lineares obtidas a partir de modelos fatoriais e de análise de componentes principais. Santos *et al.*, (1996) construíram um modelo fundamentado em análise discriminante capaz de obter indicações sobre a saúde financeira de empresas industriais. Sanvicente e Minardi (1998) desenvolveram um trabalho utilizando análise discriminante. Horta (2001) elaborou modelos de previsão de insolvência para empresas utilizando as técnicas estatísticas de análise discriminante e regressão logística na etapa de seleção de atributos. Morozini *et al.*, (2006) utiliza análise dos componentes principais para evidenciar os principais índices entre os selecionados para o estudo. Silva Brito *et al.*, (2009) utilizaram a técnica estatística de 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 DA PESQUISA

Este estudo tem como objetivo propor uma estratégia de balanceamento da base de dados originada de demonstrativos contábeis, em conjunto com a utilização de comitês de classificadores, que melhore a capacidade de caracterizar as empresas que podem vir a se tornar insolventes.

Quanto à metodologia utilizada para alcançar o objetivo da pesquisa, pode-se classificá-la sob dois enfoques, segundo Vergara (2005): quanto aos fins e quanto aos meios. Quanto aos fins, o presente estudo classifica-se como metodológica e aplicada. Quanto aos meios, a presente pesquisa enquadra-se como experimental e bibliográfica.

3.1 Técnicas de tratamento de bancos desbalanceados

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

- a) Remoção de exemplos da classe majoritária - *under-sampling*;
- 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*, nesse trabalho (SMOTE: Synthetic Minority Over-sampling Technique) o método de *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.

3.2 Uma estratégia para solucionar o problema do desequilíbrio de classes para a predição de empresas insolventes – SEID

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

Um dos principais modelos para trato de base de dados desbalanceadas baseia-se em procedimentos randômicos de diminuição dos dados da classe majoritária (*under-sampling*), incremento dos dados da classe minoritária por meio da replicação randômica com reposição (*over-sampling*), e na combinação das duas estratégias. Neste caso, não se tem a geração de novas instâncias, simplesmente o balanceamento é feito com a manipulação da base de dados original. O SMOTE tem como estratégia a inserção de novas instâncias geradas artificialmente na classe minoritária. A maior dificuldade é a falta de garantia que se tem das instâncias sintéticas pertencerem realmente à classe a que foram associadas.

Deve-se destacar que estas estratégias 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 um maior nível de acerto nesta classe; ii) a decomposição da classe majoritária de forma a torná-la de dimensão mais próxima a classe minoritária.

É importante ressaltar que a obediência a estes dois objetivos traz como característica adicional a diminuição da aleatoriedade na obtenção do balanceamento. Dai, denomina-se tal modelo de *Semi-Deterministic Ensemble Strategy for Imbalanced Data* (SEID).

A forma definida para se levar em conta estes dois objetivos conjuntamente foi por meio de um comitê de classificadores. Um procedimento de comitê apresenta, naturalmente, uma facilidade de implementação dos objetivos para cada classe descrito acima. No caso da necessidade de redundância das instâncias minoritárias tem-se a facilidade de utilização de todas suas instâncias em cada base do comitê. No caso das instâncias majoritárias, onde se pretende particionar ou decompor seus elementos, podem-se colocar parcelas de suas instâncias 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 de dimensão consideradas inadequadas.

A seguir, apresenta-se a estratégia proposta para predição de insolvência de empresas:

Considera-se inicialmente a composição do conjunto de treinamento $Str = Str_m \cup Str_M$, ou seja, formado pela união de instâncias da classe minoritária (Str_m) e da classe majoritária (Str_M) com $\#(Str_M) > \#(Str_m)$, onde $\#(*)$ é a cardinalidade do conjunto. Os conjuntos de treinamento gerados para a obtenção dos classificadores base serão balanceados com n_{ic} instâncias de cada classe. Para que se obtenham conjuntos de treinamentos de acordo com o modelo proposto, adota-se para o valor mínimo de instância por classe n_{ic} :

$$n_{ic} \geq \max(\#(Str_m), \#(Str_M)/n_{cb}) \quad (3.1)$$

Com n_{cb} sendo o número de classificadores base usados no comitê de classificadores e o operador $\max(*)$ assume o maior valor entre os avaliados. Quanto maior o valor de n_{ic} mais próximo o algoritmo se torna do algoritmo de *bagging*. A seguir, apresenta-se o pseudo-código do SEID.

Pseudo-código: comitê de classificadores para base de dados desbalanceadas (SEID)

```

início
  Defina o número de classificadores base  $n_{cb}$ 
  Defina o número de instâncias para cada classe  $n_{ic}$ 
  % construção dos  $n_{cb}$  classificadores base
  para  $i=1, n_{cb}$ 
    % classe minoritária
     $Str_i \leftarrow Str_m$ 
    % completar, quando necessário, aplicando um processo de bootstrap na classe
minoritária
    para  $j = \#(Str_m) + 1, n_{ic}$ 
       $Str_i \leftarrow Str_i \cup j$ -ésima instância obtida aplicando bootstrap na amostra  $Str_m$ 
    fim
    % classe majoritária
    para  $j = 1, \#(Str_M) / n_{cb}$ 
       $Str_i \leftarrow Str_i \cup j$ -ésima instância obtida de  $Str_M$  sem reposição
    fim
    % completar, quando necessário, aplicando um processo de bootstrap na classe
majoritária
    para  $j = \#(Str_M) / n_{cb} + 1, n_{ic}$ 
       $Str_i \leftarrow Str_i \cup j$ -ésima instância obtida aplicando bootstrap na amostra  $Str_M$ 
    fim
  fim
  Treine os  $n_{cb}$  classificadores base
  %classificação de novas instâncias
  Aplique técnica de votação majoritária para classificar os dados de teste
fim.

```

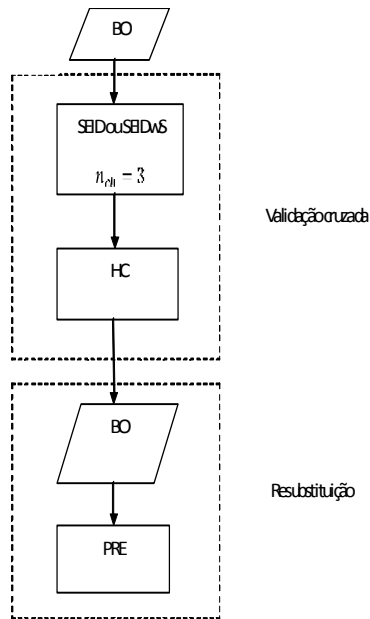


Figura 3.1 - Fluxograma referente aos procedimentos para se chegar aos resultados após os balanceamentos da base de dados original.

3.3 Validação do algoritmo proposto

A validação do algoritmo proposto será realizada em duas etapas visando atender dois objetivos, (i) testar a metodologia proposta nesta pesquisa em base de dados diferente daquelas aqui estudadas; (ii) comparar os resultados gerados pelo SEID com outras pesquisas realizadas nesse tema.

O cumprimento da primeira etapa foi feito testando os algoritmos SEID em três bases de dados originadas do Repositório UCI para Aprendizado de Máquina². Tais bases de dados são utilizadas para testes de estudos sobre modelagem de previsão de insolvência (*Japanese Credit Screening, Australian Credit Approval, German Credit Data*).

3.3.1 Validação do SEID nas bases do UCI Repositório para Aprendizado de Máquina

Nesta subseção são apresentados os resultados da validação do SEID através de três bases do UCI, o procedimento é o mesmo apresentado na Figura 3.1.

Árvore de decisão foi o classificador utilizado (devido aos resultados apresentados na Tabela 3). A Tabela 1 apresenta os resultados dos testes do SEID e do algoritmo SMOTE, neste algoritmo o k (*vizinhos mais próximos*) usado foi igual a 5. Os *softwares* utilizados foram o WEKA 3.5.6 (Witten e Frank, 2005) e o Matlab 7.1. As métricas de avaliação são apresentadas na subseção 3.5.

² <http://archive.ics.uci.edu/ml/>

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

Bases de dados do UCI	Nº de atributos	Classe	Instâncias	Base original		SEID		SMOTE	
				F	AUC	F	AUC	F	AUC
Japanese Credit Screening	15	INS	383	0,382	0,572	0,726	0,88	0,772	0,905
		SOL	307	0,77	0,572	0,896	0,88	0,902	0,905
Australian Credit Approval	14	INS	383	0	0,474	0,74	0,884	0,786	0,926
		SOL	307	0,96	0,474	0,966	0,884	0,984	0,926
German Credit Data	20	INS	300	0,547	0,73	0,881	0,94	0,811	0,933
		SOL	700	0,813	0,73	0,931	0,94	0,927	0,933

Fonte: Elaborada pelo autor.

A Tabela 2 apresenta as comparações dos alguns estudos publicados sobre o tema na literatura específica utilizando como parâmetros acurácia (classificação correta nas duas classes), Erro Tipo I (classifica instância não falidas no grupo das falidas) e Erro Tipo II (classifica instância falidas no grupo das não falidas). As comparações foram feitas através dos melhores resultados encontrados pelos autores. Os estudos utilizados para comparação são de Tsai e Wu (2008), Tsai (2009) e Nanni e Lumini (2009).

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

	SEID	Tsai e Wu	Tsai	Nanni e 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

Fonte: Elaborada pelo autor.

Na tabela 2, os resultados mostram a eficácia do algoritmo SEID. A comparação mostra que SEID obteve melhores resultados na acurácia, nos Erros Tipo I e II, e que em todos esses parâmetros há um ganho do SEID sobre os outros estudos. No Erro Tipo II o SEID obteve melhores resultados sobre os outros testes em dois das três bases de dados. Somente na base *Japanese Credit Screening* do estudo de Nanni e Lumini os resultados ficaram um pouco inferiores, (9,92 X 9,4).

3.4 Base de dados de empresas brasileiras

Foram obtidos 22 indicadores contábeis anuais das empresas classificados de acordo com grupos de índices contábeis-financeiros: liquidez, endividamento e rentabilidade. Estas empresas foram classificadas como concordatária ou falida na BOVESPA durante o período de 1996 a 2006. Para cada empresa classificada como insolvente, 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 do ativo, sempre que possível compatível, e pertencente ao mesmo setor de atividade, buscando respeitar, localização geográfica e idade. 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 erro, objetivando, também ficar mais próximo da realidade econômica. Essa base totalizou 175 empresas, com 147 classificadas como solventes e 28 classificadas como insolventes durante o período em estudo. A reduzida dimensão da amostra se deve principalmente a não obrigatoriedade de um grande número de empresas de publicar seus demonstrativos contábeis.

Foram criadas 1.610 instâncias sendo 1.470 referentes a empresas solventes e 140 empresas insolventes. 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 e (ii) utilizar demonstrativos contábeis sem a influência da inflação.

3.5 Métricas de avaliação

Das métricas alternativas existentes para lidar com o problema do desequilíbrio de classes citadas por Joshi, *et al.*, 2001; Käuck, 2004 e Gary, 2004 foram escolhidas: matriz de confusão (MC), área sob a curva ROC (AUC) e medida F (F). Já para a avaliação do classificador serão utilizadas validação cruzada com 10 sub-amostras e resubstituição (Braganeto, *et al.*, 2004).

4 RESULTADOS

Nos sub-itens seguintes serão apresentados os resultados dos classificadores aqui aplicados na base de dados de empresas brasileiras para determinar o melhor classificador para essa base de dados. A seguir é aplicado o SEID na base dados estudada e comparado os resultados encontrados com aqueles da base original e com a aplicação do SMOTE.

4.1 Aplicação de classificadores na base de dados

As técnicas empregadas para a classificação das empresas são: 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. 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 valor do AUC.

Tabela 3 - Resultados dos classificadores no treinamento da base de dados original

Classe	RL				SVM				MLP				AD			
	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	75	65	0,607	0,907	85	55	0,756	0,804	89	51	0,804	0,757	80	60	0,661	0,935
S	32	1438	0,932	0,907	0	1470	0,982	0,804	4	1466	0,982	0,757	22	1448	0,972	0,935

Fonte: Elaborada pelo autor.

Dos classificadores testados AD e RL foram os que obtiveram os melhores resultados nos testes. O AD teve um desempenho bem melhor do que o RL. Em todos os classificadores, as empresas solventes obtiveram menores erros de classificação. Isto pode ser visto tanto nas matrizes de confusão como nas medidas F. Estes resultados contrariam o objetivo principal

desta classificação que é a de descobrir melhores conhecimentos na classe de empresas insolventes.

4.2 Aplicação das estratégias SEID e comparação dos resultados encontrados

Nesta seção, a estratégia SEID desenvolvida para a predição de insolvências em empresas será aplicada na base de dados. Na prática, a aplicação completa do SEID 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 subamostras, tanto para as sub-bases do SEID quanto para o SMOTE. Agora, com a utilização do SEID completo a validação será feita pelo método da ressubstituição tanto para o SEID como para o SMOTE.

Nos resultados apresentados na Tabela 5.5 as bases de dados que obtiveram os melhores resultados na predição foram os modelos sujeitos ao balanceamento, ou seja, o SEID, com vantagem para o modelo que utiliza a seleção de características. Destaca-se, principalmente nos modelos balanceados um ganho de eficácia na classificação na classe das insolventes (I), atendendo o interesse preponderante desta técnica.

Já na Tabela 4 são comparados os resultados encontrados na base original, sem tratamento e aplicando duas técnicas de balanceamento, SEID e o SMOTE.

Tabela 4 – Comparação da base original, o SEID e o SMOTE.

Classe	BASE ORIGINAL				SEID				SMOTE			
	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	80	60	0,661	0,935	125	15	0,964	0,971	122	18	0,911	0,993
S	22	1448	0,972	0,935	3	1467	0,978	0,971	2	1468	0,991	0,993

Fonte: Elaborada pelo autor.

Pela Tabela 4 pode ser evidenciada a capacidade do SEID em melhor caracterizar aquelas amostra das empresas insolventes, apresenta maior F do que o modelo gerado pelo SMOTE, mostrando assim um resultado mais eficaz para caracterizar as empresas insolventes, mesmo que o valor do AUC e o F das solventes do SMOTE tenham sido superiores aos do SEID.

5 CONCLUSÕES E FUTUROS ESTUDOS

Esta pesquisa apresentou e testou uma estratégia para solucionar um dos problemas existentes na modelagem de previsão de insolvência de empresas pouco estudado, o desequilíbrio entre as classes de empresas classificadas como solventes e as empresas classificadas como insolventes. Em ambientes econômicos normais a classe de empresas classificadas como solventes é bem maior do que a classe de empresas classificadas como insolventes. Na grande maioria das pesquisas existentes a amostra estudada é composta com o número igual de empresas solventes e insolventes. Tal paridade entre as classes de empresas pouco representa a realidade do ambiente econômico dessas empresas distorcendo, assim a amostra a ser estudada e priorizando a classificação das empresas solventes. Nesta pesquisa buscou-se, através da estratégia apresentada, adequar a base de dados com o ambiente econômico dessas empresas.

Foi elaborada uma base de dados original com índices calculados a partir de demonstrativos contábeis de empresas classificadas como solventes e insolventes pelo BOVESPA observando seus setores econômicos no período de 1996 e 2006. Nessa base de dados original foi aplicada a estratégia aqui apresentada – SEID – que gera três sub-bases 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 que posteriormente foi testado na base de dados original gerando, assim três resultados referentes a essas classificações. Na etapa seguinte foi realizada a votação majoritária dos resultados encontrados das três classificações com a base original, obtendo assim o resultado final das classificações.

Foram feitas validações utilizando dados públicos de base de dados com fins de se testar novas técnicas de modelagem e muito utilizada na literatura específica. Tais bases se referem a testes com fins de modelagem de previsão de insolvência e neste estudo foram utilizadas três: *japanese credit*, *australian credit* e *german credit*, conforme demonstrado na Tabela 1. Na outra validação são comparados os resultados obtidos pela estratégia SEID com outros estudos contemporâneos publicados na literatura específica. Para finalizar a etapa de validação são gerados resultados utilizando o SEID e outra estratégia bem referenciada na literatura específica – o SMOTE em uma base de dados obtida na BOVESPA.

Nas validações realizadas os resultados obtidos revelam que a estratégia apresentada – o SEID – é bem competitiva com o SMOTE, tendo capacidade de classificar aquelas empresas pertencente à classe das insolventes e com resultados ainda melhores na base de dados onde o desequilíbrio entre as classes é mais acentuado (*german credit*), apresentado na Tabela 1, representando melhor o ambiente econômico dessas empresas.

Na comparação feita dos resultados encontrados aplicando o SEID com outras estratégias publicadas na literatura específica os resultados são bem animadores, na Tabela 2, pode ser evidenciado a capacidade do SEID em relação a outros trabalhos contemporâneos. Na grande maioria dos resultados obtidos o SEID foi mais eficiente do que as outras estratégias, há ganhos na capacidade de classificar aquelas empresas pertencente à classe das insolventes.

Quando da aplicação da estratégia SEID e do SMOTE na base de dados de empresas brasileiras os resultados são ainda mais convincentes na diferença da capacidade de classificar as empresas insolventes. O SEID obteve melhor capacidade de classificar aquelas empresas pertencente à classe das insolventes do que o SMOTE, mostrado na Tabela 4. Nesta Tabela é evidenciada a importância do tratamento da base dados para equacionar o problema do desequilíbrio das classes e melhorar a capacidade do modelo de previsão na classificação das empresas pertencentes à classe das insolventes.

Este estudo ilustra a importância do desenvolvimento de estratégias para equacionar o problema existente na modelagem de previsão de insolvência de empresas que é o desequilíbrio de classes, além disso, o estudo apresenta uma estratégia no qual tem plenas condições de competir com outra estratégia (SMOTE), elaborada e citada na literatura específica que ataca o problema do desequilíbrio entre as classes. Cabe ressaltar que o SEID minoria o efeito estocástico do modelo em relação ao SMOTE. Diante disso pode ser visto como uma contribuição às pesquisas sobre previsão de insolvência de empresas tema pouco explorado na literatura no Brasil.

Uma possível extensão ao presente estudo seria a inclusão de novas técnicas de comitês de classificações, a inclusão de variáveis qualitativas na base de dados, o que possivelmente melhoraria a capacidade preditiva do modelo.

REFERENCIAS BIBLIOGRÁFICAS

- 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 Manoel 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.
- 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.
- BRAGANÇA, Luiz Augusto de; BRAGANÇA, Sérgio Luiz de. “Rating” previsão de concordatas e falências no Brasil”. *VII Congresso ABAMEC/1984*.
- BRAGA-NETO, U.; HASHIMOTO, R.; DOUGHERT, Edward R. Nguyen, DANH V.; CARROLL, Rymond J. “Is cross-validation better than resubstitution for ranking genes?” Vol. 20 nº 2, 2004, p. 253-258. DOI: 10.1093/bioinformatics/btg399.
- CANBAS S, A.; CABUK, S.B.; KILIC, “Prediction of commercial bank failure via multivariate statistical analysis of financial structure: The Turkish case”, *European Journal of Operational Research*, v. 166, 2005, p. 528–546.
- CHAUDHURI, Arindam; KAJAL De. “Fuzzy Support Vector Machine for Bankruptcy Prediction”. *Applied Soft Computing Journal* (2010), doi:10.1016/j.asoc.2010.10.003.
- 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.
- CHEN, Hsueh-Ju; HUAND, Shaio Yan; LIN, Chi-Shie. “Alternative diagnosis of corporate bankruptcy: A neuro fuzzy approach”. *Expert Systems with Applications*, v. 36, Issue 4, may. 2009, p. 7710-7720.
- CHEN, Wun-Hwa; SHIH, Jen-Ying. “A study of Taiwan's issuer credit rating systems using support vector machines”. *Expert Systems with Applications*, v. 30, Issue 3, april 2006, p. 427-435.
- CHEN,Wei-Sen; DU, Yin-Kuan. “Using neural networks and data mining techniques for the financial distress prediction model”. *Expert Systems with Applications*, v. 36, Issue 2, Part 2, march 2009, p. 4075-4086.
- CHYE, Koh Hian; CHIN Tan We. “Credit scoring using data mining techniques”. *Singapore Management Review*, v. 26, Nº 2, july 2004.

- CORNETT, M. M. ;MARCUS, A. J.;SAUNDERS, A.;TEHRANIAN, H., “The impact of institutional ownership on corporate operating performance”. *Journal of Bank & Finance*, Volume 31, Issue 6, 2007, p. 1771-1794.
- DIMITRAS A. I.;SLOWINSKI R.; SUSMAGA, R. Zopounidis, C. “Business failure prediction using rough sets”. *European Journal of Operational Research*, v. 114, Issue 2, 16 april 1999, p. 263-280.
- DING, Yongsheng; SONG, Xinping, ZEN, Yueming. “Forecasting financial condition of Chinese listed companies based on support vector machine”. *Expert Systems with Applications*, Volume 34, Issue 4, May 2008, p. 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, p.35-41.
- GESTEL, Tony Van; BAESENS, Bart; MARTENS, David., “From linear to non-linear kernel based classifiers for bankruptcy prediction”. *Neurocomputing* 73 (2010), p. 2955–2970.
- HORTA, Rui Américo Mathiasi. *Utilização de indicadores contábeis na previsão de insolvência: Análise empírica de uma amostra de empresas comerciais e industriais brasileiras*. Dissertação apresentada no programa de mestrado em Ciência Contábeis da Universidade Estadual do Rio de Janeiro em 2001.
- 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, aug. 2007, p. 434-440.
- HUANG., C. L., CHEN M. C., WANG., C.J. “Credit scoring with a data mining approach based on support vector machines”. *Expert Systems with Applications*, v. 33, Issue 4, nov. 2007, p. 847-856.
- HUNG, Chihli; CHEN, Jing-Hong. “A selective ensemble based on expected probabilities for bankruptcy prediction”. *Expert systems with applications*, 2009, v. 36, Issue 3, apr. 2009, p. 3297-5309.
- JAPKOWICZ N.; STEPHEN, S., “The Class Imbalance Problem: A Systematic Study”. *Intelligent Data Analysis*, v. 6, Number 5, p. 429-450, nov. 2002.
- JONES S.; HENSHER, D.A. “Predicting firm financial distress: A mixed logit model”, *Accounting Review*, v. 79, Issue 4, 2004, p. 1011–1038.
- JOSHI, M. V. *Learning Classier Models for Predicting Rare Phonemena*. PhD thesis, University of Minnesota, Twin Cities, Minnesota, USA, 2002.
- KANITZ, Stephen Charles. *Como prever falências*. São Paulo: Mc Graw-Hill do Brasil, 1978.
- 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*, In Press, Corrected Proof, Available online 1 april 2009.
- LACHERR. C.; COATS, P. K.; SHARMA, S.C.; FANT, L. F. “A neural network for classifying the financial health of a firm”, *European Journal of Operations Research*, v. 85, 1995, p. 53-65.
- LEE, Tian-Shyug; CHENG, I-Fei. “A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines”. *Expert Systems with Applications*, v. 28, Issue 4, may. 2005, p. 743-752.

- 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, Jae H. Lee; YOUNG-CHAN. “Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters”. *Expert Systems with Applications*, v. 28, Issue 4, may. 2005, p. 603-614.
- 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 Vol. 2 Número 41, 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.
- NWOGUGU, Michel. “Decision-making, risk and corporate governance: New dynamic models/algorithms and optimization for bankruptcy decisions”. *Applied Mathematics and Computation*, v. 179, Issue 1, 1 aug. 2006, p. 386-401.
- ODOM, M.; SHARDA, R. “A neural network model for bankruptcy prediction”. In: *Proceedings of the international joint conference on neural networks*, Vol. 2, IEEE Press, Alamos, CA, 1990, p. 163–168.
- OHLSON, J.A. “Financial ratios and the probabilistic prediction of bankruptcy”. *Journal of Accounting Research*, v. 18, 1980, p.109-131.
- PARK, Cheol-Soo; HAN, Ingoo. “A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction”. *Expert Systems with Applications*, v.23, Issue 3, oct. 2002, p. 255-264.
- PREMACHANDRA, I.M.; BHABRA, Gurmeet Singh, Sueyoshi Toshiyuki. “DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique”. *European Journal of Operational Research*, v. 193, Issue 2, 1 mar. 2009, p. 412-424.
- RAVI, V.; KURNIAWAN, H.; THAI, Peter Nwee Kok.; KUMAR, P. Ravi. “Soft computing system for bank performance prediction”. *Applied Soft Computing*, v. 8, jan. 2008, p.305-315.
- ROY, J.; COSSET, C. “The determinants of country risk ratings”, *Journal of International Business studies*, First Quarter (1990), p. 135–139.
- SANTOS, Samuel Cruz Dos. *Um modelo de análise discriminante múltipla para previsão de inadimplência em empresa*. Dissertação de Mestrado, Departamento de Administração, PUC/RJ. Rio de Janeiro: PUC, 1996.
- SANVICENTE, Antônio Zoratto.; MINARDI, Andréa Maria A. F. *Identificação de indicadores contábeis significativos para previsão de concordata de empresas*. Disponível: <http://www.risktech.br/artigos/artigos_técnicos/index.html>. Acesso em: 23/10/2005.
- SAUNDERS, A. ALLEN; DELONG, G. L. “Issues in the credit risk modeling of retail markets”. *Journal of Bank & Finance*, v. 28, 2004.

- SAUNDERS, A.; CORNETT M. M. “Financial Institutions Management: A Risk Management Approach”. *Paperback edition*. McGraw-Hill/Irwin, 2008.
- SAUNDERS, ANTHONY. *Medindo o risco de crédito*. Rio de Janeiro: Qualitymark Editora, 2000.
- SHARDA R.; WILSON, R. L. “Neural network experiments in business-failure forecasting: Predictive performance measurement issues, International”. *Journal of Computational Intelligence and Organizations*, v. 1, Issue 2, 1996, p. 107-117.
- SHIN, Kyung-Shik; LEE, Yong-Joo; KIM, Hyun-Jung. “An application of support vector machines in bankruptcy prediction model”. *Expert Systems with Applications*, v. 28, Issue 1, jan. 2005, p. 127-135.
- 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, Vol. 20 Número 51, p. 28-43, Setembro/Dezembro, 2009.
- SILVA, José Pereira da. *Gestão e análise de risco de crédito*. 5ª Ed. São Paulo: Atlas, 2006.
- TAM K. Y.; KIANG, M. Y. “Managerial applications of neural networks: The case of bank failure predictions”, *Management Science*, v. 38, Issue 7, 1992, p. 926-947.
- TAY Francis E. H.; SHEN, Lixiang. “Economic and financial prediction using rough sets model”, *European Journal of Operational Research*, v. 141, Issue 3, sep. 2002, p. 641-659.
- TSAI, Chih-Fong. “Feature selection in bankruptcy prediction”. *Knowledge-Based Systems, Volume 22, Issue 2, March 2009, Pages 120-127*.
- 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.
- VERIKAS, Antanas; KALSYTE, Zivile; BACAUSKIENE, Marija; GELZINIS, Adas. “Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey”. *Soft Comput* v.14, 2010. p. 995–1010.
- WEST, David; DELLANA, Scott; QIAN, Jingxia. “Neural network ensemble strategies for financial decision applications”. *Computers & Operations Research*, Volume 32, Issue 10, October 2005, Pages 2543-2559.
- WILSON, R. L.; SHARDA, R. “Bankruptcy prediction using neural networks”, *Decision Support Systems*, v. 11, jun. 1994, p. 545-557.
- WITTEN, Ian .H.; FRANK, Eibe. *Data Mining: Practical Machine Learning Tools and Techniques*. The Morgan Kaufmann Series in Data Management Systems, 2ª ed. 2005.
- YOUN, Hyewon; ZHENG GUYU, L.; WAUNG, S.; LAI, K. K. “Predicting Korean lodging firm failures: An artificial neural network model along with a logistic regression model”. *International Journal of Hospitality Management*, Available online 26, jul. 2009.
- YU, L. WAUNG; LAI, K. K. “Credit risk assessment with a multistage neural network ensemble learning approach”. *Expert Systems with Applications*, v. 34, fev. 2008, p. 1434-1444.
- ZHANG, Guoqiang Zhang; MICHAEL Y. HU; EDDY, Patuwo W.; DANIEL C. Indro. “Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis”. *European Journal of Operational Research*, v. 116, jul. 1999, p. 16-32.