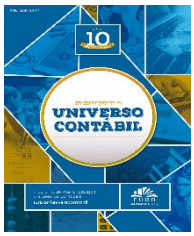

Revista Universo Contábil, ISSN 1809-3337
Blumenau, v. 11, n. 3, p. 43-62, jul./set., 2015

doi:10.4270/ruc.2015321
Disponível em www.furb.br/universocontabil

PREVISÃO DE INSOLVÊNCIA NO SETOR DE MATERIAIS BÁSICOS APLICANDO MINERAÇÃO DE DADOS[1]

FORECAST OF INSOLVENCY IN THE BASIC MATERIALS SECTOR APPLYING DATA MINING

PREDICCIÓN DE INSOLVENCIA EN EL SECTOR DE MATERIALES BÁSICOS DE APLICACIÓN DE LA MINERÍA DE DATOS

Rui Américo Mathiasi Horta
Doutor em Engenharia Civil (COPPE/UFRJ)
Professor Adjunto do Depto de Finanças e Controladoria da UFJF
Universidade Federal de Juiz de Fora
Endereço: Rua José Lourenço Kelmer, s/n
Campus Universitário - Bairro São Pedro
CEP: 36.036-900 - Juiz de Fora – MG
E-mail: rui.horta@ufjf.edu.br
Telefone: (32) 2102 - 3521

Carlos Cristiano Hasenclever Borges
Doutor em Engenharia Civil (COPPE/UFRJ)
Professor Adjunto do Depto de Ciências da Computação da UFJF
Universidade Federal de Juiz de Fora
Endereço: Rua José Lourenço Kelmer, s/n - Campus Universitário - Bairro São Pedro
CEP: 36.036-900 - Juiz de Fora – MG
E-mail: cchb@lncc.br
Telefone: (32) 2102-3327 Ramal 4024

Francisco José dos Santos Alves
Doutor em Ciências Contábeis (USP)
Professor Adjunto do Depto de Finanças da UERJ
Endereço: Rua São Francisco Xavier, 524 – Maracanã – 9º Andar
CEP: 20.550-900 - Rio de Janeiro – RJ
E-mail: fjalves@globo.com
Telefone: (21) 2334-0294

RESUMO

Este estudo tem como objetivo selecionar variáveis em base de dados de empresas do setor de materiais básicos aplicando técnicas de mineração de dados em problemas de previsão de

[1]Artigo recebido em 29.10.2014. Revisado por pares em 10.07.2015. Reformulado em 04.11.2015. Recomendado para publicação em 10.11.2015 por Carlos Eduardo Facin Lavarda. Publicado em 27.11.2015. Organização responsável pelo periódico: FURB.

insolvência utilizando técnicas de balanceamento da base de dados com (etapa de) seleção de atributos. A partir dessas variáveis selecionadas visa-se obter as implicações contábeis que expliquem sobre a descontinuidade dessas empresas. Esta pesquisa é de natureza aplicada com abordagem quantitativa, quanto ao objetivo, é descritiva. A base de dados utilizada foi originada de demonstrativos contábeis de empresas listadas na BM&FBOVESPA entre os anos de 1996 e 2012. Esse setor foi escolhido devido a sua relevância para a economia brasileira em termos de competitividade e faturamento. As variáveis selecionadas foram: EOCpOT, EOAT, GAF, MB, EBITDA, MO e TERFIN. Os resultados obtidos mostraram que as empresas deste setor se tornam insolventes não somente porque perdem a capacidade (financeira) de se endividarem, mas também porque perdem a capacidade operacional de gerar caixa.

Palavras-chave: Previsão de insolvência. Seleção de variáveis contábeis. Mineração de dados. Setor de materiais básicos – Brasil.

ABSTRACT

This study aims to select variables in the sample of companies in the basic materials sector by applying data mining techniques in insolvency forecasting problems using database balancing techniques and attributes selection selection of attributes. From these variables an analysys for the financial implications to explain about the discontinuity of these companies is determined. This is an applied research research with quantitative approach; the aims, is descriptive. The database used was derived from financial statements of companies listed on BM&FBOVESPA between 1996 and 2012. This sector was chosen for their relevance to the Brazilian economy in terms of competitiveness and billing. The selected variables were: EOCpOT, EOAT, GAF, MB, EBITDA, MO and TERFIN. The results showed that companies in this sector become insolvent not only because they lose the ability (financial) to borrow, but also because they lose operational ability to generate cash.

Keywords: *Prediction of insolvency. Selection of accounting variables. Data mining. Sector of basic material - Brazil.*

RESUMEN

Este estudio tiene como objetivo seleccionar variables en la muestra de empresas en el sector de materiales básicos mediante la aplicación de técnicas de minería de datos en problemas de predicción de insolvencia utilizando técnicas de balanceo de la base de datos (paso) de selección de atributos. A partir de estas variables reciben las consecuencias financieras para explicar sobre la discontinuidad de estas empresas. Esta investigación es de carácter aplicado con enfoque cuantitativo; los objetivos, es descriptivo. La base de datos utilizada fue derivado de estados financieros de las empresas que cotizan en BOVESPA entre 1996 y 2012. Este sector fue escogido por su importancia para la economía brasileña en términos de competitividad y de facturación. Las variables seleccionadas fueron: EOCpOT, EOAT, GAF, MB, EBITDA, MO y Terfin. Los resultados mostraron que las empresas de este sector se vuelven insolventes no sólo porque pierden la capacidad (financiera) para pedir prestado, sino también porque pierden capacidad operativa para generar efectivo.

Palabras clave: *Predicción de la insolvencia. La selección de las variables contables. Minería de datos. Sector de materiales básicos - Brasil.*

1 INTRODUÇÃO

Previsão de insolvência é um tema sempre atual e de grande relevância sempre estimulando reflexões e inovações. Além disso, vem adquirindo mais importância nas áreas acadêmicas e empresariais relativas à Contabilidade e Finanças. De fato, a previsão de insolvência no ambiente econômico contemporâneo é primordial por dois motivos: a) permite à empresa antecipar uma situação financeira difícil, de forma que haja tempo hábil para serem adotadas medidas capazes de reverter a situação impedindo a ocorrência de grandes custos sociais e financeiros, consequências de uma descontinuidade; b) a grande maioria das entidades dependem de recursos provenientes de terceiros para operar e incrementar suas atividades. Expectativas de descontinuidade, muitas vezes infundadas, leva ao aumento da taxa de risco em empréstimos bancários.

Razões não faltam para justificar o aumento qualitativo e quantitativo dos estudos sobre o tema. Por exemplo, em vários países a maioria das estatísticas sobre falências mostraram significativo crescimento. Além disso, nas últimas décadas o ambiente econômico geral das empresas, na grande maioria dos países, tem mudado com enorme velocidade e experimentado tendências adversas como aumento do custo do capital, acirramento da concorrência, mudanças de padrões de consumo, etc. Houve um aumento na disponibilização de recursos financeiros para o crédito nos últimos anos em vários mercados, sobretudo no brasileiro. Cresceu, também, a cautela associada à implementação, em vários países, de normas internacionais de contabilidade e finanças (IFRS), tais como Basiléia II e III, Solvência II e Sarbanes-Oxley.

Como sempre ocorre, apesar das inúmeras pesquisas na área, há ainda questões pouco exploradas como a não estacionariedade e instabilidade dos dados, a seleção da amostra e dos atributos e o desequilíbrio entre a quantidade de dados de empresas solventes e insolventes (BALCAEN; OOGHE, 2006; VERIKAS et al.,2010; SUN et al., 2014).

Uma dessas questões pouco exploradas é o problema do desequilíbrio de tamanho das classes inicialmente disponíveis quando se olha a separação entre empresas solventes e insolventes. Sun et al., (2014, p. 53) afirma que o problema do mundo real da previsão de descontinuidade de empresas consiste em lidar com conjuntos de dados desbalanceados. Cabe reconhecer, imediatamente, que esta situação é natural, porque, em qualquer sociedade, o número de insolventes é muito inferior ao de solventes, independentemente do período que se analisa. Por isso mesmo é, também, muito frequente e, nessa medida, é de interesse a utilização de um tratamento analítico adequado para evitar que "os modelos de predição sejam pouco efetivos, predizendo bem somente o que ocorre com a classe majoritária" (JAPKOWICZ; STEPHEN, 2002, p. 431). No caso da dicotomia "solvente / insolvente", ressalta-se que, a classe minoritária é exatamente a que demanda mais atenção, com necessidade de atenção mais precisa.

São vários os especialistas em contabilidade e finanças que discutem o que vem a ser efetivamente insolvência de empresas, havendo definições bem distintas para entender este conceito. Entretanto, em um aspecto a grande maioria concorda, a insolvência está relacionada à existência de um potencial risco na continuidade da empresa. Na visão de Ross *et al.*, 2013 insolvência é a dificuldade financeira máxima que uma empresa pode estar sujeita. Nesta situação a propriedade dos ativos da empresa é legalmente transferida dos acionistas aos credores. Já para Damodaran, 2002 a insolvência ocorre quando os fluxos de caixa de uma empresa venham a se tornar insuficientes para atender aos seus compromissos de dívida (juros e principal). Em Lopes e Martins, 2007 a insolvência está mais relacionada com a incapacidade da empresa de se endividar do que com o seu desempenho operacional. Para Matarazzo, 2003 a insolvência de uma empresa ocorre pela incapacidade de solver suas obrigações, ou seja, pela falta de dinheiro no momento de vencimento de uma dívida.

Neste artigo duas são as questões-chaves o que o diferencia dos demais estudos. Primeiramente, utiliza somente variáveis contábeis obtidos em demonstrativos contábeis de empresas do setor de material básico e em segundo lugar, utiliza técnicas de balanceamento nesta base de dados. Até agora não se conhece nenhum estudo sobre o previsão de insolvência de empresas brasileiras do setor de material básico. O objetivo do presente artigo é preencher esse vazio na literatura sobre previsão de insolvência de empresas brasileiras do setor de materiais básicos.

Este artigo dá continuidade a um conjunto de estudos e publicações referentes a esta metodologia (SEIDwS) aplicada em dados contábeis de empresas brasileiras listadas na BM&FBOVESPA pertencentes a alguns setores econômicos. Os setores já estudados e geraram publicações são os de consumo cíclico (HORTA et al., 2014) e consumo não cíclico (HORTA et al., 2013).

O artigo está organizado em cinco seções, incluindo esta Introdução. A seção 2 apresenta a revisão bibliográfica que dará suporte ao desenvolvimento da pesquisa. Na terceira seção descrevem-se os procedimentos metodológicos adotados. Na seção 4, apresentam-se os resultados obtidos. Na quinta e última seção são expostas algumas conclusões da pesquisa e sugeridos futuros estudos.

2 REVISÃO DA LITERATURA

A previsão de insolvência tornou-se assunto mais pesquisado e difundido na década de 60, notadamente através do modelo chamado Escore-Z (ALTMAN, 1968). O mesmo Altman em conjunto com Haldeman e Narayanan (1977) desenvolveram 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 empresas. West (1985) utilizou análise fatorial para selecionar e especificar as variáveis empregadas. Canbas, Cabuk e Kilic (2005) combinaram Análise Discriminante Linear (ADL), regressão logística (RL), *probit* e análise de componentes principais em sua modelagem da insolvência.

Mais recentemente, Shin, Lee e Kim (2005) investigaram a eficácia da aplicação de SVM (Máquinas de Vetor Suporte) para o problema de previsão de falências, mostrando que o classificador SVM, para os testes deles, supera MLP (redes neurais *multilayerpercption*) em problemas de previsão de falências de empresas. Min, Lee e Han (2006) propuseram métodos para melhorar o desempenho de SVM em dois aspectos: a seleção de atributos e a otimização de parâmetros. Para Hua et al. (2007), que o aplicaram ao problema de previsão de falências, o classificador provou ser superior aos métodos concorrentes, tais como as MLP, as múltiplas abordagens de ADL e a RL. Ding, Song e Zen (2008) desenvolveram um modelo de previsão de insolvência utilizando SVM para empresas chinesas de alta tecnologia. Kim e Sohn (2009) elaboraram um modelo SVM para prever insolvência em pequenas e médias empresas sulcoreanas no setor de tecnologia.

Alguns autores, além de classificar, também desenvolveram metodologias que são aplicadas na construção do preditor ou na base de dados estudada, visando obter melhores resultados nas classificações. Por exemplo, Atiya (2001) desenvolveu um estudo sobre previsão de insolvência em que aplica MLP em um caso de bancos de dados desbalanceados obtendo resultados motivadores. Em busca de maior precisão nas previsões, West, Dellana e Qian (2005) investigaram três estratégias de combinação de classificadores para aplicação a decisões financeiras, incluindo previsão de insolvência. Hung e Chen (2009) aplicaram um modelo de probabilidade híbrida, baseado em comitê de classificadores, para previsão de insolvência utilizando votação majoritária e votação ponderada. Tsai e Wu (2008) compararam o desempenho de um classificador simples de MLP com o de múltiplos classificadores, também

baseados em MPL. Fazendo aplicação de comitês de classificadores, Yu e Lay, (2008) utilizam MLP para avaliar o risco de crédito. Ravi et al. (2008) elaboraram e testaram modelos utilizando 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 italianas.

Wu, Gaunt e Gray (2010) testaram cinco modelos bem conhecidos na literatura internacional sobre falência. Descobriram que cada um dos modelos contém informações exclusivas sobre a probabilidade de falência, mas o seu desempenho varia com o tempo estudado. Em decorrencia, construíram um novo modelo compreendendo as variáveis-chave de cada um dos cinco modelos estudados anteriormente, e adicionaram uma nova variável no modelo representando o grau de diversificação dentro da empresa. O novo modelo superou os modelos existentes em uma variedade de testes. Olson, Delen e Meng (2012) aplicaram uma variedade de ferramentas de mineração de dados para bases de falência, com o objetivo de comparar a precisão e o número de regras. Por esses dados, árvores de decisão foram relativamente mais preciso em relação a redes neurais e máquinas de vetor suporte.

Tinoco e Wilson (2013) investigaram empiricamente a utilidade de combinar dados contábeis, dados de mercado, e macro-econômicos para explicar o risco de crédito corporativo. Os resultados mostram a utilidade da combinação desses dados em modelos de previsão de falência. Tsai, Hsu e Yen (2014) compararam um conjunto de classificadores (*ensemble*) utilizando três técnicas de classificação amplamente utilizados, MLP, máquinas de vetor de suporte (SVM), e árvores de decisão (AD) baseado em dois métodos de combinação bem conhecidos, incluindo *bagging* e *bosting* e diferentes números de classificadores combinados. Os resultados, utilizando três conjuntos de dados públicos mostram que o conjunto AD utilizando o método de *boosting* apresentou um melhor desempenho. O teste de Wilcoxon também demonstrou que AD *ensembles* por *boosting* teve um desempenho significativamente diferente dos outros classificadores ensembles. Sun, Li e Huang (2014) fizeram uma revisão sobre recentes estudos sobre previsão de insolvência discutindo abordagens sobre definições, modelos, amostragem.

No Brasil ainda é notória a escassez de pesquisas desenvolvidas com o propósito de encontrar parâmetros para previsão de insolvência, além da persistente escassez de dados adequados e confiáveis para a realização desse tipo de estudos. Essa situação começa a mudar, mas, em comparação à facilidade de obtenção de dados que ocorre em outros países, ainda se está bem longe de poder desenvolver tais estudos com fluidez. A seguir são revistos alguns trabalhos que acabaram conquistando destaque no Brasil.

Tidos como destacados precursores, Elizabetsky (1976), Kanitz (1978) e Matias (1978) trabalharam em modelos de previsão de insolvência utilizando análise discriminante. A metodologia dos trabalhos seguintes de Altman, Baidya e Dias (1979) ou Pereira (1982) - também recorreu à ferramenta estatística de análise discriminante, assim como Sanvicente e Minardi (1998). Horta (2001) elaborou modelos de previsão de insolvência em que aplicou as técnicas estatísticas de análise discriminante e regressão logística na etapa de seleção de atributos utilizando dados obtidos em demonstrativos contábeis de empresas brasileiras evidenciando a capacidade desses dados para previsão de insolvência. Morozini, Olinquevitch e Hein (2006) utilizam análise dos componentes principais para combinar os principais índices dentre os selecionados para o estudo. Silva Brito, Assaf Neto e Corrar, (2009) utilizaram regressão logística para examinar se eventos de *default* de empresas abertas no Brasil podem ser adequadamente previstos por um sistema de classificação de risco de crédito baseado em índices contábeis. Horta (2010), utilizando dados contábeis de empresas brasileiras, num pioneirismo, propõe uma metodologia que ataca e resolve o problema do desbalanceamento entre as classes de empresas solventes e insolventes existente em estudos de previsão de insolvência - *Semi-Determinístic Ensemble Strategy for Imbalanced Data* with attribute Selection (SEIDwS).

3 METODOLOGIA DA PESQUISA

Este estudo tem como objetivo investigar a previsão de insolvência em empresas brasileiras do setor de materiais básicos, levando em consideração o desbalanceamento causado pelo número de empresas insolventes em relação ao quantitativo de empresas solventes deste segmento.

Para atender esse objetivo, foram selecionadas variáveis a partir de uma amostra de empresas do setor de material básico, e aplicadas técnicas de mineração de dados para previsão de insolvência, em especial, as técnicas de balanceamento de bases de dados com com posterior seleção de atributos. A partir dessas variáveis contábeis, buscou-se analisar as implicações dos achados da pesquisa sobre a descontinuidade dessas empresas. Na seção 3.1 os procedimentos passam a ser descritos em detalhes.

Esta pesquisa é de natureza aplicada com abordagem quantitativa; quanto ao objetivo, é descritiva. A metodologia desta pesquisa é a mesma aplicada nos estudos feitos nas bases de dados de empresas de outros setores econômicos (consumo cíclico e consumo não cíclico) citados em parágrafos anteriores. Vejamos, agora, os passos metodológicos cumpridos para alcançar o objetivo.

3.1 Base de dados e métricas de avaliação

A partir dos demonstrativos contábeis publicados no BM&FBOVESPA, foram obtidos 22 indicadores de liquidez, endividamento e rentabilidade (ver os Anexos) de empresas do setor de materiais básico. Esses índices econômico-financeiros para Silva (2006, p. 190), "têm por objetivo fornecer-nos informações que não são fáceis de serem visualizadas de forma direta nas demonstrações financeiras".

Cada uma destas empresas havia sido classificada como concordatária ou falida na BM&FBOVESPA, durante o período de 1996 a 2012. Para cada empresa classificada como insolvente foi adicionada uma quantidade superior de empresas de capital aberto, com controle privado nacional, financeiramente saudáveis (no sentido de que não hávia solicitação de concordata por parte da empresa no período considerado). O estabelecimento de uma quantidade superior de empresas adimplentes para cada inadimplente, ademais, baseia-se na hipótese de que quanto maior a quantidade de dados existentes, menor a probabilidade de erros de classificação, além de representar melhor a realidade econômica.

A base das empresas insolventes foi composta com dados referentes aos demonstrativos contábeis dos cinco anos anteriores ao ano em que a empresa foi declarada insolvente. De acordo com Altman, Giancarlo e Franco (1994) e com Hung e Chen (2009), as empresas insolventes começam a apresentar características ou indícios de insolvência cerca de cinco anos anteriores ao ano em que ocorre efetivamente a falha. No grupo das insolventes há 6 empresas totalizando 30 instâncias (empresas x anos dos demonstrativos).

O grupo das solventes há 32 empresas totalizando 351 instâncias, na proporção de quase 12 para 1. Os dados sobre empresas solventes se referem aos demonstrativos contábeis de até dez anos, facilitando assim uma melhor caracterização dessas empresas. Adotou-se, também, (i) a presença na amostra de dados contábeis do ano 2005 em que ocorreu a mudança na Lei de Falências, (ii) utilizar demonstrativos contábeis sem a influência da inflação e (iii) ter uma base de dados de um período de tempo com o ambiente econômico de muitas mudanças e transformações para essas empresas brasileiras.

Para se chegar a essa proporção adotou-se a seguinte estratégia: primeiro foram obtidos um maior número possível de empresas, através de seus demonstrativos contábeis, classificadas como insolventes, a seguir foram obtidos o maior número possível de empresas pertencentes ao setor econômico de materiais básicos. Com isso buscou-se ajustar a base de dados ao ambiente econômico no qual ocorreram as insolvências, ou seja, a quantidade de empresas que

apresentam problemas na sua saúde financeira é bem inferior àquelas de boa saúde financeira.

A reduzida dimensão da amostra final, se comparada com estudos publicados na literatura internacional se deve, principalmente, à não obrigatoriedade, para um grande número de empresas brasileiras, de publicar demonstrativos contábeis.

A desproporção ou o desbalanceamento apresentado entre o número de empresas classificadas como insolventes e solventes caracteriza um problema que requer o uso de uma técnica de balanceamento de amostras e no qual, neste estudo, se utilizará o SEIDwS.

Neste setor econômico, as empresas apresentam, normalmente, valores proporcionalmente altos em seus ativos não circulantes. O setor econômico de materiais básicos, segundo a classificação da Bovespa, era composto, em 2012, por empresas, subdivididas em 6 subsetores e 11 segmentos. A maioria das empresas analisadas se concentra nos subsetores de Siderurgia e Metalurgia e Químico; a menor concentração de empresas está no subsetor de Materiais Diversos. Destaca-se a concentração das empresas nos segmentos de Petroquímicos, Artefatos de Ferro e Aço e Siderurgia, sendo os segmentos de Químicos Diversos, Madeira e Fertilizantes e Defensivos os de menor participação relativa no setor de Materiais Básicos.

Das métricas de avaliação alternativas existentes para lidar com o problema do desequilíbrio de classes citadas por Joshi et al. (2001); Käuck (2004) e Gary (2004) foram escolhidas três: Matriz de Confusão (MC), Medida F e Área sob a curva ROC (AUC). A Matriz de Confusão é uma tabela em que são representados os TP (verdadeiros positivos), TN (verdadeiros negativos), FP (falsos positivos), FN (falso negativos), e que permite calcular as percentagens de classificações corretas e incorretas. A Medida F é a razão média entre precisão e *recall*, e mede a capacidade de reconhecer os exemplos negativos e positivos (Witten; Frank, 2011, p. 175). Por definição, uma curva ROC é um gráfico bidimensional em que o eixo horizontal representa a taxa de erro da classe negativa (1-*Spec*) e no eixo vertical os valores de sensitividade. (HAN; KAMBER; PEI, 2011, p. 372).

Para a avaliação do classificador serão utilizadas validação cruzada, a validação cruzada requer que os dados originais na base de dados sejam utilizados para treinamento e teste, neste trabalho foram adotados 10 subconjuntos para aprendizagem, e resubstituição. Na validação por resubstituição ocorre a construção da hipótese de classificação com todos os dados para em seguida aplicar esta mesma hipótese de classificação por sua vez em cada uma das observações. (BRAGA-NETO et al. 2004). Também será utilizada a técnica da votação majoritária para refinar os resultados. A técnica da votação majoritária é um método simples e eficaz de combinação. Ela escolhe o rótulo de classe que é apoiada pela maioria dos múltiplos classificadores (LI HUI; JIE SUN, 2009).

3.2 Técnicas de tratamento de bases desbalanceados

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

Uma forma de diminuir o problema de classes desbalanceadas em uma base de dados é equilibrar artificialmente a distribuição das classes no conjunto de exemplos. Duas abordagens principais são utilizadas neste estudo: (a) remoção de exemplos da classe majoritária - u*nder-sampling* e (b) inclusão de exemplos da classe minoritária *over-sampling*.

Alguns trabalhos recentes têm buscado superar as limitações existentes, tanto nos métodos de *under-sampling*, quanto nos de *over-sampling*. Por exemplo, Chawla et al. (2002) combinam métodos de *under* e *over-sampling* através do algoritmo chamado SMOTE (*Synthetic Minority Over-sampling Technique*), em que a etapa 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, uma técnica bastante citada na literatura específica, servirá como comparativo com a metodologia aqui proposta.

3.3 A estratégia para a predição de empresas insolventes

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

Vale recordar que um dos principais modos para tratar uma base de dados desbalanceado 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.

O modelo desenvolvido, a ser aplicado neste trabalho, 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 é denominado *Semi-Determinístic Ensemble Strategy for Imbalanced Data* (SEID).

A forma definida para levar em conta os dois objetivos conjuntamente foi utilizar um comitê de classificadores (*ensemble classifier*) (TSAI; WU, 2008; NANNI; LUMINI, 2009). Em termos práticos, um comitê de classificadores é composto por vários classificadores individuais, cada um gerado com dados/parâmetros diferentes, que devem ser levados em conta no processo de indução baseando-se em alguma estratégia de combinação dos resultados individuais. Os modelos mais representativos de comitê de classificadores são os algoritmos de *bagging* (BREIMAN, 1996) e *boosting* (SCHAPIRE, 1990). No algoritmo de *bagging*, são gerados um determinado número de classificadores individuais por meio de bases obtidas com o mesmo número de instâncias da base original geradas através da escolha das instâncias via distribuição uniforme com reposição da base original. No algoritmo de *boosting*, busca-se aumentar o nível de predição focando-se no desenvolvimento de classificadores individuais que tenham um enfoque maior na classificação das instâncias que se apresentam com maior dificuldade de discriminação.

Um procedimento de comitê apresenta, naturalmente, uma facilidade de implementação visando o enfoque esperado para cada classe, tal como descrita acima. No caso da necessidade de redundância das instâncias minoritárias, tem-se a facilidade de utilizá-las em cada base para a geração dos classificadores individuais que compõem o comitê. No caso das instâncias majoritárias, em que se pretende particionar ou decompor em subconjuntos, podem-se distribuir suas instâncias em subases 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 da base, pois uma estratégia de comitê lida bem com bases de dados menos completas, 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.

Vale ressaltar que esta estratégia para balanceamento baseada em comitê permite o uso de um procedimento de seleção de características de forma diferenciada, descrita mais adiante. O método para predição em bases desbalanceadas, aplicado na determinação de processo de insolvência em empresas. O modelo foi inicialmente aplicado na predição de insolvência de empresas listadas na BM&FBOVESPA sem distinção de setor por Horta, De Lima e Borges (2008, p. 208).

Considera-se, inicialmente, a composição do conjunto de treinamento:

$$Str = Str_m \cup Str_M ,$$

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

Os conjuntos de treinamento gerados para a obtenção dos classificadores individuais serão balanceados com n_{ic} instâncias em cada classe, a saber, majoritária e minoritária. Para que se obtenham conjuntos de treinamento com as características previstas, adota-se como valor mínimo para o número de instâncias por classe n_{ic} o seguinte valor:

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

onde n_{bc} é o número de classificadores bases (individuais) usados no comitê de classificadores e o operador max (*) calcula o maior valor entre os argumentos. 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. A expectativa do algoritmo é que poucos classificadores bases sejam necessários para a geração de um comitê de classificadores de qualidade adequada. Adotou-se três classificadores para a formação do comitê. A figura 1 apresenta o pseudo-código do comitê de classificadores.

Figura 1: 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, se 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.

Fonte: Autores

3.4 Seleção de atributos

Apesar de parecer "óbvia" ou "sempre necessária", a seleção de atributos é uma opção metodológica de fundamental importância em mineração de dados (MD), sendo frequentemente realizada como uma etapa de pré-processamento. Os principais objetivos da seleção de atributos para previsão de insolvência, segundo Piramuthu (2006), são o desenvolvimento de modelos compactos, o uso e refinamento do modelo de classificação ou predição para avaliação, e a identificação de índices financeiros relevantes.

Avaliar o subconjunto de atributos selecionados é medir quão bom é determinado atributo, segundo um dado critério de avaliação (p. ex., informação, distância, dependência, consistência, precisão). Em outras palavras, avaliar o subconjunto é determinar como ele interage com o algoritmo de aprendizado. Essa interação pode ser subdividida, basicamente, em duas abordagens principais - Filtro e Encapsulada (KOHAVI; JOHN, 1997). A abordagem *Wrapper* foi utilizada neste artigo. Para Somol et al. (2005) a abordagem Encapsulada é mais eficiente quando se trata de estudos sobre insolvência de empresas.

Além disso, neste trabalho também foram utilizadas duas abordagens de direção de busca (WITTEN; FRANK, 2011, p. 293): seleção *forward* e seleção aleatória. Estas abordagens foram escolhidas por serem bastante citadas na literatura especifica.

3.5 Uma estratégia de predição de insolvência com seleção de atributos

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 fluxograma do modelo para predição de insolvência com a inclusão do procedimento de seleção de atributos é mostrado a seguir.

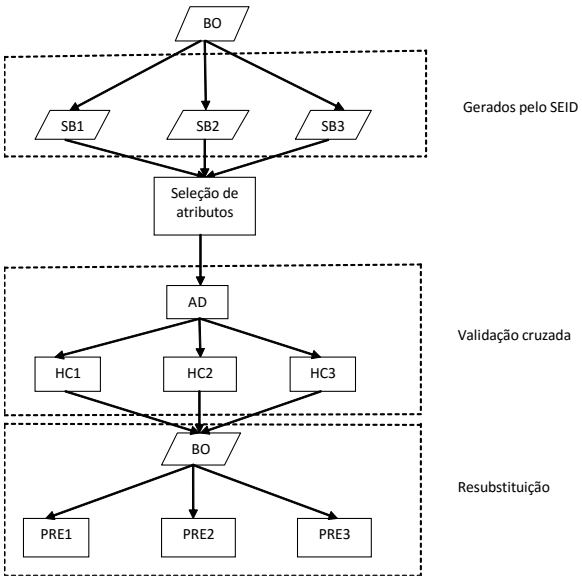

Figura 2.1 - Fluxograma referente aos procedimentos para se chegar aos resultados após os balanceamentos e a seleção de atributos da base de dados original.

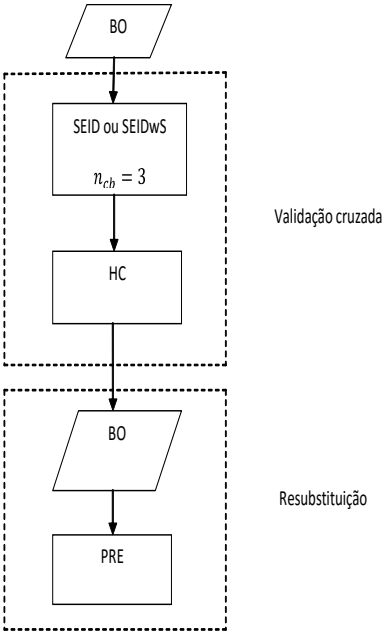

Figura 2.2 – Procedimento de classificação com o SEID ou SEIDwS.

Legenda das siglas na Figura 2.1 e 2.2 – BO: Base de dados original; SB: Subase gerados pelo SEID; AD: Classificador árvore de decisão; HC: Modelos gerados após a seleção de atributos e a aplicação do classificador; PRE: Resultados encontrados após testar os modelos gerados na base de dados original.

3.6 Validação do algoritmo proposto

As Figuras 2.1 e 2.2 oferecem uma visualização dos procedimentos do SEIDwS. A base de dados original (BO) aplicando o SEIDwS é subdividida em três subases por meio da técnica de seleção de atributos. Nessas subases 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 (UC *Irvine Machine Learning Repository*) - *Japanese Credit Screening, Australian Credit Approval, German Credit Data*. Essas bases são normalmente utilizadas para testes de estudos sobre modelagem de previsão de insolvência (ver http://archive.ics.uci.edu/ml).

3.6.1 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 2.1. O classificador AD foi o utilizado.

O Quadro 1 apresenta os resultados dos testes do SEIDwS e do algoritmo SMOTE, neste algoritmo o valor "*k*" foi utilizado com o valor 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; 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 recombinação 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 (c*onsistency*) e o CFS (Seleção de atributos baseado em correlação). Na abordagem *wrapper* os algoritmos indutores foram RL, MLP, 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 - bases de dados sobre insolvência do UCI

Bases de dados do UCI	Nº de	Classe	Instâncias	Base original		SEIDwS		SMOTE	
	atributos			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

Fonte: Autores.

A Tabela 1 apresenta os resultados dos testes feitos com algoritmos de balanceamentos (SEIDwS e SMOTE). O número de atributos antes da seleção era para as bases de dados do UCI Japanese, Australian e German, respectivamente, 15, 14 e 20.

Pelos resultados apresentados 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 (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 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.

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 base mais desbalanceada, os resultados foram os melhores.

Vale aqui ressaltar que os resultados apresentados nos Quadros 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 as 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 especifica.

Tabela 2 – Comparação dos resultados do SEIDwS - bases de dados UCI com outros estudos

	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

Fonte: Autores.

Para se chegar aos resultados, apresentados no item seguinte, primeiramente foi elaborada uma base de dados secundários, inédita, contendo índices calculados a partir de demonstrativos contábeis de empresas do setor econômico de material básico, previamente classificadas como solventes e insolventes pela BM&FBOVESPA no período de 1996 a 2012. 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.

4 RESULTADOS

Nesta seção são apresentados os resultados das aplicações à base de dados de interesse. Esta base se refere a variáveis obtidas em demonstrativos contábeis de empresas do setor de material básico. As duas aplicações descrevem, primeiro, resultados sem a estratégia SEIDwS seção 4.1) e, a seguir, a aplicação dos classificadores após o uso da estratégia SEIDwS (§ 4.2). Na terceira subseção aparecem os resultados da votação majoritária e na subseção 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 (MC), medida F e AUC.

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

Classificador	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	25	5	0,757	0,907	22	8	0,83	0,928	25	5	0,847	0,923	24	6	0,873	0,942
S	11	340	0,977	0,907	1	350	0,987	0,928	4	347	0,987	0,923	1	350	0,99	0,942

Fonte: Elaborado pelos autores.

Para a base de dados do setor econômico de empresas de materiais básicos, em relação à Matriz de Confusão, a medida F e área AUC apresentaram pouca diferença entre os classificadores. Entretanto, SVM e AD foram aqueles que foram capazes de melhor classificar tanto as empresas solventes quanto as empresas insolventes. 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: EOCpOT, EOAT, GAF, MB, TERFIN. A interpretação dessas variáveis será apresentada nos itens seguintes.

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 2.1).

Nas aplicações com a abordagem *wrapper* (conforme Figura 2.1) foram testados para o método de busca GS (*GeneticSearch*) e GD (*GreedyStepwise*). GS obteve os melhores resultados. O classificador utilizado foi o AD. Os resultados encontrados estão na Tabela 4.

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

	SB1_MB30WRAPPER+GS				SB2_MB30WRAPPER+GS				SB3_MB30WRAPPER+GS			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	30	0	0,968	0,984	30	0	0,968	0,984	30	0	0,937	0,967
S	2	349	0,997	0,984	2	349	0,997	0,984	4	347	0,994	0,967

Fonte: Autores.

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 sete foram selecionadas pela abordagem encapsulada. As variáveis selecionadas foram: EOCpOT, EOAT, GAF, MB, EBITDA, MO e TERFIN.

Com a aplicação da estratégia SEIDwS foram selecionadas mais duas variáveis, EBITDA e MO. Como a estratégia SEIDwS melhora a caracterização das empress potencialmente insolventes pode-se deduzir que nas empresas pertecentes ao setor econômico de material básico as variáveis que representam os desempenhos operacionais 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; 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 2.1 e 2.2. 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	25	5	0,847	0,937	30	0	0,968	0,984
S	4	347	0,987	0,937	2	349	0,999	0,984

Fonte: Autores.

4.4 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	25	5	0,847	0,937	30	0	0,968	0,984	29	1	0,951	0,989
S	4	347	0,987	0,937	2	349	0,999	0,984	1	350	0,999	0,989

Fonte: Autores.

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

No mesmo quadro a comparação do SEIDwS com o SMOTE evidencia 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 (0,984 x 0,989). Diante desses resultados pode-se inferir a capacidade bastante competitiva da estratégia aqui apresentada (SEIDwS) em relação ao SMOTE, uma técnica já consagrada e bem referenciada na literatura específica.

5 CONCLUSÕES E FUTUROS ESTUDOS

Esta pesquisa aplicou e testou uma estratégia para tratar um problema pouco estudado em modelagem para previsão de insolvência 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 não representa 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 aplicado, adequar a base de dados ao ambiente econômico das empresas.

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.

Nas implicações para a análise contábil do estudo devemos considerar que em relação às variáveis contábeis selecionadas, prevaleceram as originadas no Balanço Patrimonial, sobretudo aquelas que aferem a composição (estrutura) das fontes passivas de recursos das empresas (EOCpOT, EOAT, GAF, TERFIN). Depois da aplicação da estratégia SEIDwS somam-se a estas variáveis MO e EBITDA, representativas do desempenho operacional das empresas.

Com a seleção de MO e EBITDA pode-se inferir que, para a amostra estudada, a insolvência se relaciona não somente a aspectos financeiros, relativos à incapacidade da empresa de se endividar (EOCpOT, EOAT, GAF, TERFIN), mas também a aspectos de seu desempenho operacional. As variáveis MO e EBITIDA representam o desempenho e a capacidade operacional de geração de caixa da empresa, ou seja, sua eficiência financeira determinada pelas estratégias operacionais adotadas. Em outras palavras, na amostra estudada empresas do setor de materiais básicos tornam-se insolventes não somente porque perdem a capacidade (financeira) de se endividar, mas também porque perdem a capacidade operacional de gerar caixa. 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.

Outra contribuição deste trabalho às pesquisas sobre previsão de insolvência foi a utilização de dados secundários obtidos exclusivamente em demonstrativos contábeis de

empresas brasileiras, acrescenta-se a isso o estudo de uma base de dados originada de empresas de um setor econômico específico. Durante algum tempo os dados contábeis, no Brasil, foram tratados com desconfiança, de modo que pesquisas e conclusões como a presente servem para reiterar a conveniência daquela utilização para a análise da evolução da cultura contábil e econômica de empresas em nosso país.

Agradecemos a FAPEMIG e ao CNPQ, respectivamente, pelo apoio concedido à pesquisa APQ 00592-14 e a pesquisa cujo Processo é 445335/2014-9.

REFERÊNCIAS

ALTMAN, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. **Journal of Finance**, v. 23, n. 4, p. 589-609, 1968. http://dx.doi.org/10.1111/j.1540-6261.1968.tb00843.x

_____; HALDEMAN, R. G.; NARAYANAN, P. Zeta Analysis: A new model to identify bankruptcy risk of corporations, **Journal of Banking and Finance**, v. 1, p. 29–54, 1977. http://dx.doi.org/10.1016/0378-4266(77)90017-6

_____; BAIDYA, T. K. N.; DIAS, L. M. R. **Previsão de problemas financeiros em empresas**. **Revista de Administração de Empresas**, v. 19, jan./mar., p. 17-28, 1979. http://dx.doi.org/10.1590/S0034-75901979000100002

_____; GIANCARLO, M.; VARETTO, F. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). **Journal of Banking & Finance,** v. 18, n. 3, p. 505-529, 1994. http://dx.doi.org/10.1016/0378-4266(94)90007-8

ATIYA, A. F. Bankruptcy prediction for credit risk using neural network: a survey and new results. **IEEE transactions on neural networks**, v. 12, n. 4, p. 929-935, 2001. http://dx.doi.org/10.1109/72.935101

BALCAEN, S.; OOGHE, H. 35 Years of studies on business failure: on overview of the classical statistical methodologies and their related problems. **The British Accounting Review**, v. 38, n. 1, p. 63-93, 2006. http://dx.doi.org/10.1016/j.bar.2005.09.001

BRAGA-NETO, U.; HASHIMOTO, R.; DOUGHERTY, E. R.; NGUYEN, D. V.; CARROLL, R. J. Is cross-validation better than resubstitution for ranking genes? **Bioinformatics**, v. 20, n. 2, p. 253-258, 2004. http://dx.doi.org/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, p. 528–546, 2005. http://dx.doi.org/10.1016/j.ejor.2004.03.023

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, p. 321-357, 2002. http://dx.doi.org/10.1613/jair.953

_____; JAPKOWICZ, N.; KOLCZ, A. Editorial: Special Issue on Learning from Imbalanced Datasets. *SIGKDD* **Explorations**, v. 6, n. 1, p. 1-6, 2004. http://dx.doi.org/10.1145/1007730.1007733

DAMODARAN, Aswath. **Finanças corporativas aplicadas**. Tradução Jorge Ritter. Porto Alegre: Bookman, 2002.

DING, Y.; SONG, X., ZEN,Y. Forecasting financial condition of Chinese listed companies based on support vector machine. **Expert Systems with Applications**, v. 34, n. 4, p. 3081-3089, 2008. http://dx.doi.org/10.1016/j.eswa.2007.06.037

ELIZABETSKY, R. **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. Mining with Rarity: A Unifying Framework, **SIGKDD Explorations**, v. 6, Issue 1, 2004, p.7-19.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** concepts and techniques. 3rd ed. Waltham: Morgan Kaufmann, 2011. 744 p.

HORTA, Rui Américo Mathiasi. **Utilização de indicadores contábeis na previsão de insolvência: Análise empírica de uma amostra de empresas comerciais e industriais brasileiras.** 2001. 108 p. Dissertação Mestrado em Ciências Contábeis – Faculdade de Ciências Contábeis da Universidade Estadual do Rio de Janeiro.

HORTA, Rui Américo Mathiasi. **Uma metodologia de mineração de dados para a previsão de insolvência de empresas brasileiras de capital aberto**. 2010. 152 p. Doutorado em Engenharia Civil – COPPE - Universidade Federal do Rio e Janeiro.

HORTA R.A.M., DE LIMA B.S.L.P., BORGES C.C.H. A semi-deterministic ensemble strategy for imbalanced datasets (SEID) applied to bankruptcy prediction. In: Data mining IX: data minig, protection, detection and other security technologies. WIT transactions on information and communication technologies, v. 40, Spain, 2008, p. 205–213.

HORTA, Rui Américo Mathiasi; Carlos Cristiano Hasenclever Borges; JORGE, M. J. Descontinuidade de empresas brasileiras do setor de consumo cíclico: Uma metodologia para balanceamento de base de dados utilizando técnicas de *data mining*. **Revista Ambiente Contábil**, v. 6, p. 99-121, 2014.

HORTA, Rui Américo Mathiasi; JORGE, M. J.; ALVES, F. J. Descontinuidade de empresas brasileiras do setor de consumo não cíclico: Uma metodologia para balanceamento de base de dados utilizando técnicas de *data mining*. **Revista de Informação Contábil** (UFPE), v. 07, p. 63-83, 2013.

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, p. 434-440, 2007. http://dx.doi.org/10.1016/j.eswa.2006.05.006

HUNG, Chihli; CHEN, Jing-Hong. A selective ensemble based on expected probabilities for bankruptcy prediction. **Expert systems with applications**, v. 36, Issue 3, p. 3297-5309, 2009. http://dx.doi.org/10.1016/j.eswa.2008.06.068

IUDÍCIBUS, S. de. **Análise de Balanços**. 9. ed. São Paulo: Atlas, 2008. 258 p.

JAPKOWICZ N.; STEPHEN, S. The Class Imbalance Problem: A Systematic Study. **Intelligent Data Analysis**, v. 6, Issue 5, p. 429-449, 2002.

JOSHI, M. V. **Learning Classier Models for Predicting Rare Phonemena**. PhD thesis, University of Minnesota, Twin Cites, Minnesota, USA, 2001.

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, p. 838-846, 2010. http://dx.doi.org/10.1016/j.ejor.2009.03.036

KOHAVI, R..; JOHN, G. H. Wrappers for feature subset selection**. Artif. Intell.,** v.97, 1997. p. 273-324. http://dx.doi.org/10.1016/S0004-3702(97)00043-X

LI HUI, JIE SUN. Majority voting combination of multiple case-based reasoning for financial distress prediction. **Expert Systems with Applications**, v.36, p. 4363-4373, apr, 2009. http://dx.doi.org/10.1016/j.eswa.2008.05.019

LOPES, Alexsandro Broedel; MARTINS, Eliseu. **Teoria da contabilidade: uma nova abordagem** – 2ª. Reimpressão – São Paulo: Atlas, 2007.

MARTIN, D. Early warning of bank failure: A logit regression approach, **Journal of Banking and Finance**, v.1, p. 249–276, 1977. http://dx.doi.org/10.1016/0378-4266(77)90022-X

MATARAZZO, Dante Carmine. **Análise financeira de balanços: abordagem básica e gerencial.** 6ª Ed. São Paulo: Atlas, 2003.

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, p. 652-660, oct. 2006. http://dx.doi.org/10.1016/j.eswa.2005.09.070

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. http://dx.doi.org/10.1590/S1519-70772006000200007

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, p. 3028-3033, mar. 2009. http://dx.doi.org/10.1016/j.eswa.2008.01.018

OHLSON, J.A. Financial ratios and the probabilisitic prediction of bankruptcy. **Journal of Accounting Research**, v. 18, p.109-131, 1980. http://dx.doi.org/10.2307/2490395.

OLSON, David L. DELEN, Dursun, MENG, Yanyan. Comparative analysis of data mining methods for bankruptcy prediction. **Decision Support Systems**, v. 52, n. 2, January 2012, p. 464-473. http://dx.doi.org/10.1016/j.dss.2011.10.007

PIRAMUTHU S. On preprocessing data for financial credit risk evaluation. **Expert Systems with Applications**, v. 30, p.489-497, 2006. http://dx.doi.org/10.1016/j.eswa.2005.10.006

RAVI, V.; KURNIAWAN, H.; THAI, Peter Nwee Kok,; KUMAR, P. Ravi. Soft computing system for bank performance predition. **Applied Soft Computing**, v. 8, p. 305-315, jan. 2008. http://dx.doi.org/10.1016/j.asoc.2007.02.001

ROSS, Stephen A.; WESTERFIELD, Randolph, W. Jaffe; JORDAN, Bradford; LAMB, Roberto. **Fundamentos de Administração Financeira** – Porto Alegre: Editora Mc Graw Hill, 9ª edição, 2013.

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/artigostécnicos/index.html>. Acesso em: 23/10/2005.

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, p. 127-135, jan. 2005.

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, n. 51, p. 28-43, Setembro/Dezembro, 2009.

SILVA, José Pereira da. **Gestão e análise de risco de crédito**. 5. ed. São Paulo: Atlas, 2006. 448 p.

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, p. 985-999, 2005.

SUN, Jie; LI, Hui; HUANG, Qing-Hua; HE, Kai-Yu. Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. **Knowledge-Based Systems**, v. 57, p. 41-56, February 2014.

TINOCO, Mario Hernandez, WILSON Nick. Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. **International Review of Financial Analysis**, v. 30, p. 394-419, December 2013.

TSAI, C. F. Feature selection in bankruptcy prediction. **Knowledge-Based Systems**, v. 22, n. 2, p. 120-127, mar. 2009. http://dx.doi.org/10.1016/j.knosys.2008.08.002

_____; WU J. W. Using neural network ensembles for bankruptcy prediction and credit scoring. **Expert Systems with applications**, v. 34, n. 4, p. 2639-2649, May 2008.

_____; HSU, Yu-Feng; YEN, David C. A comparative study of classifier ensembles for bankruptcy prediction. **Applied Soft Computing**, v. 24, p. 977-984, November 2014.

VERIKAS, Antanas; KALSYTE, Zivile; BACAUSKIENE, Marija; GELZINIS, Adas. Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. **Soft Comput** v.14, p. 995-1010, 2010. http://dx.doi.org/10.1007/s00500-009-0490-5

WEISS, G. M.; McCARTHY, K.; 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, Fordham University, Bronx, NY, USA, SREA Press, p. 35-41, 2007.

WEST, David; DELLANA, Scott; QIAN, Jingxia. Neural network ensemble strategies for financial decision applications. **Computers & Operations Research**, v. 32, n. 10, p. 2543-2559, oct. 2005. http://dx.doi.org/10.1016/0378-4266(85)90021-4

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

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

WU, Y.; GAUNT, C. GRAY, S. A comparison of alternative bankruptcy prediction models. **Journal of Contemporary Accounting & Economics**, v. 6, n. 1, p. 34-45, June 2010. http://dx.doi.org/10.1016/j.jcae.2010.04.002

YU, L. WAUNG; LAI, K. K. Credit risk assessment with a multistage neural network ensemble learning approach. **Expert Systems with Applications**, v. 34, p. 1434-1444, fev. 2008. http://dx.doi.org/10.1016/j.eswa.2007.01.009

ZHOU, Ligang. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. **Knowledge-Based Systems**, v. 41, p. 16-25, mar. 2013. http://dx.doi.org/10.1016/j.knosys.2012.12.007

ANEXO 1 – VARIÁVEIS CONTÁBEIS COLETADAS

Liquidez Corrente - LC, Liquidez Seca – LS, Liquidez Imediata – LI, Liquidez Geral – LG, Endividamento Oneroso de Curto Prazo sobre o Ativo Total – EOCpOT, Endividamento Oneroso sobre Patrimônio Líquido – EOPL, Endividamento Total sobre o Patrimônio Líquido – ETPL, Endividamento Oneroso sobre Ativo Total – EOAT, Endividamento Total sobre Ativo Total – ETAT, 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.