http://dx.doi.org/10.18593/race.v14i1.4760

Disponível em:
http://editora.unoesc.edu.br/index.php/race

Race, Joaçaba, v. 14, n. 1, p. 171-196, jan./abr. 2015

DESCONTINUIDADE DE EMPRESAS BRASILEIRAS DO SETOR DE MATERIAL BÁSICO: NO PERÍODO COMPREENDIDO PRÉ E PÓS A CRISE DO *SUBPRIME*

Discontinuance of brazilian companies of basic materials sector: the period pre and post the subprime crisis

Rui Américo Mathiasi Horta
E-mail: rui.horta@ufjf.edu.br
Doutor em Engenharia Civil pela Universidade Federal do Rio de Janeiro; Professor Adjunto do Departamento de Finanças e Controladoria da Universidade Federal de Juiz de Fora; Rua José Lourenço Kelmer, s/n, Campus Universitário, São Pedro, 36036-900, Juiz de Fora, Minas Gerais, Brasil.

Francisco José dos Santos Alves
E-mail: fjalves@globo.com
Doutor em Controladoria e Contabilidade pela Universidade de São Paulo; Professor Adjunto do Departamento de Finanças na Universidade do Estado do Rio de Janeiro.

Carlos Cristiano Hasenclever Borges
E-mail: cchb@lncc.br
Doutor em Engenharia Civil pela Universidade Federal do Rio de Janeiro; Professor Adjunto do Departamento de Ciências da Computação da Universidade Federal de Juiz de Fora.

Adriano Rodrigues
E-mail: adriano@facc.ufrj.br
Doutor em Controladoria e Contabilidade da Universidade de São Paulo; Professor Adjunto do Departamento de Ciências Contábeis da Universidade Federal do Rio de Janeiro.

Artigo recebido em 09 de maio de 2014. Aceito em 04 de agosto de 2014

Resumo

Descontinuidade de empresas é um tema cada vez mais estudado no campo da contabilidade, das finanças, dos negócios e da computação em decorrência do considerável impacto social causado pelo fracasso corporativo de uma entidade. Bancos, investidores, auditores, gerentes, fornecedores, empregados e muitos outros têm grandes interesses na acurácia da previsão de continuidade de uma companhia. Ainda há questões pouco estudadas na modelagem de previsão de insolvência. O desequilíbrio ou desbalanceamento dos dados sobre insolvência é uma dessas questões; em ambientes econômicos típicos o número de empresas classificadas como solventes é bem maior do que o daquelas classificadas como insolventes. O objetivo deste estudo foi comparar as variáveis contábeis selecionadas nas amostras das empresas do setor de material básico antes e após a crise do *subprime*, aplicando tecnologias de aprendizagem de máquinas em problemas de previsão de insolvência, utilizando técnicas de balanceamento da base de dados com (etapa de) seleção de atributos e, a partir disso, obter informações contábeis que levem a explicações das diferenças ocorridas na descontinuidade das empresas estudadas. Esta pesquisa é de natureza aplicada com abordagem quantitativa; quanto ao objetivo, é descritiva. A base de dados foi originada de demonstrativos contábeis de empresas brasileiras do setor econômico de material básico, listadas na Bovespa e na Serasa entre os anos 1994 e 2006 e 2007 e 2012. Os resultados obtidos evidenciaram a alteração de algumas variáveis selecionadas para a caracterização daquelas empresas que se tornaram descontínuas. Empresas do setor de material básico, no período estudado, 2007 a 2012, que se adequaram às exigências do mercado, privilegiando a liquidez em detrimento ao operacional, lograram sucesso em sua continuidade.
Palavras-chave: Descontinuidade de empresas. Seleção de variáveis contábeis. Balanceamento de base de dados. Tecnologias de aprendizagem de máquina. Setor de material básico. Crise s*ubprime* Brasil.

Discontinuance of brazilian companies of basic materials sector: the period pre and post the subprime crisis

Abstract

Discontinuity of companies is a topic increasingly studied in the field of accounting, finance, business and computing due to the considerable social impact caused by failure of a corporate entity. Banks, investors, auditors, managers, suppliers, employees and many others have great interests in the accuracy of prediction of continuity of a company. There is still little studied issues in predictive modeling of insolvency. The imbalance or unbalance data insolvency is one of those issues, in typical economic environments the number of companies classified as solvents is much greater than those classified as insolvent. The aim of this study was to compare the accounting variables selected in the samples of companies in the basic materials sector before and after the subprime crisis by applying machine learning technologies in problems of insolvency prediction, using balancing techniques with database (step a) selection of attributes. Thereafter, obtain accounting information leading to explanations of the differences arising in the discontinuity of the companies

studied. This research is of an applied nature with a quantitative approach; about the objectives, it is exploratory and explanatory. The database was derived from financial statements of Brazilian companies in the basic materials economic sector, listed on the BOVESPA and SERASA between the years 1994-2006 and 2007-2012. The results showed alteration of certain selected to characterize those companies that have become discontinuous variables. Companies in the basic materials sector, during the study period from 2007 to 2012, that suited the requirements of the market, focusing on liquidity rather than the operational, managed to succeed in its continuity.

Keywords: Discontinuance of business. Selection of accounting variables. Balancing database. Machine learning technologies. Basic material sector. Subprime crisis – Brazil.

1 **INTRODUÇÃO**

Cada vez mais o desenvolvimento de estudos sobre modelagem para previsão de descontinuidade de empresas vem adquirindo importância nas áreas relativas à Contabilidade, Finanças, Negócios e Computação. De fato, a previsão de insolvência permite 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.

Vários fatores têm concorrido para o aumento quantitativo e qualitativo dos estudos sobre o tema. Por exemplo, em vários países, a maioria das estatísticas sobre falências mostrou significativo crescimento. Além disso, nas últimas décadas o ambiente econômico geral das empresas, na maioria dos países, tem mudado com enorme velocidade e experimentado tendências adversas. Cresceu, também, a cautela associada à implementação de normas internacionais de contabilidade (IFRS) e finanças, 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 a instabilidade dos dados, a seleção da amostra e o desequilíbrio entre as classes (BALCAEN; OOGHE, 2006; TSAI; WU, 2008; NANNI; LUMINI, 2009; GESTEL; BAESENS; MARTENS, 2010; ZHOU, 2013; SUN et al., 2014).

Uma dessas questões pouco exploradas é o problema do desequilíbrio de tamanho das classes inicialmente disponíveis quando se observa a separação entre empresas solventes e insolventes. Cabe reconhecer que essa situação é natural, porque, normalmente, em qualquer sociedade, a classe de empresas insolventes é muito inferior à de solventes, independentemente do período que se analisa. Por isso mesmo se requer 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.

A solução para o problema de desbalanceamento em classificação de dados pode ser considerada relativamente nova, entre "[...] as respostas que surgiram quando as ideias relacionadas à aprendizagem de máquina (*machine learning*) tornaram-se uma tecnologia efetivamente aplicada e amplamente utilizada em áreas como negócios, indústria, linguística, bioinformática entre muitas outras." (CHAWLA; JAPKOWICZ; KOLZ, 2004, p. 1).

Por outro lado, a crise provocada pela concessão de créditos hipotecários com elevado risco, acrescido do uso de derivativos cuja aplicação nesse tipo de financiamento encontrava-se desregulamentada pelo Estado, gerou nos Estados Unidos mais de 26 milhões de americanos desempregados. Cerca de 4 milhões de famílias perderam suas casas e outros 4 milhões sofreram ações judiciais de despejo (UNITED STATES OF AMERICA, 2011, p. 15).

Essa crise evidenciada a partir de 2007 desencadeou um efeito dominó no mercado financeiro e, posteriormente, na economia real. As operações com derivativos e a falência de grandes bancos de investimento geraram um efeito de restrição de crédito e impactaram negativamente todas as principais economias mundiais (CROUHY; JARROW; TURNBULL, 2008).

O efeito chegou ao Brasil ainda em 2007: o Ibovespa encontrava-se acima de 60.000 pontos e chegou a ficar abaixo de 30.000. O Governo brasileiro teve que estabelecer uma série de políticas para evitar a crise, como, por exemplo, alterações de taxas de juros e também dos tributos. Apenas em 2009, o principal índice da BM&FBovespa retornou a patamares próximos do período anterior à crise (COSTA; REIS; TEIXEIRA, 2012).

Nesse contexto, este estudo teve por objetivo comparar as variáveis contábeis selecionadas nas amostras das empresas do setor de material básico de antes e após a crise do *subprime* (2007), aplicando tecnologias de aprendizagem de máquinas em problemas de previsão de insolvência, utilizando técnicas de balanceamento da base de dados com (etapa de) seleção de atributos e, a partir disso, obter informações contábeis que levem a explicações das diferenças ocorridas na descontinuidade das empresas do setor econômico estudado.

Esta pesquisa é de natureza aplicada com abordagem quantitativa; quanto aos objetivos, é exploratória e explicativa. Utilizaram-se dados obtidos em

demonstrativos contábeis de empresas classificadas na Bolsa de Valores de São Paulo (Bovespa) e Serasa, pertencentes ao setor econômico de material básico.

A utilização de uma base empírica apoiada em demonstrativos contábeis se justifica plenamente pelo pressuposto de que

> [...] na previsão de insolvências, os principais indicadores macroeconômicos (p. ex., inflação, juros, impostos, etc.), juntamente com as características das empresas (concorrência, gestão, capacidade produtiva, produto, etc.), estão devidamente refletidos naqueles demonstrativos, de tal modo que a futura situação financeira da empresa possa ser prevista usando dados deles para alimentar técnicas de modelagem avançadas. (GESTEL; BAESENS; MARTENS, 2010, p. 2956).

Já em relação à crise do *subprime*, ela começa a ficar bem evidente por meio da queda dos preços dos imóveis nos Estados Unidos, desencadeando um efeito dominó no mercado financeiro e, posteriormente, na economia real. O *subprime* é um termo empregado para designar uma forma de crédito hipotecário (*mortgage*) para o setor imobiliário, surgida nos Estados Unidos e destinada a tomadores de empréstimos que representam maior risco. Esse crédito imobiliário tem como garantia a residência do tomador e, muitas vezes, era acoplado à emissão de cartões de crédito. De uma forma ampla, o *subprime* é um crédito de risco concedido a um tomador que não oferece garantias suficientes para se beneficiar da taxa de juros mais vantajosa (*prime rate*). As operações com derivativos e a falência de grandes bancos de investimento geraram um efeito de restrição de crédito e impactaram negativamente todas as principais economias mundiais (CROUHY; JARROW; TURNBULL, 2008). O efeito chegou ao Brasil ainda em 2007: o Ibovespa encontrava-se acima de 60.000 pontos e chegou a ficar abaixo de 30.000 pontos. O Governo brasileiro teve que estabelecer uma série de políticas para evitar a crise, como, por exemplo, alterações de taxas de juros e também dos tributos. Apenas em 2009, o principal índice da BM&FBovespa retornou a patamares próximos do período anterior à crise (COSTA; REIS; TEIXEIRA, 2012).

O artigo está organizado em cinco seções, incluindo esta Introdução. A seção 2 apresenta a revisão bibliográfica que forneceu 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 análises e conclusões da pesquisa sendo também sugeridos futuros estudos.

2 **FUNDAMENTAÇÃO TEÓRICA**

A previsão de insolvência tornou-se o assunto mais pesquisado e difundido na década de 1960, notadamente por meio do modelo chamado Escore-Z (ALTMAN, 1968). Altman, Haldeman e Narayanan (1977) desenvolveram um novo modelo de classificação de insolvência, chamado Zeta, uma atualização e aprimoramento do modelo Escore-Z original; em ambos os estudos foi utilizada a análise discriminante.

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. Canbas, Cabuk e Kilic (2005) combinaram análise discriminante linear (LDA), regressão logística (RL), *probit* e análise de componentes principais em sua modelagem da insolvência.

Mais recentemente, estratégias baseadas em tecnologias computacionais começaram a ser aplicadas visando à detecção de insolvência. Shin, Lee e Kim (2005) investigaram a eficácia da aplicação de Máquinas de Vetor Suporte (SVM) para o problema de previsão de falências, mostrando que o classificador SVM supera as redes neurais (ANN) 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.

Alguns autores, visando aumentar a eficácia da predição, também desenvolveram metodologias específicas no uso dos classificadores ou na manipulação das bases de dados. Por exemplo, Atiya (2001) desenvolveu um estudo sobre previsão de insolvência em que aplica redes neurais em um caso de bancos de dados desbalanceados. Em busca de maior precisão nas previsões, Tsai e Wu (2008) compararam o desempenho de um classificador simples de ANNs com o de múltiplos classificadores, também baseados em ANNs. Fazendo aplicação de comitês de classificadores, 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. Hung e Chen (2009) aplicaram um modelo de probabilidades híbridas, baseado em comitê de classificadores, para previsão de insolvência, utilizando votação majoritária e votação ponderada.

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 estudo. Essa

situação começa a ser mudada, 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 de maior relevância aplicados em dados de empresas brasileiras.

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 – por exemplo, Altman, Baidya e Dias (1979) – também recorreu à ferramenta estatística de análise discriminante, bem como Sanvicente e Minardi (1998). Morozini, Olinquevitch e Hein (2006) utilizam análise dos componentes principais para combinar os principais índices entre 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, propõe-se a resolver o problema do desbalanceamento entre as classes de empresas solventes e insolventes existente em estudos de previsão de insolvência.

3 **METODOLOGIA DA PESQUISA**

Este estudo tem como objetivo comparar as variáveis selecionadas nas amostras das empresas do setor de material básico de antes e após a crise do *subprime* (2007), aplicando tecnologias de aprendizagem de máquinas em problemas de previsão de insolvência, utilizando técnicas de balanceamento da base de dados com (etapa de) seleção de atributos e a partir disso, obter informações contábeis que levem a explicações das diferenças ocorridas na descontinuidade das empresas do setor econômico estudado. Esta pesquisa é de natureza aplicada com abordagem quantitativa; quanto ao objetivo, é descritiva. A seguir são apresentados os passos metodológicos cumpridos para alcançar o objetivo.

3.1 BASE DE DADOS E MÉTRICAS DE AVALIAÇÃO

Foram obtidos nos demonstrativos contábeis de empresas, publicados na Bovespa, 23 indicadores contábeis anuais das empresas do setor de material básico, classificadas de acordo com grupos de índices contábeis-financeiros: liquidez, endividamento, rentabilidade e ciclo operacional (APÊNDICE A). Na visão de

Wolk, Dodd e Rozycki (2013, p. 304), "[...] variáveis contábeis são muito usadas para discriminar empresas que apresentam tendências de se tornarem insolventes daquelas solventes."

Importa, aqui, evidenciar a importância e os principais motivos da escolha de um conjunto de empresas pertencentes a um mesmo setor econômico. Para Iudícibus (2009, p. 91), "[...] os demonstrativos contábeis de empresas do mesmo setor econômico apresentam semelhanças devido a suas estruturas patrimoniais e econômicas. Indicadores tais como liquidez, endividamento e rentabilidade deveriam apresentar valores bem próximos, em termos da média setorial."

Compõem este setor empresas dos subsetores de embalagem, madeira e papel, materiais diversos, mineração, químicos, siderurgia e metalurgia. Nesse setor econômico, as empresas apresentam, normalmente, valores proporcionalmente altos em seus ativos não circulantes; em empresas, sobretudo dos setores de madeiras e papel, químicos, siderurgia e metalurgia, esses ativos são substanciados pelos ativos imobilizados (instalações, equipamentos, máquinas, etc.). Empresas desse setor foram as escolhidas em decorrência do baixo índice do tamanho de seus ativos intangíveis (MOURA et al., 2013, p. 131) e, por consequência disso, um dos setores mais vulneráveis aos efeitos da crise.

Na montagem da base de dados, cada uma das empresas foi classificada como concordatária, em recuperação judicial ou falida na Bovespa, durante o período de 2007 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 havia solicitação de concordata por parte da empresa no período considerado). O estabelecimento de uma quantidade superior de empresas adimplentes para cada inadimplente baseia-se na hipótese de "[...] que algum evento que dependa de um conjunto de variáveis aleatórias cujo número aumenta sem limite e cada uma das quais tem apenas um efeito insignificante no conjunto, ocorrerá com probabilidade tão próxima de um quanto desejado" (GNEDENKO, 2008, p. 299) além de representar melhor a realidade econômica.

Foi utilizada análise de dados em painel, pois segundo Pindick e Rubinfeld (2004) e Gujarati (2006 apud. FÁVERO et al. (2009, p. 382), as principais características da análise de dados em painel são: maior número de observações para se trabalhar, com consequente aumento do número de graus de liberdade e eficiência dos parâmetros, redução de problemas de multicolinearidade de variáveis explicativas e existência da dinâmica intertemporal. Em modelos de previsão de

insolvência elaborados com dados em painel, cada empresa fornece dados contínuos durante períodos (anos) (HUNG; CHEN, 2009, p. 5301). Na análise de dados em painel pela facilidade de acesso a uma maior quantidade de dados obtidos nos demonstrativos contábeis, em razão da criação de instância em quantidade bem maior do que o número de empresas, as aplicações e os estudos nessas bases acabam por apresentar melhores resultados. Disso, a sua preferência e utilização (BALCAEN; OOGHE, 2006, p. 69).

Estudos tradicionais de previsão de continuidade utilizam os métodos de estatística convencionais, como análise discriminante múltipla, *logit* e *probit*. No entanto esses métodos apresentam algumas hipóteses restritivas, como a linearidade, a normalidade e a independência dos preditores ou variáveis de entrada. Considerando que a violação dessas premissas ocorre com frequência com o uso de dados financeiros (YEH; CHI; LIN, 2014, p. 98), abordagens de aprendizagem de máquina (AM), como a árvore de decisão (AD), são menos vulneráveis a essas violações. AM pode ser uma alternativa de solução para problemas de classificação, uma vez que ela demonstrou ter capacidade preditiva superior aos métodos estatísticos convencionais de previsão de continuidade (YEH; CHI; LIN, 2014, p. 99).

No presente estudo foram utilizadas técnicas de aprendizagem de máquina e o SEIDWS (HORTA, 2010) para solucionar o problema do desequilíbrio entre as classes de empresas classificadas como solventes e insolventes, além de minorar os problemas de classificação quando da utilização de técnicas tradicionais. Com isso, buscou-se representar melhor o ambiente econômico das empresas. Vale ressaltar que não se conhece estudos com dados contábeis de empresas brasileiras que utilizaram tais ferramentas, seja em separado ou em conjunto.

Na base de dados estudada há 10 instâncias representando as empresas insolventes e 180 representando as solventes do mesmo setor, na proporção de 18 para 1. Para se chegar a essa proporção, adotou-se a seguinte estratégia: primeiro foram obtidos um maior número possível de empresas classificadas como insolventes e que apresentavam demonstrativos contábeis confiáveis e adequados de serem estudados (demonstrativos contábeis publicados na Bovespa/CVM), a seguir, foi obtido o maior número possível de demonstrativos contábeis de empresas classificadas como solventes e que pertencessem ao setor econômico de material básico. Com isso, buscou-se adequar 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 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, Giancarlo e Varetto (1994, p. 508) e com Hung e Chen (2009, p. 5297), as empresas insolventes começam a apresentar características ou indícios de insolvência cerca de cinco anos anteriores ao ano em que ocorre efetivamente a falha.

Os dados sobre empresas solventes totalizaram o período de dez anos, facilitando assim uma melhor caracterização dessas empresas. Pretendeu-se, também: uma adequação após o ano 2005 em que ocorreu a mudança na Lei de Falências no Brasil; utilizar dados obtidos em demonstrativos contábeis elaborados de acordo com as normas internacionais, Leis n. 11.638/07 e 11.941/09 e os procedimentos do Comitê de Pronunciamento Contábeis e, sobretudo, adequar a base de dados a um período de tempo com o ambiente econômico de muitas mudanças e transformações para as empresas brasileiras (crise do *subprime*).

Das métricas de avaliação alternativas existentes para lidar com o problema do desequilíbrio de classes citadas por Kück (2004, p. 68) e Gary (2004, p. 9), foram escolhidas três (APÊNDICE B): Matriz de Confusão (MC), Medida F e Área sob a curva ROC (AUC).

Para a avaliação dos classificadores, foram utilizadas validação cruzada e com resubstituição (BRAGA-NETO et al., 2004, p. 254). Também foi utilizada a técnica da votação majoritária na combinação dos classificadores gerados. A técnica da maioria dos votos é um método simples e eficaz de combinação (LI; JIE, 2009, p. 4366).

3.2 TÉCNICAS DE TRATAMENTO DE BANCOS DESBALANCEADOS

A abordagem baseada em amostras é amplamente usada para resolver o problema de desequilíbrio de classe. A ideia da amostragem é modificar a distribuição das unidades de forma que a classe minoritária seja mais bem representada no conjunto de treinamento.

A maneira mais simples para amenizar o desequilíbrio entre instâncias de cada classe (neste estudo, a classe de empresas solventes e a classe das empresas insolventes) em uma base de dados é balancear artificialmente a distribuição das classes no conjunto de dados. Duas abordagens padrão são utilizadas neste estudo: a remoção de exemplos da classe majoritária – u*nder-sampling* e a inclusão de exemplos

da classe minoritária – *over-sampling*. Ambos os modelos são baseados na retirada ou na colocação de dados na base de forma randômica.

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

O modelo desenvolvido, aplicado neste trabalho, busca diminuir este componente estocástico, visando a utilização dos dados da classe minoritária de forma mais intensa ou redundante, pois se busca maior nível de acerto nessa classe, como é intuitivamente desejável em problemas de previsão de insolvência, e a decomposição da classe majoritária de forma a torná-la de dimensão "aceitavelmente" mais próxima à classe minoritária. É importante ressaltar que a obediência a esses dois objetivos acarretam, como característica adicional, a diminuição da aleatoriedade na obtenção do balanceamento. Assim, esse modelo é denominado *Semi-Determinístic Ensemble Strategy for Imbalanced Data* (SEID). A forma definida para considerar 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 considerados 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*, é gerado um determinado número de classificadores individuais por meio de bases obtidas com o mesmo número de instâncias da base original geradas por meio 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 dos objetivos para cada classe, como descrita anteriormente. 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 sub-bases diferentes para gerar os classificadores que formam o comitê. Dessa 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 disso, 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 essa estratégia para balanceamento baseada em comitê permite o uso de um procedimento de seleção de características de forma diferenciada, descrita mais adiante.

O modelo foi inicialmente aplicado na predição de insolvência de empresas listadas na Bovespa sem distinção de setor por Horta; De Lima e Borges (2008, p. 208).

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

$$Str = Str_m \; Str_M, \tag{1}$$

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 $\#(\star)$ significa número de instâncias do conjunto.

Os conjuntos de treinamento gerados para a obtenção dos classificadores individuais serão balanceados com 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 o seguinte valor:

$$\#(\ Str_m), \#(\ Str_M)/), \tag{2}$$

onde 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 , mais próximo o algoritmo se torna do algoritmo de *bagging*, ou seja, é um algoritmo que cria amostras repetidamente a partir de um conjunto de dados de acordo com uma distribuição uniforme de distribuição. A expectativa do algoritmo é que poucos classificadores bases sejam necessários para a geração de um comitê de classificadores de qualidade adequada. A seguir, apresenta-se o pseudocódigo do comitê de classificadores.

Figura 1 – Algoritmo SEID

Pseudocódigo: comitê de classificadores para base de dados desbalanceada (SEID)
Início
Defina o número de classificadores bases $n_{\_cb}$
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.

Fonte: os autores.

3.4 SELEÇÃO DE ATRIBUTOS

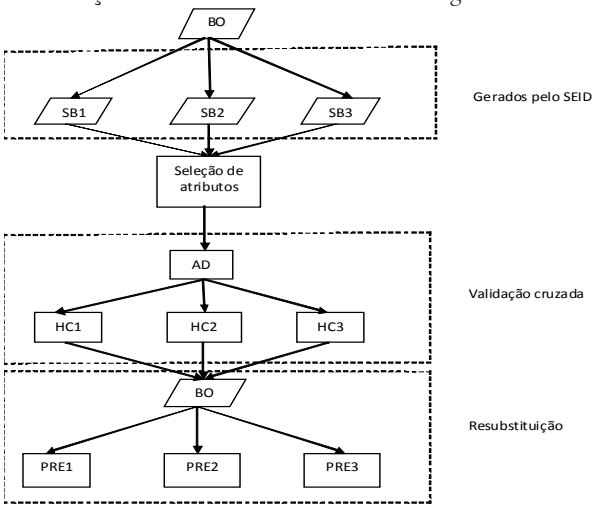
Apesar de parecer "óbvia" ou "sempre necessária", a seleção de atributos é uma opção metodológica de fundamental importância em aprendizagem de máquina, sendo frequentemente realizada como uma etapa de pré-processamento. O objetivo principal da seleção de atributos, para Meisel (1972, p. 162), é a obtenção de um espaço de características com baixa dimensionalidade, retenção de informações suficientes, aumento da separabilidade na função espaço, por exemplo, em diferentes categorias, removendo efeitos em razão das características ruidosas, e comparabilidade dos recursos entre os exemplos na mesma categoria. Os principais objetivos da seleção de atributos para previsão de insolvência, segundo Piramuthu (2006, p. 489), 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.

Neste trabalho, foram utilizadas duas abordagens de busca (WITTEN; FRANK; HALL, 2011, p. 293): seleção *forward* e seleção aleatória. Já para a avaliação do subconjunto de atributos selecionados, utilizou-se a abordagem encapsulada (w*rapper)*. Para Somol et al. (2005, p. 997), a abordagem w*rapper* deve ser preferida quando se trata de estudos sobre insolvência de empresas.

3.5 A ESTRATÉGIA DE PREDIÇÃO DE INSOLVÊNCIA COM SELEÇÃO DE ATRIBUTOS

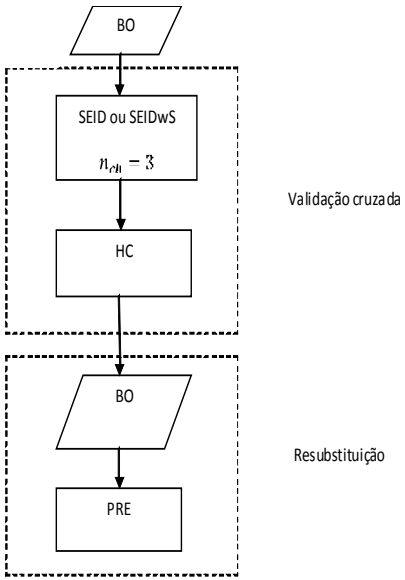
Apresenta-se, nesta seção, uma técnica para seleção de atributos a ser acoplada ao modelo de predição desenvolvido (SEID), completando a proposta deste trabalho. A ideia é considerar a aplicação dos métodos de seleção de forma individualizada nas bases que compõem o comitê, configurando o modelo proposto *Semi-Determinístic Ensemble Strategy for Imbalanced Data with attribute Selection* (SEIDwS). O fluxograma do modelo para a predição de insolvência com estratégia de seleção de atributos é apresentado nos Fluxogramas 1 e 2 a seguir. Deve-se ressaltar que o comitê de classificadores é composto por três sub-bases, nesse caso, a saber, SB1, SB2 e SB3.

Fluxograma 1 – Procedimentos para se chegar aos resultados após os balanceamentos e a seleção de atributos da base de dados original



Fonte: os autores.

Fluxograma 2 – Procedimento de classificação Com Seid ou Seidws



Fonte: os autores.

Nota: Legenda das siglas nos Fluxogramas 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.

Nota: As siglas na Fluxograma 2.2 são as mesmas da Fluxograma 2.1, entretanto, aqui, é sintetizado o processamento da estratégia apresentada.

4 **RESULTADOS**

Nesta seção, são apresentados os resultados das aplicações à base de dados aqui construída. Esta base se refere a variáveis obtidas em demonstrativos contábeis de empresas do setor de material básico.

4.1 COMPARAÇÃO DA BASE DE DADOS ORIGINAL COM A BASE BALANCEADA

Na Tabela 1 é feita uma comparação da base original com os resultados encontrados após a aplicação da estratégia SEIDwS utilizando o modelo *wrapper*.

Tabela 1 – Comparação dos resultados

	Base original			**SEIDwS**		
Classe	**MC**	**F**	**AUC**	**MC**	**F**	**AUC**
I	7 3	0,873	0,895	10 0	0,983	0,981
S	3 177	0,938	0,895	2 178	0,957	0,981

Fonte: os autores.

Pela Tabela 1, a partir da análise da MC, F e AUC, evidencia-se que aquelas empresas classificadas como insolventes (I) obtiveram, com a aplicação do SEIDwS, um resultado bem eficiente, com 10 das 10 instâncias classificadas corretamente, um valor de F bem considerável, 0,983, e um AUC que pode ser considerado bem eficiente, 0,981.

Dessa forma, conclui-se que o balanceamento com a seleção de atributos e um comitê de classificadores (SEIDwS) melhoram a capacidade de caracterização das empresas classificadas como insolventes, os resultados da MC, F e AUC evidenciam esses ganhos (BASE ORIGINAL x SEIDwS).

4.2 COMPARAÇÃO DAS VARIÁVEIS SELECIONADAS NAS BASES BALANCEADAS ANTES E PÓS CRISE *SUBPRIME*

Nesta subseção, é aplicada a estratégia SEIDwS desenvolvida para a predição de insolvências em empresas. Na prática, a aplicação completa do SEIDwS é obtida

com o uso da votação majoritária (LI; JIE, 2009) em relação aos resultados dos modelos das sub-bases obtidas na definição da instância que está sendo avaliada. Dessa forma, as sub-bases passam a representar um comitê de classificadores, conforme descrito anteriormente.

Das 23 variáveis totais, sete foram selecionadas: EOCpOT, EOAT, GAF, MB, EBITDA, MO, TERFIN no período entre os anos de 1996 e 2006 (HORTA et al., 2012). Já no período que compreende os anos 2007 até 2012, as variáveis selecionadas foram: EOCpOT, GAF, LI, LC e MB. Os resultados obtidos são mostrados na Tabela 2, apresentada a seguir.

Tabela 2 – Resultados referentes às variáveis selecionadas de acordo com os períodos pré e pós crise *subprime*

Período de 1996 a 2006	**Período de 2007 a 2012**
EOCpOT	EOCpOT
GAF	GAF
MB	MB
EBTIDA	LI
MO	LC
EOAT	-
TERFIN	-

Fonte: os autores.

Pelos resultados apresentados na Tabela 2 ficou demonstrada a mudança de algumas das variáveis selecionadas nos dois períodos estudados, foram selecionadas duas variáveis referentes à liquidez e não foram selecionadas algumas das variáveis representativas do desempenho operacional das entidades, com isso, pode-se entender da prevalência do financeiro em detrimento do operacional após a crise do *subprime*.

5 ANÁLISES DOS RESULTADOS, CONCLUSÕES E FUTUROS ESTUDOS

Observou-se no estudo que prevaleceram aquelas variáveis originadas do Balanço Patrimonial, a única exceção foi a variável MB; isso evidencia a importância patrimonial e financeira dessas entidades durante o período estudado, 2007 a 2012 (Tabela 2).

Dessas variáveis selecionadas, três se repetem (EOCpOT, GAF e MB); nos dois períodos sob análise evidencia-se a importância da adoção contínua de uma política de endividamento em níveis prudentes e gerenciáveis, em especial quanto ao seu endividamento oneroso (EOCpOT e GAF).

Já a presença da variável MB pode mostrar a importância permanente do grau de eficiência da gerência da empresa para usar materiais e mão de obra no processo de produção, refletindo a relação entre preços, quantidade produzida e custos.

No que se refere ao período antes crise *subprime* (1996-2006), mais duas variáveis relacionadas aos aspectos operacionais estavam presentes, MO e EBTIDA (Tabela 2), o que pode levar à conclusão de que o desempenho operacional (vendas e custo) perdeu relevância após a crise de 2007 para o financeiro (liquidez). Essas duas variáveis representam a eficiência operacional ainda não distorcida pelos cálculos de financiamento e impostos.

As outras variáveis selecionadas (LI e LC) representam a importância da liquidez para essas empresas nesse período (pós-crise *subprime*). Ganhou importância uma melhor adequação da estrutura financeira nas entidades, sobretudo a da liquidez, evidenciando a necessidade de adaptações às mudanças ocorridas no ambiente econômico em que operavam (queda no nível de liquidez no mercado). Adquire relevância a relação com os credores em detrimento ao dos gestores (operacional) no período anterior, pois tais credores estão interessados em financiar as necessidades de um negócio sucedido, que funcionará como esperado e terão os seus recursos com retorno garantidos.

Ao mesmo tempo, eles têm que considerar as possíveis consequências negativas da inadimplência e da liquidação. No período pós-crise *subprime*, a proteção de que dispõem os credores se centra no crédito de curto prazo concedido para uma empresa financiar suas operações. Os credores consideram os ativos atuais que podem ser prontamente convertidos em caixa, na suposição de que eles representariam uma segurança imediata contra a inadimplência.

A presença da variável LI na seleção de atributos da base de dados pós-crise *subprime* evidencia cada vez mais a importância da necessária capacidade da entidade de pagar suas obrigações de curto prazo valendo-se de suas disponibilidades em caixa, bancos ou aplicações no mercado financeiro de curtíssimo prazo, objetivando que tais empresas obtenham sucesso na sua continuidade.

Comparando as variáveis selecionadas, apresentadas na Tabela 2, pode-se inferir que houve uma necessidade de mudança de estratégia gerencial relevante após a crise do *subprime* nas empresas brasileiras do setor de material básico; saiu do foco operacional e foi para o foco financeiro ou do credor. Ter uma relação "mais amigável" com os credores ganhou dimensão nesse período, cuidar da capacidade de pagamento ganhou mais importância para aquelas empresas que obtiveram o

sucesso da continuidade. Tais fatores continuam sendo preponderantes nas empresas atualmente, talvez até com mais preponderância em relação ao período estudado.

Esta pesquisa apresentou e aplicou uma estratégia para solucionar um problema pouco estudado em modelagens para descontinuidade de empresas – o desequilíbrio entre as classes de empresas classificadas como solventes e as empresas classificadas como insolventes. Na maioria das pesquisas existentes a amostra estudada é uma *paired sample*, ou seja, composta com número igual de empresas solventes e insolventes. Essa paridade entre as classes de empresas representa mal a realidade do ambiente econômico, distorcendo a utilidade da amostra e, comprovadamente, priorizando a classificação correta somente para as empresas solventes.

Durante algum tempo os dados contábeis, no Brasil, foram tratados com desconfiança, de modo que pesquisas e conclusões como as presentes servem para reiterar a conveniência daquela utilização para a análise da evolução econômica de empresas em nosso País confirmando, assim, o valor preditivo presente na informação contábil-financeira.

Outra contribuição que pode ser considerada deste trabalho diz respeito ao uso exclusivo de dados originados de demonstrativos contábeis de empresas pertencentes a somente um setor econômico, com isso, fica bem mais relacionado os motivos da insolvência daquelas empresas, não havendo influências de variáveis mais significativas para empresas de outros setores econômicos.

Esta pesquisa evidenciou mudanças geradas pelo novo ambiente macroeconômico das empresas (crise do *subprime*) pertencentes ao setor econômico de material básico, durante o período estudado, gerando necessidades de readequações gerenciais, prevalecendo as questões de liquidez visando manter a sua continuidade.

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. Com isso, o SEIDwS deverá melhorar, ainda mais, a classificação daquelas empresas que podem vir a se tornarem insolventes, além de discriminar a importância dessas variáveis.

REFERÊNCIAS

ALTMAN, E. I.; BAIDYA, T. K. N.; DIAS, L. M. R. Previsão de problemas financeiros em empresas. **Revista de Administração de Empresas**, v. 19, p. 17-28, jan./mar. 1979.

ALTMAN, E. I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. **Journal of Finance**, v. 23, n. 4, p. 589-609, 1968.

ALTMAN, E. I.; 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.

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, p. 29-54, 1977.

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.

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.

BRAGA NETO, U. et al. Is cross-validation better than resubstitution for ranking genes? **Bioinformatics**, v. 20, n. 2, p. 253-258, 2004.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123-140, 1996.

CANBAS, S.; CABUK, A.; KILIC, S. B. Prediction of commercial bank failure via multivariate statistical analysis of financial structure: the turkish case**. Journal of Operational Research**, v. 166, p. 528-546, 2005.

CHAWLA, N. V. et al. Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002.

CHAWLA, N. V.; JAPKOWICZ, N.; KOLCZ, A. Editorial: special issue on learning from imbalanced datasets. **SIGKDD Explorations**, v. 6, n. 1, p. 1-6, 2004.

COSTA, F. M. da; REIS, J. S. dos; TEIXEIRA, A. M. C. **Repec**, Brasília, DF, v. 6, n. 2, p. 141-153, abr./jun. 2012.

CROUHY, M. G.; JARROW, R. A.; TURNBULL, S. M. **The subprime credit crisis of 07**: working paper. New York: Cornell University 2008.

ELIZABETSKY, R. Um **modelo matemático para decisão no banco comercial**. 1976. Dissertação (Mestrado em Engenharia de Produção)– Universidade de São Paulo, São Paulo, 1976.

FÁVERO, L. P. et al. **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

GARY M. W. Mining with rarity: a unifying framework. **Sigkdd Explorations**, v. 6, n. 1, p. 7-19, 2004.

GESTEL, T. V.; BAESENS, B.; MARTENS, D. From linear to non-linear kernel based classifiers for bankruptcy prediction. **Neurocomputing**, v. 73, p. 2955-2970, 2010.

GNEDENKO, B. V. **A teoria da probabilidade**. Rio de Janeiro: Ciência Moderna, 2008.

GUJARATI, D. N. **Econometria básica**. 4. ed. Rio de Janeiro: Elsevier, 2006.

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, p. 205-213, 2008.

HORTA, R. A. M. et al. Previsão de insolvência com seleção de atributos no setor de material básico: uma metodologia para balanceamento de bases de dados. In: CONGRESSO NACIONAL DE ADMINISTRAÇÃO E CIÊNCIAS CONTÁBEIS, 3., 2012, Rio de Janeiro. **Anais**... Rio de Janeiro, 2012.

HORTA, R. A. M. **Uma metodologia de mineração de dados para a previsão de insolvência de empresas brasileiras de capital aberto**. 2010. 152 p. Tese (Doutorado em Engenharia Civil)–Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

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, p. 847-856, 2007.

HUNG, C.; CHEN, J.-H. A selective ensemble based on expected probabilities for bankruptcy prediction. **Expert systems with applications**, v. 36, p. 5297-5309, 2009.

IUDÍCIBUS, S. de. **Análise de balanços**. 10. ed. São Paulo: Atlas, 2009.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: a systematic study. **Intelligent Data Analysis**, v. 6, p. 429-449, 2002.

KANITZ, S. C. **Como prever falências**. *São Paulo: Mc* Graw-Hill do Brasil, 1978.

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

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

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

MATIAS, A. B. **Contribuição às técnicas de análise financeira**: um modelo de concessão de crédito. 1978. 82 p. Trabalho de Conclusão de Curso (Graduação em Economia e Administração)–Universidade de São Paulo, São Paulo, 1978.

MEISEL, W. S. **Computer-oriented approaches to pattern recognition**. New York: Academic Press, 1978.

MIN, S.-H.; LEE, J.; HAN, I. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. **Expert Systems with Applications**, v. 31, p. 652-660, Oct. 2006.

MOROZINI, J. F.; OLINQUEVITCH, J. L.; HEIN, N. Seleção de índices na análise de balanços: uma aplicação da técnica estatística 'ACP'. **Revista Contabilidade e Finanças USP**, São Paulo, v. 2, n. 41, maio/ago. 2006.

MOURA, G. D. de et al. Relação entre ativos intangíveis e governança corporativa em companhias abertas listadas na BM&FBovespa. **Revista de Contabilidade e Controladoria**, Curitiba, v. 5, n. 1, p. 120-138, jan./abr. 2013.

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

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

PENMAN, S. **Análise das demonstrações financeiras e security valuation**. Rio de Janeiro: Elsevier, 2013.

PINDYCK, R. S.; RUBINFELD, D. L. **Econometria**: modelos e previsões. 4. ed. Rio de Janeiro: Elsevier, 2004.

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

RAVI, V. et al. Ravi. Soft computing system for bank performance predition. *Applied Soft Computing*, v. 8, p. 305-315, jan. 2008.

SANVICENTE, A. Z.; MINARDI, A. M. A. F. **Identificação de indicadores contábeis significativos para previsão de concordata de empresas**. 2005. Disponível: <http://www.risktech.br/artigos/artigos técnicos/index.html>. Acesso em: 23 out. 2012.

SCHAPIRE, R. E. The strength of weak learnability. **Machine Learning**, v. 5, p. 197-227, 1990.

SHIN, K.-S.; LEE, Y.-J.; KIM, H.-J. An application of support vector machines in bankruptcy prediction model. **Expert Systems with Applications**, v. 28, p. 127-135, jan. 2005.

SILVA BRITO, G. A.; ASSAF NETO, A.; CORRAR, L. J. 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, set./dez. 2009.

SOMOL, P. et al. Filter *versus* wrapper-based feature selection for credit scoring. **International Journal of Intelligent Systems**, v. 20, n. 10, p. 985-999, 2005.

SUN, J. et al. 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, Feb. 2014.

TSAI, C. F. Feature selection in bankruptcy prediction. **Knowledge-Based Systems**, v. 22, p. 120-127, 2009.

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

UNITED STATES OF AMERICA. The financial crisis inquiry report. **Public Affairs**, p. 15, 2011.

WEISS, G. M.; MCCARTHY, K.; BIBI, Z. cost-sensitive learning vs. sampling: which is best for handling unbalanced classes with unequal error costs? **Proceedings of the 2007 International Conference on Data Mining**, New York, p. 35-41, 2007.

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

WITTEN, I. H.; FRANK, E.; HALL, M. A. Data mining: practical machine learning tools and techniques. In: KAUFMANN, M. **Series in data management systems**. 3. ed. 2011.

WOLK, H. I.; DODD, J. L.; ROZYCKI, J. J. Accounting theory: conceptual issues in a political and economic environment. 8. ed. London: Sage Publications, 2013.

YEH, C.-C.; CHI, D.-J.; LIN, Y.-R. Going-concern prediction using hybrid random forests and rough set approach. **Information Sciences**, v. 254, p. 98-100, Jan. 2014.

ZHOU, L. Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods. **Knowledge-Based Systems**, v. 41, p. 16-25, Mar. 2013.

AGRADECIMENTOS

Agradecemos à Fapemig pelo apoio concedido à pesquisa APQ 00916/12.

APÊNDICE A – VARIÁVEIS CONTÁBEIS COLETADAS

Liquidez corrente (LC), Liquidez seca (LS), Liquidez Imediata (LI), Liquidez Geral (LG), Endividamento Oneroso sobre Patrimônio Líquido (EOPL), Endividamento Total sobre o Patrimônio Líquido (EOAT), Endividamento Oneroso de Curto Prazo sobre Ativo Total (EOCpOT), Grau de Alavancagem Financeira (GAF), Imobilizado dos Recursos Permanentes (IMCP), Margem Bruta (MB), Margem Operacional (MO), Margem 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), Prazo médio de estocagem de matéria-prima (PME), Prazo Médio de Fabricação (PMF), Prazo médio de venda (PMV).

APÊNDICE B – MÉTRICAS DE AVALIAÇÃO

1 MATRIZ DE CONFUSÃO

Os diferentes tipos de erros e acertos realizados por um classificador podem ser sintetizados em uma matriz de confusão. Na Tabela 3, é mostrada uma matriz de confusão para um problema que possui duas classes rotuladas como positiva e negativa.

Tabela 3 – Matriz de confusão para duas classes de problemas

	Predição Positiva A	**Predição Negativa B**
Classe Positiva	Verdadeiro Positivo (TP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: os autores.

2 ÁREA SOB A CURVA ROC (AUC)

Por definição, uma curva ROC é um gráfico bidimensional em que o eixo horizontal representa a taxa de erro da classe negativa (1-*Spec*) e no eixo vertical os valores de sensitividade. O desempenho de um classificador é medido pela área sob a curva ROC (AUC).

Pontos na diagonal representam classificadores aleatórios, de acordo com a probabilidade a priori de cada classe. Acima (abaixo) da diagonal encontram-se classificadores com desempenho melhor (pior) que o classificador aleatório.

3 MEDIDA F

Para se entender a métrica Medida F é importante conhecer a definição de sua composição. *Recall* e *Precision* ou precisão são duas métricas amplamente usadas em aplicações onde a detecção bem sucedida de uma classe é considerada mais significativa do que outras classes. Uma definição formal dessas métricas é apresentada a seguir:

Precisão, p = (TP + FP)/FP
Recall, r = TP / (TP + FP)

A precisão determina a fração de registros que realmente acabam sendo positivos no grupo que o classificador declarou como classe positiva. Quanto maior a precisão, menor o número de erros positivos no grupo que o classificador declarou como classe positiva. O *recall* mede a fração de exemplos positivos previstos corretamente pelo classificador. Classificadores com valores altos de *recall* têm poucos exemplos positivos mal classificados como a classe negativa. Na verdade, o valor de *recall* é equivalente à taxa de positivos verdadeiros.

O desafio dos algoritmos de classificação é elaborar um modelo que maximize tanto a precisão quanto o *recall*.

A métrica entre precisão e *recall* é concedida pelas fórmulas a seguir:

$$Medida\ F = \frac{2\ x\ recall\ x\ precisão}{recalll + precisão} = \frac{2xTP}{2xTP + P\quad + N}$$

Como citar este artigo

ABNT

HORTA, Rui Américo Mathiasi et al. Descontinuidade de empresas brasileiras do setor de material básico: no período compreendido pré e pós a crise do subprime. **RACE**, Revista de Administração, Contabilidade e Economia, Joaçaba: Ed. Unoesc, v. 14, n. 1, p. 171-196, jan./abr. 2015. Disponível em: <http://editora.unoesc.edu.br/index.php/race>. Acesso em: dia/mês/ano.

APA

Horta,R. A. M., Alves, F. J. dos S., Borges, C. C. H., & Rodrigues, A. (2015). Descontinuidade de empresas brasileiras do setor de material básico: no período compreendido pré e pós a crise do subprime. *RACE, Revista de Administração, Contabilidade e Economia, 14*(1), 171-196.Recuperado em dia/mês/ano, de http://editora.unoesc.edu.br/index.php/race