REVISTA AMBIENTE CONTÁBIL
http://www.periodicos.ufrn.br/ojs/index.php/ambiente
http://www.ojs.ccsa.ufrn.br/index.php/contabil
http://www.atena.org.br/revista/ojs-2.2.3-06/index.php/Ambiente
ISSN 2176-9036
Artigo recebido em: 31.07.2013. Revisado por pares em: 28.10.2103. Reformulado em: 01.12.2013. Avaliado pelo sistema double blind review.

DESCONTINUIDADE DE EMPRESAS BRASILEIRAS DO SETOR DE CONSUMO CÍCLICO: UMA METODOLOGIA PARA BALANCEAMENTO DE BASE DE DADOS UTILIZANDO TÉCNICAS DE DATA MINING

DISCONTINUANCE OF BRAZILIAN COMPANIES SECTOR CONSUMPTION CYCLIC: A METHODOLOGY FOR BALANCING OF DATABASE USING DATA MINING TECHNIQUES

INTERRUPCIÓN DEL BRASILEÑO SECTOR EMPRESARIAL DE CONSUMO CÍCLICO: UNA METODOLOGÍA PARA EQUILIBRAR LA BASE DE DATOS CON TÉCNICAS DE MINERÍA DE DATOS

Autores

Rui Américo Mathiasi Horta
Professor Doutor - Universidade Federal de Juiz de Fora/Depto Finanças e Controladoria. Endereço: Rua José Lourenço Kelmer, s/n - Bairro São Pedro - Juiz de Fora/ MG – Brasil. Telefone (32) 2102-3521.
E-mail: rui.horta@ufjf.edu.br

Carlos Cristiano Hasenclever Borges
Professor Doutor - Universidade Federal de Juiz de Fora/Depto Ciência da Computação. Endereço: Rua José Lourenço Kelmer, s/n - Bairro São Pedro - Juiz de Fora/ MG – Brasil. Telefone (32) 2102-3327.
E-mail: cchb@lncc.br

Marcelino José Jorge
Professor Doutor - Instituto de Pesquisa Clínica Evandro Chagas – IPEC/FIOCRUZ. Av. Brasil, 4365 – Manguinhos - Rio de Janeiro/RJ – Brasil. Telefone (21) 3865-9595.
E-mail: marcelino.jorge@ipec.fiocruz.br

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

RESUMO

Descontinuidade de empresas é um tema que cada vez mais vem sendo estudado no campo da contabilidade e finanças devido ao considerável número de partes do tecido social afetadas pelo fracasso corporativo de uma entidade. Bancos, investidores, governos, auditores, gerentes, fornecedores, empregados e muitos outros têm grandes interesses na acurácia da previsão de insolvência de uma companhia. No Brasil os estudos sobre o tema ainda sofrem o efeito de estarem disponíveis apenas em bases de dados com dimensões reduzidas, quase sempre devido à qualidade dos dados. Mas 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 é propor um novo procedimento para balanceamento da base de dados em problemas de previsão de insolvência com (etapa de) seleção de atributos. Foi então construída uma estratégia de *data mining* com a dupla virtude de selecionar atributos e de resolver o problema do desequilíbrio. A base de dados foi originada de demonstrativos contábeis de empresas brasileiras do setor econômico de consumo cíclico, listadas na BOVESPA entre os anos de 1996 e 2011. Os resultados obtidos e as validações realizadas evidenciam o sucesso da estratégia proposta, melhorando a capacidade do modelo de previsão na classificação das empresas pertencentes à classe das insolventes e, assim consolidando-a como bastante competitiva com outras estratégias apresentadas na literatura específica. Dada a natureza do exercício empírico, fica evidente que a vantagem do novo procedimento não depende do setor estudado.

Palavras-chave: Descontinuidade de empresas. Seleção de variáveis contábeis. Balanceamento de base de dados. *Data mining*. Setor de consumo cíclico – Brasil.

ABSTRACT

Discontinuity of companies is an issue that increasingly is being studied in the field of accounting and finance due to the considerable number of parts of the social fabric affected by the failure of a corporate entity. Banks, investors, governments, auditors, managers, suppliers, employees and many others have great interests in the accuracy of prediction of insolvency of a company. In Brazil, studies on the subject are still suffering the effects of being available only in databases with reduced dimensions, mostly due to data quality. But there is little studied issues in predictive modeling of insolvency. The balance or imbalance of data on insolvency is one of those issues in economic environments typical number of companies classified as solvent is much higher than those classified as insolvent. The aim of this study is to propose a new procedure for balancing of database problems in insolvency prediction with (step) feature selection. Was then constructed a strategy of data mining with the double virtue of selecting attributes and solve the problem of the imbalance. The database was derived from financial statements of Brazilian companies in the consumer cyclical economic sector, listed on the BOVESPA between the years 1996 and 2011. The results and validations performed demonstrate the success of the proposed strategy, improving the ability of the prediction model for the classification of companies belonging to the class of insolvent and thus consolidating it as quite competitive with other strategies presented in the specific literature. Given the nature of the empirical exercise, it is clear that the advantage of the new procedure does not depend on the studied sector.

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

Keywords: Discontinuity of companies; Brazilian companies from the consumer discretionary sector, accounting variables, Data mining; Balancing database.

RESUMEN

La discontinuidad de las empresas es un tema que se está estudiando cada vez más en el campo de la contabilidad y las finanzas, debido al número considerable de partes del tejido social afectado por la falla de una entidad corporativa. Bancos, inversores, gobiernos, auditores, gerentes, proveedores, empleados y muchos otros tienen grandes intereses en la exactitud de la predicción de la insolvencia de una empresa. En Brasil, los estudios sobre el tema siguen sufriendo los efectos de estar disponible sólo en bases de datos de dimensiones reducidas, casi siempre debido a la calidad de los datos. Pero hay pocos temas estudiados en modelos de predicción de la insolvencia. El desequilibrio o insolvencia datos desequilibrio es uno de esos problemas, en entornos económicos típicos del número de empresas clasificadas como disolventes es mucho mayor que los clasificados como insolvente. El objetivo de este estudio es proponer un nuevo procedimiento para el equilibrio de los problemas de base de datos en la predicción de la insolvencia con (paso) de selección de características. A continuación, se construyó una estrategia de minería de datos con la doble virtud de la selección de atributos y resolver el problema de desequilibrio. La base de datos se deriva de los estados financieros de las empresas brasileñas en el sector de la economía de consumo cíclico, que se enumeran en la Bovespa entre los años 1996 y 2011. Los resultados y validaciones demuestran el éxito de la estrategia propuesta, la mejora de la capacidad del modelo de pronóstico para la clasificación de las empresas que pertenecen a la clase de insolvencia y por lo tanto la consolidación como bastante competitivo con otros enfoques sugeridos en la literatura. Dada la naturaleza del ejercicio empírico, está claro que la ventaja del nuevo procedimiento no depende de el sector estudiado.

Palabras clave: Negocios discontinuados. La selección de las variables de contabilidad. Equilibrio de base de datos. La minería de datos. Sector Consumidor Discrecional - Brasil.

1 INTRODUÇÃO

Cada vez mais o desenvolvimento de estudos sobre modelagem para previsão de insolvência vem adquirindo importância nas áreas acadêmicas e empresarial relativas a Contabilidade e Finanças. 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 à situação, impedindo a ocorrência de grandes custos sociais e financeiros.

Vários fatores têm concorrido para o aumento quantitativo dos estudos sobre o tema. Por exemplo, em vários países a maioria das estatísticas sobre falências mostrou significativo crescimento quantitativo e qualitativo. 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. Cresceu, também, a cautela associada à implementação, em vários países, de normas internacionais de contabilidade e finanças, tais como Basiléia II e III, Solvência II, Sarbanes-Oxley e IFRS.

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, o desequilíbrio entre as classes (BALCAEN; OOGHE, 2006; RAVI; KURNIAWAN; THAI; KUMAR, 2008; TSAI; WU, 2008; NANNI; LUMINI, 2009; VERIKAS; KALSYTE;

BACAUSKIENE; GELZINIS, 2010; GESTEL; BAESENS; MARTENS, 2010; ZHOU L., 2013).

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. Cabe reconhecer imediatamente que esta situação é natural, porque, normalmente, em qualquer sociedade, o número de insolventes é muito inferior ao de solventes, independentemente do período que se analisa. Por isso mesmo, a situação é, também, muito freqüente e, nessa medida, 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 do chamado problema do desbalanceamento em classificação de dados pode ser considerada relativamente nova, dentre "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 tão diversas quanto negócios, indústria, linguística, bioinformática entre muitas outras" (CHAWLA; JAPKOWICZ; KOLZ, 2004, p. 1).

Valendo-se de dados contábeis este estudo tem por objetivo propor um novo procedimento para balanceamento da base de dados em problemas de previsão de insolvência com (etapa de) seleção de atributos. A metodologia proposta é empiricamente ilustrada e avaliada quando são utilizados dados obtidos em demonstrativos contábeis de empresas classificadas na BOVESPA (Bolsa de Valores de São Paulo), pertencentes ao setor econômico de consumo cíclico. Mostra-se que a aplicação da metodologia proposta melhora, em diversos sentidos, a capacidade de classificar aquelas empresas do setor que podem vir a se tornar insolventes. Dada à natureza do exercício empírico, fica evidente que a vantagem do novo procedimento não depende do setor estudado e nem da área.

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).

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. FUNDAMENTAÇÃO TEÓRICA

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; HALDEMAN; 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 utilizado 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; 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 aprendizagem de máquina começaram a ser aplicadas visando à detecção de insolvência. SHIN; LEE; 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 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. 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 ANNs, as múltiplas abordagens de LDA e a RL. DING; SONG; ZEN, (2008) desenvolveram um modelo de previsão de insolvência em 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, visando aumentar a eficácia da predição, também desenvolveram metodologias especificas 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, WEST; DELLANA; QIAN (2005) investigaram três estratégias de combinação de classificadores para aplicação em decisões financeiras, incluindo previsão de insolvência. HUNG; CHEN (2009) aplicaram um modelo de probabilidades híbridas, baseado em comitê de classificadores, para previsão de insolvência utilizando votação majoritária e votação ponderada. HUANG *et al*., (2007) investigaram três estratégias para construção de modelos híbridos, baseados no SVM, para *credit scoring* e compararam suas performances com ANNs, algoritmo genético (GA) e árvore de decisão (AD). 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, YU e LAY, (2008) utilizam ANNs 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.

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 – p. ex., 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; HEIN (2006) utilizam análise dos componentes principais para combinar os principais índices dentre os selecionados para o

estudo. SILVA BRITO; ASSAF NETO; 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.

3 METODOLOGIA DA PESQUISA

Este estudo tem como objetivo propor uma estratégia que permite o balanceamento da base de dados em conjunto com um procedimento especifico visando uma seleção de atributos/variáveis mais relevantes originados de demonstrativos contábeis para uma amostra de empresas pertencentes ao setor econômico de consumo cíclico, visando melhorar a capacidade de caracterizar aquelas empresas que podem vir a tornar-se insolventes. A seguir serão 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 no BOVESPA 23 indicadores contábeis anuais das empresas do setor de consumo cíclico, classificadas de acordo com grupos de índices contábeis-financeiros: liquidez, endividamento, rentabilidade e ciclo operacional (ver o Anexo). Esses grupos de índices contábeis-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". GLANTZ (2007, p. 92) afirma que "os índices servem como medidas relativas ou interações entre números e são usados para, (i) esclarecer a relação entre contas ou itens que constam dos relatórios financeiros (análise estrutural), (ii) comparar o desempenho do tomador aos níveis históricos (análise de séries temporais) e (iii) comparar seu desempenho por meio de *benchmarks* (referencias de excelência) ou médias do setor (análise de *cross-section*)". Na visão de WORK *et. al*. (2013, p. 304) "variáveis contábeis são muito usadas para discriminar empresas que apresentam tendências de se tornarem insolventes daquelas solventes".

Vale aqui evidenciar a importância e os principais motivos da escolha de um conjunto de empresas pertencentes a um mesmo setor econômico. Para IUDÍCIBUS (2008, 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. Empresas que apresentem índices bem distintos das médias setoriais devem evidenciar situações em que haja anomalias, especialmente econômicas ou financeiras". SILVA (2006, p. 190) afirma que "é possível comparar o índice financeiro de uma empresa com o mesmo índice relativo a outras empresas de mesma atividade econômica, para sabermos como está a empresa em relação às principais concorrentes".

O setor de consumo cíclico é composto por empresas que dependem de um determinado ciclo da economia para obterem ganhos mais expressivos e é composto por companhias do setor de comércio/varejo, hotéis, tecidos, calçados, lazer etc. No período estudado, segundo a BOVESPA, empresas deste setor representavam em média 17,8% do total, o maior dos setores.

Neste setor, as empresas se caracterizam por apresentar os valores de seus ativos imobilizados proporcionalmente mais baixos em relação aos seus ativos circulantes, e também por apresentarem valores pouco significativos em seus ativos. Empresas deste setor, durante o período estudado, passaram por várias alterações em seus ambientes macro e micro

econômicos influenciando em suas estratégias gerenciais e operacionais refletindo em seus demonstrativos contábeis.

Na montagem da base de dados cada uma das empresas havia sido classificada como concordatária, em recuperação judicial ou falida na BOVESPA, durante o período de 1996 a 2011. 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). Sempre que possível, o tamanho do ativo dessas empresas adicionadas era compatível com o de falidas correspondentes. Além disso, as adicionadas, todas elas listadas na BOVESPA, buscando respeitar, também, localização geográfica e idade. 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 (Lei dos grandes números, GNEDENKO 2008, p. 297) além de representar melhor a realidade econômica.

Foi utilizada análise de dados em painel, pois segundo PINDYCK e RUBINFELD (2004) e GUJARATI (2006) *apud* FÁVERO *et. al.* (2009, p.382) as principais características da análise de dados em painel são: (i) maior número de observações para se trabalhar, com consequente aumento do número de graus de liberdade e eficiência dos parâmetros, (ii) redução de problemas de multicolinearidade de variáveis explicativas e (iii) existência da dinâmica intertemporal. Em modelos de previsão de insolvência elaborados com dados em painel cada empresa fornece dados contínuos durante períodos (anos) (HUNG e CHEN, 2009, p. 5301). Na análise de dados em painel pela facilidade de acesso a uma maior quantidade de dados obtidos nos demonstrativos contábeis, devido à criação de instância em quantidade bem maior do que o número de empresas, as aplicações e os estudos nestas bases acabam por apresentar melhores resultados. Daí a sua preferência e utilização (BALCAEN e 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 tem algumas hipóteses restritivas, como a linearidade, normalidade e independência dos preditores ou variáveis de entrada. Considerando que a violação dessas premissas ocorre com freqüência com o uso de dados financeiros (YEH; CHI; LIN, 2014, p.98), abordagens de mineração de dados (MD), tais como a árvore de decisão (AD) são menos vulneráveis a essas violações. Além disso, a MD tem como objetivo identificar novos, potencialmente úteis e compreensíveis correlações válidas e padrões nos dados. MD pode ser uma alternativa solução para problemas de classificação, uma vez que a MD demonstrou ter capacidade preditiva superior aos métodos estatísticos convencionais de previsão de continuidade (YEH; CHI; LIN, 2014, p.99).

Na base de dados há 50 instâncias representando as empresas insolventes e 474 representando as solventes do mesmo setor, na proporção de quase 10 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, a seguir foram obtidos o maior número possível de demonstrativos contábeis de empresas classificadas como solventes e que pertencessem ao setor econômico de consumo cíclico. 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 reduzida dimensão da amostra final se deve, principalmente, à não obrigatoriedade, para um grande número de empresas, de publicar demonstrativos contábeis. A base foi composta por dados referentes aos demonstrativos contábeis dos cinco anos anteriores ao ano

em que a empresa foi declarada insolvente. De acordo com ALTMAN; GIANCARLO; FRANCO, (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 dez anos, facilitando assim uma melhor caracterização dessas empresas. Pretendeu-se, também, (i) uma adequação ao ano (2005) em que ocorreu a mudança na Lei de Falências no Brasil, (ii) utilizar demonstrativos contábeis sem a influência da inflação e (iii) adequar a base de dados a um período de tempo com o ambiente econômico de muitas mudanças e transformações para as empresas brasileiras.

Das métricas de avaliação alternativas existentes para lidar com o problema do desequilíbrio de classes citadas por JOSHI, *et al*., 2001; KAUCK, 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 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 dos classificadores foram utilizadas validação cruzada e com resubstituição. 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 no mesmo conjunto de dados (BRAGA-NETO; HASHIMOTO; DOUGHERTY; NGUYEN; CARROLL, 2004). Também será 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. 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 BANCOS 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.

A maneira mais simples para amenizar o desequilíbrio entre instancias de cada classe 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 (b) 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.

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 de pouca eficácia na replicação de dados na base.

O algoritmo SMOTE, uma técnica bastante citada na literatura específica, servirá como comparativo com a metodologia aqui proposta.

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

3.3 UMA 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.

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

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, L., 1996) e *boosting* (SCHAPIRE, R.E., 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 dos objetivos 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 sub- bases diferentes para gerar os classificadores que formam o comitê. Desta forma, a partição não prejudica nem a representatividade dos dados da classe majoritária, que devem compor pelo menos uma base de dados do comitê, nem a dimensão 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.

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

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.

Vamos apresentar, agora, 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 BOVESPA sem distinção de setor (HORTA; DE LIMA; 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 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 pseudo-código do comitê de classificadores.

Figura 1 -Algoritmo SEID

Pseudo-código: *comitê de classificadores para base de dados desbalanceadas* (SEID)
início
Defina o número de classificadores base n_{cb}
Defina o número de instâncias para cada classe n_{ic}
% construção dos n_{cb} *classificadores base*
para i=1, n_{cb}
% classe minoritária
$Str_i \leftarrow Str_m$
% completar, quando necessário, aplicando um processo de bootstrap na classe minoritária
para j = $\#(Str_m) + 1$ *,*n_{ic}
$Str_i \leftarrow Str_i \cup$ *j-ésima instância obtida aplicando bootstrap na amostra* Str_m
fim
%classe majoritária
para j = $1, \quad \#(Str_M) / n_{cb}$
$Str_i \leftarrow Str_i \cup$ *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}

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

$Str_i \leftarrow Str_i \cup$ 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: elaborada pelo 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 (DM), sendo frequentemente realizada como uma etapa de pré-processamento. Os principais objetivos da seleção de atributos para previsão de insolvência, segundo PIRAMUTHU (2006, p. 489), são (i) o desenvolvimento de modelos compactos, (ii) o uso e refinamento do modelo de classificação ou predição para avaliação, e (iii) a identificação de índices financeiros relevantes.

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

Determinar 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, determinar o subconjunto é avaliar como ele interage com o algoritmo de aprendizado. Essa determinação pode ser obtida, basicamente, em duas abordagens principais - Filtro e Encapsulada (*Wrapper)* (KOHAVI; JOHN, 1997). A abordagem *Wrapper* foi utilizada neste artigo. Para SOMOL *et al*. (2005, 997) a abordagem *Wrapper* deve ser preferida quando se trata de estudos sobre insolvência de empresas.

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 a predição de insolvência com estratégia de seleção de atributos é apresentado nas figuras 3.1 e 3.2, abaixo. Deve-se ressaltar que o comitê de classificadores é composto por 3 sub-bases neste caso, a saber, SB1, SB2 e SB3.

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

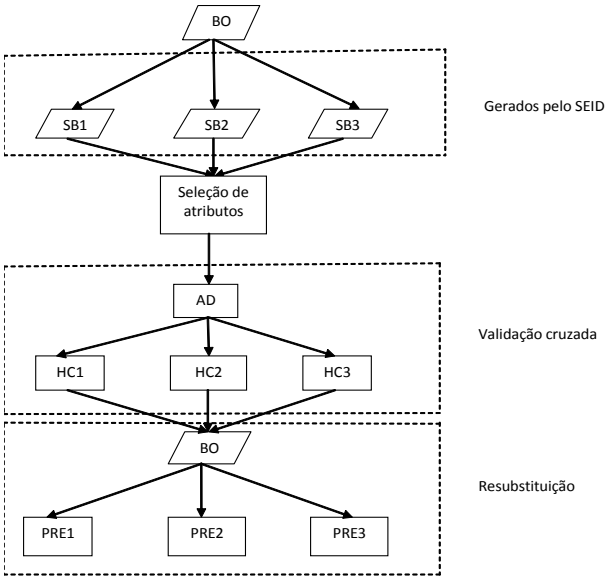

Figura 2.1 - Fluxograma referente aos procedimentos para se chegar aos resultados após os balanceamentos e a seleção de atributos da base de dados original.

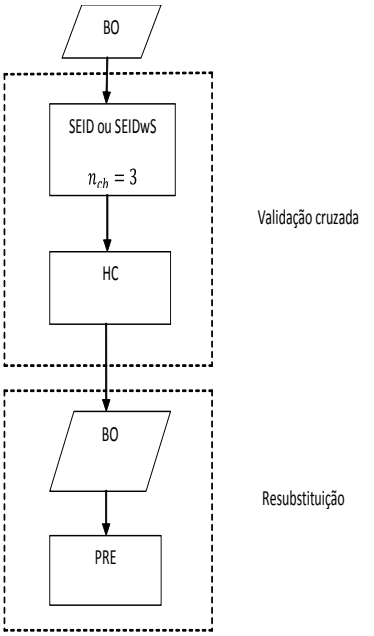

Figura 2.2 – Procedimento de classificação com o SEID ou SEIDwS.
As siglas na Figura 2.2 são as mesmas da Figura 2.1, entretanto aqui é sintetizado o processamento da estratégia apresentada.

Legenda das siglas nas 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.

3.5 VALIDAÇÃO DO ALGORITMO PROPOSTO

As Figuras 3.1 e 3.2 apresentam o funcionamento do SEIDwS. A base de dados original (BO) aplicando o SEIDwS é subdividida em três sub-bases por meio da técnica de seleção de atributos. Nessas sub-bases é feita a classificação, gerando três modelos (por validação cruzada) para serem testados na base de dados original (resubstituição). A seguir aplica-se a técnica da votação majoritária.

A validação do algoritmo proposto será realizada em duas etapas, visando atender dois objetivos - (i) testar os algoritmos aqui propostos em bases de dados diferentes daquelas aqui estudadas; (ii) comparar os resultados gerados pelo SEIDwS com outras pesquisas realizadas nesse tema.

O cumprimento da primeira etapa (i) consistiu em testar o algoritmo SEIDwS em três bases de dados originadas do Repositório UCI para Aprendizagem de Máquina - *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).

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

3.5.1 VALIDAÇÃO DO SEIDWS NAS BASES DO REPOSITÓRIO UCI PARA APRENDIZADO DE MÁQUINA

Nesta subseção são apresentados os resultados da validação do SEIDwS através das três bases do UCI, o procedimento é o mesmo apresentado na Figura 3.1. O classificador AD foi o utilizado e a sua escolha será justificada na Tabela 3.

A Tabela 1 apresenta os resultados dos testes do SEIDwS e do algoritmo SMOTE, neste algoritmo o parâmetro *k* (o número de vizinhos mais próximos) usado foi igual a 5. As bases de dados utilizadas do UCI foram as que normalmente são utilizadas para testes na previsão de insolvência, nelas as classes são discriminadas em empresas insolventes (INS) e solventes (SOL).

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

A abordagem seleção aleatória utilizou o algoritmo de busca *Genetic Selection* (GS). Em GS foram usados os valores do tamanho da população e do número das gerações iguais a 20, e a probabilidade de *crossover* e mutação igual a 0.6 e 0.033, respectivamente. Na abordagem filtro, as técnicas de avaliação de atributos (Witten e Frank, 2011, p.422) foram medidas de consistência (c*onsistency*) e o CFS (Seleção de atributos baseado em correlação). Na abordagem *wrapper* os algoritmos indutores foram RL, ANNs, SVM e AD. Os resultados apresentados nas tabelas seguintes, que obtiveram os melhores resultados, utilizaram GS, *wrapper* e AD.

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

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

Fonte: elaborada pelos autores.

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

Pelos resultados apresentadas na Tabela 1 o SEIDwS apresentou resultados bem promissores e competitivos com o SMOTE como estratégia de balanceamento quando testado nas bases de dados do UCI.

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

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: elaborada pelos 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 BOVESPA no período de 1996 a 2011. A essa base de dados original foi aplicada a estratégia aqui apresentada, denominada SEIDwS, que, após gerar três sub-bases, selecionou atributos priorizando os índices daquelas empresas classificadas como insolventes e não descartando os índices das empresas solventes no conjunto das sub-bases.

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

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

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

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. As duas aplicações descrevem, primeiro, resultados sem a estratégia SEIDwS (§ 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, medida F e AUC.

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

	RL				SVM				MLP				AD			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	41	9	0,811	0,921	31	19	0,738	0,813	41	9	0,877	0,923	42	8	0,884	0,942
S	10	464	0,98	0,921	3	471	0,977	0,813	3	471	0,987	0,923	3	471	0,988	0,942

Fonte: elaborada pelos autores.

Para a base de dados do setor econômico de empresas de consumo cíclico, em relação a matriz de confusão, a medida F e área AUC apresentaram pouca diferença entre os classificadores. Entretanto, MLP e AD foram aqueles que foram capazes de melhor classificar tanto as empresas solventes (S) quanto as empresas insolventes (I). 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 na seção 3.4 (sem a aplicação do SEIDwS) foram: EOCpOT, GAF, LC, MB.

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 3.1).

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

Nas aplicações com a abordagem *wrapper* (conforme Figura 3.1) 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				SB2				SB3			
Classe	MC		F	AUC	MC		F	AUC	MC		F	AUC
I	50	0	0,925	0,958	50	0	0,909	0,95	50	0	0,793	0,908
S	8	466	0,991	0,958	10	464	0,989	0,95	26	448	0,971	0,908

Fonte: elaborada pelos autores.

Os resultados evidenciam a influencia do balanceamento seguido da seleção de atributos com a abordagem *wrapper* refletindo num ganho de desempenho em relação a classificação sem o balanceamento e sem a aplicação de técnicas de seleção de atributos (Tabela 3).

Das 23 variáveis totais seis foram selecionadas pela abordagem *wrapper*. As variáveis selecionadas foram: EOCpOT, GAF, LC, MB, LI e LG.

Com a aplicação da estratégia SEIDwS foram selecionadas mais duas variáveis, LI e LG. Como a estratégia SEIDwS melhora a caracterização das empress potencialmente insolventes pode-se deduzir que nas empresas pertecentes ao setor econômico de consumo cíclico as variáveis que representam os desempenhos de liquidez dessas empresas são importantes para a discriminação das solventes e insolventes.

4.3 BALANCEAMENTO E SELEÇÃO DE CARACTERÍSTICAS PARA BASE DE DADOS COM VOTAÇÃO MAJORITÁRIA -SEIDWS

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

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

Os atributos selecionados continuam os mesmos para cada sub-base. Porém, a votação majoritária deve aumentar a robustez na predição obtida para as instâncias avaliadas. O procedimento foi exposto nas Figuras 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	42	8	0,884	0,942	50	0	0,947	0,991
S	3	471	0,988	0,942	9	465	0,99	0,991

Fonte: elaborada pelos autores.

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

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	42	8	0,884	0,942	50	0	0,947	0,991	49	1	0,986	0,993
S	3	471	0,988	0,942	9	465	0,99	0,991	3	471	0,986	0,993

Fonte: elaborada pelos autores.

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

No mesmo quadro a comparação do SEIDwS com o SMOTE pode ser evidenciado a melhor capacidade do SEIDwS de caracterizar aquelas empresas classificadas como insolventes (I) em relação ao SMOTE – MC e F. Somente na classificação geral (AUC) o SEIDwS obteve um resultado inferior em relação ao SMOTE (0,991 x 0,993). Diante desses resultados pode ser concluído da capacidade bem competitiva existente entre a estratégia aqui apresentada (SEIDwS) e o SMOTE, técnica bem referenciada na literatura específica.

5 CONCLUSÕES E FUTUROS ESTUDOS

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

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.

Implicações metodológicas:

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

Na comparação feita entre os resultados encontrados pela aplicação do SEIDwS e outras estratégias publicadas na literatura específica, os resultados foram bastante animadores. Na Tabela 2 está evidenciada a capacidade do SEIDwS em relação a outros trabalhos

Revista Ambiente Contábil – UFRN – Natal-RN. v. 6. n. 1, p. 99 – 121, jan./jun. 2014.

contemporâneos. Na grande maioria dos resultados obtidos o SEIDwS foi mais eficiente do que as outras estratégias, havendo ganhos na capacidade de classificar aquelas empresas pertencentes à classe das insolventes, o que era exatamente a situação que se queria melhorar.

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

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

Cabe ressaltar que o SEIDwS minora o efeito estocástico do modelo em relação ao SMOTE porque se utiliza dos dados da classe minoritária de forma mais intensa ou redundante, pois, busca-se um maior nível de acerto nesta classe Diante disso pode ser visto como uma contribuição metodológica às pesquisas sobre previsão de insolvência de empresas em bases desbalanceadas, tema ainda pouco explorado na literatura contábil no Brasil.

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

Implicações para a análise contábil:

Em relação às variáveis contábeis selecionadas, prevaleceram às originadas no Balanço Patrimonial, sobretudo aquelas que aferem a composição (estrutura) das fontes passivas de recursos das empresas (EOCpOT, GAF, LC). Depois da aplicação da estratégia SEIDwS somam-se àquelas as variáveis LI e LG, representativas da capacidade de liquidez das empresas superando em importância aquelas de estrutura.

Com a seleção de LI e LG pode-se inferir que, para a amostra estudada, a descontinuidade das empresas do setor de consumo cíclico se relaciona não somente a aspectos relativos à incapacidade da empresa de se endividar (EOCpOT, GAF), mas também a aspectos de seu desempenho de liquidez e solvência. As variáveis LI, LC e LG representam o desempenho e a capacidade da empresa manter sua capacidade de liquidez tanto no curto como no longo prazo (solvência), ou seja, sua eficiência financeira será determinada pelas estratégias adotadas pelas empresas visando “fortalecer” essas variáveis. Em outras palavras, na amostra estudada empresas podem tornar-se descontínuas não somente porque perdem a capacidade (financeira) de se endividarem, mas, sobretudo porque perdem a capacidade de manter sua liquidez e sua solvência. Apesar de potencialmente esperada antes do exercício, esta conclusão pode ser considerada outra contribuição específica deste trabalho às pesquisas sobre previsão de descontinuidade no Brasil.

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

A quarta contribuição que pode ser considerada deste trabalho diz respeito ao uso exclusivo de dados originados de demonstrativos contábeis de empresas pertencentes a

somente um setor econômico, com isso fica bem mais relacionado os motivos da insolvência daquelas empresas não havendo influências de variáveis mais significativas para empresas de outros setores econômicos.

Agradecemos a FAPEMIG pelo apoio concedido à pesquisa APQ 00916/12.

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.

_____; 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.

_____; 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.

_____; 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.

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.; 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.

BREIMAN, L. Bagging predictors. Machine Learning, vol. 24, pp. 123-140, 1996.

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.

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.

_____; JAPKOWICZ, N.; KOLCZ, A. Editorial: Special Issue on Learning from Imbalanced Datasets. ***SIGKDD Explorations***, v. 6, n. 1, p. 1-6, 2004.

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.

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.

FÁVERO, L. P.; BELFIORE, P.; DA SILVA, F. L.; CHAN, B. L. **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009 – 5ª reimpressão. 672 p.

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

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

GLANTZ, Morton. **Gerenciamento de riscos bancários**: introdução a uma ampla engenharia de crédito. Rio de Janeiro: Elsevier, 2007. 550 p.

GNEDENKO, Boris Vladimirovich. ***A*** **teoria da probabilidade.** Rio de Janeiro: Editora Ciência Moderna Ltda, 2008. 696 p.

GUJARATI, D. N. **Econometria básica**. 4. ed. Rio de Janeiro: Campus Elsevier, 2006. 819 p.

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 Administração e Finanças 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, vol 40, Spain, 2008, p. 205–213.

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.

HUANG, C. L., CHEN M. C., WANG, C.J. Credit scoring with a data mining approach based on support vector machines. ***Expert Systems with Applications****,* v. 33, Issue 4, p. 847-856, 2007.

HUNG, Chihli; CHEN, Jing-Hong. A selective ensemble based on expected probabilities for bankruptcy prediction. ***Expert systems with applications***, v. 36, Issue 3, p. 5297-5309, 2009.

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.*

KOHAVI, R..; JOHN, G. H. *Wrappers* for feature subset selection. ***Artif. Intell****., v.*97, 1997. P.273-324.

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.

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

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.

MOROZINI, João Francisco; OLINQUEVITCH, José Leônidas; HEIN, Nelson. Seleção de índices na análise de balanços: uma aplicação da técnica estatística 'ACP'. **Revista Contabilidade & Finanças - USP.** São Paulo Vol. 2 Número 41, Maio/Agosto 2006.

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

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

PINDYCK, R. S.; RUBINFELD, D. L. ***Econometria*****:** modelos e previsões. 4. ed. Rio de Janeiro: Campus Elsevier, 2004. 730 p.

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

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.

SANVICENTE, Antônio Zoratto; MINARDI, Andréa Maria A. F. *Identificação de indicadores contábeis significativos para previsão de concordata de empresas*. Disponível: <http://www.risktech.br/artigos/artigos técnicos/index.html>. Acesso em: 23/10/2005.

SCHAPIRE, R.E. **The strength of weak learnability**. Machine Learning, vol. 5, pp. 197–227, 1990.

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, Vol. 20 Número 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.

TSAI, C. F. Feature selection in bankruptcy prediction. ***Knowledge-Based Systems***, *Volume 22, Issue 2*, p. 120-127, mar. 2009.

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

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.

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,*** Volume 32, Issue 10, p. 2543-2559, oct. 2005.

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.

WOLK, H. I.; DODD, J. L.; ROZYCKI, J. J. *Accounting Theory*: Conceptual Issues in a Political and Economic Environment. **SAGE Publications**, 8th ed. 2013, 808 p.

YEH, Ching-Chiang, CHI, Der-Jang, LIN, Yi-Rong. Going-concern prediction using hybrid random forests and rough set approach. ***Information Sciences***, v. 254, 1 January 2014, Pages 98-110.

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.**

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*.

ANEXO – 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.