INTELLIGENT SYSTEMS IN ACCOUNTING, FINANCE AND MANAGEMENT
*Intell. Sys. Acc. Fin. Mgmt.* **21**, 155–167 (2014)
Published online 18 March 2014 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/isaf.1352

# INSOLVENCY PREDICTION IN THE PRESENCE OF DATA INCONSISTENCIES

A. MENDES,[a]* R. L. CARDOSO,[b] P. C. MÁRIO,[c,d] A. L. MARTINEZ[e] AND F. R. FERREIRA[f]

[a] *School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan, NSW, Australia*
[b] *The Brazilian School of Public and Business Administration, Getúlio Vargas Foundation, Rio de Janeiro, RJ, Brazil*
[c] *School of Economics, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil*
[d] *UNA University Center, Belo Horizonte, MG, Brazil*
[e] *FUCAPE Business School, Vitória, ES, Brazil*
[f] *Accounting School, Centro Universitário do Estado do Pará, Belém, PA, Brazil*

## SUMMARY

In this paper we use data inconsistencies as an indicator of financial distress. Traditional models for insolvency prediction normally ignore inconsistent data, either by removing or replacing it. Instead of removing that information, we propose a new variable to capture it; using it together with traditional accounting variables (based on financial ratios) for the purpose of insolvency prediction.

Computational tests use three datasets based on the financial results of 2033 Brazilian Health Maintenance Organizations over 7 years (2001 to 2007). Sixteen classification methods were used to evaluate whether or not the new variable impacted solvency prediction. Tests show a statistically significant improvement in classification accuracy – average results improve 1.3 ($p = 0.003$) and 1.8 ($p = 0.006$) percentage points, for 10-fold and leave-one-out cross-validations respectively. In addition, the analysis of false positives and false negatives shows that the new variable reduces the potentially harmful misclassification of false negatives (i.e. financially distressed companies being classified as financially healthy) and also reduces the estimated overall error rate.

Regarding the extensibility of the results, even though this work uses data from Brazilian companies only, the calculation of the financial ratios variables, as well as the inconsistencies, could be extended to most companies worldwide subject to governmental accounting regulations aligned with the International Financial Reporting Standards. Copyright © 2014 John Wiley & Sons, Ltd.

**Keywords:** data mining; insolvency prediction; classification; inconsistency

## 1. INTRODUCTION

This paper addresses an important topic in the analysis of insolvency prediction, namely the inconsistency of accounting data by companies. Insolvency and bankruptcy have been studied in the areas of accounting and finance for several decades. Most of these studies address these elements under different perspectives, either by trying to predict them (Altman, 1968; Ohlson, 1980; Newton, 2003; Altman & Hotchkiss, 2006), or by analysing the processes that occur during an insolvency crisis or bankruptcy (Aghion, Hart, & Moore, 1993; Hart, 2000). As a side note, in the literature, insolvency, failure and bankruptcy usually appear as synonyms; however, they refer to different moments. Insolvency is linked to a state, failure to an act, and bankruptcy has a legal meaning, as in a judicial process.

* Correspondence to: A. Mendes, School of Electrical Engineering and Computer Science, The University of Newcastle, Callaghan, NSW, Australia. E-mail: Alexandre.Mendes@newcastle.edu.au

To better understand the regulatory environment that health maintenance organizations (HMOs) are subject to in Brazil, and provide a comparison basis to international readers, we present a brief summary of the local health-care industry. Brazilian HMOs are regulated by the federal regulatory agency (the National Agency for Supplementary Health, hereafter ANS, for its original name). The detection of insolvency via the regular auditing process conducted by the ANS is very complex and fuzzy; although it is based on financial ratio parameters (see Table I for details). Since 2001, the ANS has adopted three levels of insolvency risk: *low*, *medium*, and *high*. Accounting information is analysed every quarter, and if the HMO does not break any threshold established by the ANS it is considered to be *low risk* and continues to operate normally. If the HMO breaks some of those thresholds, its risk is classified as *medium*. The HMO would then go through continuous, more rigorous analysis and could be asked to provide a so-called 'recovery plan'. In this case the company is also obliged to present financial data in a monthly basis to the ANS. Finally, if the HMO breaks the majority of thresholds established by the ANS and/or its recovery plan is not successful, its insolvency risk is reclassified to *high*. The HMO then becomes subject to direct intervention by the ANS, which may lead to the organization's discontinuity and assets liquidation. Several parameters are used to identify the financial state of an HMO, including current ratio, profitability, return on assets, and return on equity, among others, as presented in Table I.

This scenario creates the opportunity for companies in financial distress to manipulate their accounting information (Benham, 2005; Laughlin, 2007). In fact, we suspect that HMOs in financial distress manipulate their accounting information in order to introduce noise and mislead the process conducted by the ANS to evaluate their solvency status. Also, earnings management literature, which uses discretionary accruals as a proxy for manipulation, has shown that banks (Kato, Kunimura, & Yoshida, 2001), HMOs (Mensah, Considine, & Oakes, 1994) and producers in general (Jones, 1991; Navissi, 1999) manipulate their accounting information in order to mislead regulators. The originality of this work is that, differently from previous studies, it is not based on any discretionary accruals models. Instead, it relies on the consistency of HMOs' quarterly accounting data.

To our best knowledge, this study is the first attempt to consider a variable that captures accounting data inconsistency in insolvency prediction models. There are some similarities between this study and Cormier, Magnan, and Morard (2000) and Kasanen, Kinnunen, and Niskanen (1996). However, this study brings some new contributions to the area. First, Cormier *et al*. (2000) and Kasanen *et al*. (1996) focused on the accounting relevance for the capital market, providing evidence from the correlation between accounting data and price of securities, whereas this study focuses on the relevance of using accounting data for the assessment of financial distress. Second, the two aforementioned studies use discretionary accruals as proxies for unreliable accounting information, while we use a case-based investigation of quarterly financial reports while searching for inconsistencies (the parameters we used to assess inconsistency are listed in the Section 3). In addition, while those studies investigate Swiss and Finnish listed companies from different economic sectors, this paper specifically investigates Brazilian HMOs. Although mentioned before, it is important to emphasize that even though this work uses data from Brazilian companies only, the calculation of the financial ratios variables, as well as the inconsistencies, could be extended to most companies worldwide subject to governmental accounting regulations aligned with the International Financial Reporting Standards (http://www.ifrs.org).

This paper is organized as follows. In Section 2 we present the theoretical background of the concepts used in this study. In Section 3 we describe the data used in our analysis; in Section 4 we present the computational results; and in Section 5 we have the conclusions.

DOI: 10.1002/isaf

Table I. List of attributes considered in this study. It is a composition of the attributes used by the Brazilian National Agency for Supplementary Health in its regulatory process; the attributes from the works of Altman (1968) and Ohlson (1980); and the data inconsistency attribute introduced in this study

| Symbol | Attribute | Description |
|---|---|---|
| *Source: ANS (Brazilian National Agency for Supplementary Health)* | | |
| CR | Current ratio | CA/CL |
| CRe | Expanded current ratio | (CA + LTR)/(CL + NCL) |
| COD | Cost of debt | IE/(STD + LTD) |
| DE | Debt to equity ratio | TL/OE |
| EI | Expense index | (MCE + CE + AE)/OR |
| EIa | Amplified expense index | (MCE + CE + AE)/(OR – IE + II) |
| ERAd | Administration expenses to revenue ratio | AE/OR |
| ERCo | Commercial expenses to revenue ratio | CE/OR |
| ERMed | Medical care expenses to revenue ratio | MCE/OR |
| FL | Financial leverage | RE/RA |
| LS | Liability structure | CL/TL |
| LTAcs | Common size ratio for long-term assets | LTA/TA |
| NIvar | Net income variation | $(\mathrm{NI}_{q1} - \mathrm{NI}_{q4})/\mathrm{NI}_{q4}$ |
| OIA | Operating income to asset ratio | OI/TA |
| OL | Operating leverage | $\frac{\mathrm{OI}_q - \mathrm{OI}_{q-1}}{\mathrm{OI}_{q-1}} \times \frac{\mathrm{OR}_q - \mathrm{OR}_{q-1}}{\mathrm{OR}_{q-1}}$ |
| PC | Payable conversion ratio | 360 * AP/MCE |
| RAinv | Inverse of return on assets | TA/OR |
| RC | Receivables conversion ratio | 360 * AR/OR |
| ROE | Return on equity | NI/OE |
| Rvar | Revenue variation | $(\mathrm{OR}_{q1} - \mathrm{OR}_{q4})/\mathrm{OR}_{q4}$ |
| TAlog | Log of total asset | log(TA) |
| TR | Treasure to revenue ratio | (CCE – STD)/OR |
| WCN | Working capital needs | $\frac{\mathrm{CCE} - \mathrm{STD}}{(\mathrm{CA} - \mathrm{CL}) - (\mathrm{CCE} - \mathrm{STD})}$ |
| WCNcs | Common size ratio for working capital needs | $\frac{(\mathrm{CA} - \mathrm{CL}) - (\mathrm{CCE} - \mathrm{STD})}{\mathrm{TA}}$ |
| WCNr | Working capital needs to revenue ratio | $\frac{(\mathrm{CA} - \mathrm{CL}) - (\mathrm{CCE} - \mathrm{STD})}{\mathrm{OR}}$ |
| *Source: Altman (1968)* | | |
| AT | Asset turnover | OR/TA |
| CFA | Operating cash flow to asset ratio | EBIT/TA |
| DEinv | Inverse of equity to debt ratio | OE/TL |
| REcs | Common size ratio for retained earnings | (OE – CC)/TA |
| WCcs | Common size ratio for working capital | (CA – CL)/TA |
| *Source: Ohlson (1980)* | | |
| CHIN | Net income performance | $\frac{\mathrm{NI}_{q1} - \mathrm{NI}_{q4}}{\lvert \mathrm{NI}_{q1} - \mathrm{NI}_{q4} \rvert}$ |
| CRinv | Inverse of current ratio | CL/CA |
| FUTL | Operating cash flow to debt ratio | EBIT/(TA – OE) |
| INTWO | 2 years of losses | 1 if there were net losses in the past 2 years; 0 otherwise |
| NITA | Net income to asset | NI/TA |
| OENEG | Negative owner's equity | 1 if there is negative owner's equity; 0 otherwise |
| SIZE | Company size | log(TA)/GPIC |
| TLTA | Liabilities to assets ratio | TL/TA |
| WCcs | Common size ratio for working capital | (CA – CL)/TA |
| *New attribute* | | |
| DI | Data inconsistency | 0 if less than 50% of the quarters have inconsistencies; 1 otherwise |

AE: administration expenses; IE: interest expenses; OI: operating income AP: accounts payable; II: interest income; OR: operating revenue; AR: accounts receivable; LTA: long-term assets; RA: return on assets; CA: current asset; LTD: long-term debt; RE: return on equity; CC: contributed capital; LTR: long-term receivables; STD: short-term debt; CCE: cash and cash equivalent; MCE: medical care expenses; TA: total assets; CE: commercial expenses; NCL: non-current liabilities; TL: total liabilities; CL: current liabilities; NI: net income; GPIC: general price index to consumers; OE: owners' equity.

DOI: 10.1002/isaf

## 2. THEORETICAL BACKGROUND

This section presents a brief summary about the manipulation of accounting information (Section 2.1) and about insolvency prediction models (Section 2.2).

### 2.1. Manipulation of Accounting Information

The manipulation of accounting information can be understood as the choice of accounting practices or operational decisions with the goal to elaborate reports and report financial numbers different from those that would be presented if such practices were not adopted (Schipper, 1989; Healy & Wahlen, 1999; Fields, Lys, & Vincent, 2001; McKee, 2005). Therefore, the goal of portraying a specific financial position and performance is achievable through accounting practices and operating decisions. Accounting decisions involve the choice of accounting practices related to:

- identification of the phenomenon – transactions and other events;
- measurement of their effect on the company's performance and net assets;
- classification;
- accounting recognition;
- presentation and disclosure of the company's financial position.

In the literature there are numerous examples of manipulation of accounting information through misleading accounting practices. Among them, some relevant studies are Jones (1991); Dechow, Sloan, and Sweeney (1995) and Kang and Sivaramakrishnan (1995). In addition, and specifically related to Brazilian companies, the study of Martinez (2001) used a sample of non-financial companies traded at the local stock market to show that the most common manipulation of accounting information aims to avoid the reduction of the net profit, as well as to reduce its volatility (also referred to as income smoothing). Also, Fuji (2004) has shown that, in a sample composed of the 50 largest Brazilian banks, the manipulation of accounting information is concentrated on the use of the provision account for allowance for bad debts. This was aimed at reducing the political cost related to the regulation made by the Brazilian Central Bank. There are several other examples in the same line. In particular, we cite the work that preceded this, Cardoso (2005), which used a reduced set of HMOs and three years of quarterly financial information. That study showed that HMOs manipulate accounting information in order to avoid breaking financial thresholds established by the ANS (specifically to avoid reporting losses and negative owners' equity).

The second type of accounting information manipulation is by operating decisions. McKee (2005) exemplifies it with the use (or not) of special discounts, or special programs, to increase sales close to the end of a quarter in which the income goals were not achieved. Other types of operating decisions include the investment in new equipment and hiring of new staff, among others, which will impact the company's cash flow – and consequently the income and expenditures associated with these activities. There are very few studies in the literature that deal with this kind of manipulation (Martinez & Cardoso, 2009; Gunny, 2010; Zang, 2012).

In this work, we use the concept of *data inconsistencies* as an indication of data manipulation. HMOs manipulate accounting information usually in very simplistic ways (e.g. providing inconsistent data to ANS). The ANS regulates the health industry in Brazil and regulation is based on the 'market-wide cost savings from regulation' logic. Where the costs of complying with a one-size-fits-all regime are relatively low, standardization of corporate reporting can make it easier for the agency to process the information and to compare across companies (Leuz & Wysocki, 2008).

It is quite clear that considering the way accounting policies are regulated and accounting information is used by the ANS, and that information is provided in an environment characterized by uncertainty and imperfect information, HMOs' managers have different views regarding accounting information compared with other stakeholders (including ANS's staff). Therefore, it is likely that HMOs in financial distress will manipulate their accounting information in order to introduce noise and mislead the process conducted by the ANS to evaluate their solvency status (Cardoso, 2005).

Even though data inconsistencies can be attributed to the malicious manipulation of accounting information in order to mislead regulators, they can also be caused by unintended errors. In both cases, however, we can consider such inconsistencies as financial distress proxies. The first case is clear, as intended (malicious) inconsistencies are used by financially distressed companies to mislead regulators, and avoid or delay being identified as insolvent. In the second case, unintended (error) inconsistencies can also be an attribute of financial distress because insolvent companies generally lack the appropriate resources (monetary, human and technological) required to correctly report accounting information that faithfully represents their financial transactions and current economic status.

### 2.2. Insolvency Prediction Models

The literature on insolvency prediction models is quite vast. For a comprehensive literature review, we refer the reader to the work of Balcaen and Ooghe (2006). In order to check whether data inconsistency is indeed an indication of financial distress and future insolvency, we focused on two of the most traditional prediction models: Altman (1968) and Ohlson (1980). The most relevant contribution of Altman's model was the use of multivariate data analysis technology to predict insolvency. On a sample comprised of 66 companies, with half of them going bankrupt in the period between 1946 and 1965, Altman's work used discriminant analysis to identify which financial ratios could better discriminate insolvent companies from solvent ones, which resulted in a model (*Z*-score) composed of five accounting-based ratios:

- common size ratio for working capital;
- common size ratio for retained earnings;
- operating cash flow to asset ratio;
- equity to debt ratio;
- asset's turnover.

Twelve years later, Ohlson (1980) presented a new insolvency prediction model. The main improvement compared with its predecessors was the use of another statistical technique, namely the *conditional logit model*. This technique is less dependent on assumptions than the discriminant analysis and provides a probability function as result. Based on a sample of 105 listed companies that went bankrupt in the period between 1970 and1976 and 2058 solvent listed companies randomly selected, Ohlson (1980) built a model comprised of nine financial ratios: SIZE, TLTA, CCLA, CLCA, OENEG, NITA, FUTL, INTWO, CHIN. Several other attributes were included in this study based on their frequent use in other insolvency prediction studies (Charitou, Neophytou, & Charalambous, 2004; Cardoso, 2005; Min & Lee, 2005; Aziz & Dar, 2006) (see Table I for a list of attributes used in this work).

Although both models are mature and have led to numerous important studies, there is a clear gap present, evidenced by the data manipulation normally present in financial reports. The question is

DOI: 10.1002/isaf

whether or not insolvency prediction models in general could be improved by both (a) keeping inconsistent data in the analysis, without any filtering or smoothing/data replacement process, and (b) the use of an attribute that aggregates information about the reliability of accounting data.

### 2.3. Classification Methods

Sixteen classification methods were used to assess the use of data inconsistencies in insolvency prediction. The decision for such a large number of methods was to remove the focus of the study from any particular classification technique and put it in the relation

$$\text{data inconsistency} \Rightarrow \text{data manipulation} \Rightarrow \text{financial distress} \Rightarrow \text{insolvency}$$

Other studies might adopt a different approach, using instead more complex classification models and/or fewer, previously-validated attributes (Anandarajan, Lee, & Anandarajan, 2001; McKee & Lensberg, 2002; Pompe, 2005; Zhou, Lai, & Yen, 2014). Both approaches are valid depending on the goal of the study. The 16 classification methods used in this study are part of the data mining software tool Weka (Hall et al., 2009) and are listed below:

- *BayesNet*: Bayes network learning
- *ClassViaRegression*: regression models
- *FT*: functional trees
- *IB1*: nearest-neighbour classifier
- *J48*: a decision tree classifier based on the C4.5 algorithm
- *LADTree*: LogitBoost-based decision tree
- *Logistic*: multinomial logistic regression
- *LogitBoost*: additive logistic regression
- *LWL*: locally weighted learning
- *NaiveBayes*: common naive Bayes classifier
- *OneR*: traditional 1R classifier
- *PART*: separate-and-conquer partial decision list
- *RandomCommittee*: ensemble of random trees classifiers
- *RandomForest*: forest of random trees
- *SimpleLogistic*: linear logistic regression model
- *SMO*: support vector classifier.

All the methods listed above are well-established, traditional classifiers and there is plenty of literature available about them. Therefore, we refer the reader to the work of Witten and Frank (2005), which contains detailed descriptions of the methods listed.

## 3. DATA COLLECTION AND PROCESSING

The data used in this study were obtained directly from the ANS in 2008.[1] The accounting information comprises 2033 HMOs and contains all their financial reports between 2001 and 2007, plus the first two

[1] Data were obtained through an agreement with the ANS and The National Council for Scientific and Technological Development (CNPq). Process number MCT/CNPq/ANS 410612/2006-5.

quarters of 2008. Some preprocessing on the raw data collected was necessary before the analysis could be carried out, which is described next.

### 3.1. Data Processing

During the data validation stage, we detected many differences (or errors) in the HMOs' balances, which we simply called 'inconsistencies'. Instead of treating them as missing values, and filtering or smoothing them out, we decided to keep them in our data and incorporated a binary variable named DI (for *data inconsistency*) to control for their occurrence. A 0 means that less than 50% of the HMOs' quarterly information presents accounting inconsistencies, and 1 means that 50% or more of the quarterly information presents inconsistencies. The criteria to identify inconsistencies are as follows.

- Any of the following accounts has a negative balance: current assets, noncurrent assets, monetary current assets, prepaid commercial expenses, current liabilities, noncurrent liabilities, monetary current liabilities or contributed capital.
- Any of the following accounts has a balance lower than the equivalent of US$500.00: total assets, total liabilities or total revenues.
- Total assets differ from the sum of total liabilities and owners' equity.
- Total assets differ from the sum of current assets and noncurrent assets.
- Total liabilities differ from the sum of current liabilities and non-current liabilities.
- Net income differs from the sum of total revenues and total expenses.
- Total assets balance is lower than owners' equity.
- The variable SIZE (measured in accordance with Table I) has a negative value.
- Any of the following variables is a missing value: SIZE or cost of debt (COD).

As mentioned before, inconsistencies can be maliciously created to mislead the regulators, or they can derive from unintended errors while preparing the financial reports. Nevertheless, whether malicious or not, inconsistencies affect accounting information reliability, so the financial reporting no longer faithfully represents the economic phenomena that it purportedly should. The main contribution of this paper is to consider inconsistencies as an attribute to discriminate solvent from insolvent companies, either because financially distressed companies intentionally manipulate accounting data in order to mislead regulators, or because such companies lack the appropriate resources to hire skilled staff, to keep appropriate information technology systems, or to offer incentives for avoiding such unintended errors.

For each company and each quarterly report, we generate the values of the 39 attributes in Table I. Then, each attribute is averaged over the following periods:

- from 2001 until the last quarter prior to the insolvency, if the company is insolvent.
- from 2001 until the last quarter of the data collection interval, if the company is solvent.

That creates an average financial portrait of the company for the quarters when the company was still solvent. That averaged data are used by the classification methods to determine whether the company will still be solvent or not in the following quarter. The data collected from the ANS allows for that, as it specifies in which quarter the company became insolvent, or if it is still solvent.

### 3.2. Generation of Datasets

Three datasets were generated for this study using the original data from the ANS. It is important to mention, though, that the data are very unbalanced in the number of solvent and insolvent companies. For example, of the 2033 companies, we have 466 insolvent ones and 1567 solvent ones – a proportion close to 30/70. Using the complete data in our tests would cause all models to be biased towards classifying samples as solvent. Therefore, we randomly filtered out the excessive solvent companies from each dataset, forcing them to a 50/50 proportion between solvent and insolvent companies. The final datasets are as follows:

- *DI$_0$ dataset*. Originally comprised of 1239 companies with DI = 0 (i.e. HMOs with less than 50% of the quarterly information presenting accounting inconsistencies). This dataset is used to measure insolvency prediction in an ideal scenario where companies reported their quarterly information with few inconsistencies; an indication of little manipulation of accounting information. After filtering out the excess number of solvent companies, the dataset ended with 202 companies: 101 solvent and 101 insolvent.
- *DI$_{exc}$ dataset*. Originally comprised all 2033 companies, but without the DI attribute. This dataset is used as a 'control', to estimate insolvency prediction without the DI attribute in the analysis. After filtering, the dataset was reduced to 932 companies: 466 solvent and 466 insolvent.
- *DI$_{inc}$ dataset*. Same data as DI$_{\mathrm{exc}}$, but the attribute DI is included in the dataset to control for inconsistencies.

Apart from the attribute DI, *all* other financial ratio attributes considered for analysis in the three datasets are the same. In this study, a company is insolvent when it was discontinued and had its assets liquidated, after all interventions by the ANS have failed.

## 4. RESULTS

### 4.1. Prediction Accuracy

In this section, we present all computational results for insolvency prediction for the 16 classifiers and the three datasets used in this study. The results are summarized in Table II and are divided into two parts. The first three columns with results refer to a 10-fold cross-validation classification – one for each dataset, DI$_0$, DI$_{\mathrm{exc}}$ and DI$_{\mathrm{inc}}$. The next three columns correspond to the leave-one-out cross-validation tests – again, for the three datasets.

A few conclusions can be drawn from these results. First, when the analysis excludes companies with higher quantities of accounting inconsistencies (DI$_0$ dataset), prediction accuracy stays at low levels. The figures were 55.1% and 53.8% for 10-fold and leave-one-out cross-validations respectively. When those companies are included in the analysis, accuracy increases considerable, by 12.6 and 13.6 percentage points respectively for the two types of cross-validation. That is a clear indication that data inconsistencies and insolvency might indeed be connected. Finally, when we aggregate the accounting inconsistencies by introducing the variable DI, accuracy increases again by small, though statistically significant margins: 1.3 and 1.8 percentage points. Given the proximity of the numbers, a paired *t*-test was conducted using the figures for DI$_{\mathrm{exc}}$ and DI$_{\mathrm{inc}}$ and the results were $p = 0.0030$ and $p = 0.0059$ respectively.

Table II. Insolvency prediction accuracy for the 16 classification methods. The three datasets were tested using both 10-fold cross-validation and leave-one-out cross-validation. When companies with higher quantities of accounting inconsistencies are filtered out from the analysis, prediction accuracy remains at low levels – just a bit higher than 50% (see the two columns labelled $DI_0$). Once those companies are included in the analysis, accuracy increases by over 12 percentage points (columns labelled $DI_{exc}$). Finally, when we introduce the new attribute DI (which represents those inconsistencies) and use it in the analysis, accuracy increases again by a small, though statistically significant margin (columns labelled $DI_{inc}$)

| Method | 10-fold cross-validation | | | Leave-one-out cross-validation | | |
|---|---|---|---|---|---|---|
| | $DI_0$ (%) | $DI_{exc}$ (%) | $DI_{inc}$ (%) | $DI_0$ (%) | $DI_{exc}$ (%) | $DI_{inc}$ (%) |
| BayesNet | 50.5 | 62.1 | 62.6 | 51.5 | 63.7 | 63.8 |
| ClassViaRegression | 51.0 | 72.9 | 73.1 | 55.4 | 74.1 | 74.2 |
| FT | 55.9 | 70.5 | 72.4 | 50.5 | 71.2 | 72.2 |
| IB1 | 55.4 | 60.2 | 60.8 | 52.5 | 61.6 | 62.2 |
| J48 | 53.4 | 70.5 | 71.2 | 52.0 | 69.2 | 72.0 |
| LADTree | 52.5 | 54.6 | 60.7 | 52.0 | 57.2 | 60.9 |
| Logistic | 54.5 | 70.6 | 71.0 | 55.4 | 71.3 | 70.4 |
| LogitBoost | 56.9 | 74.0 | 74.0 | 53.5 | 74.1 | 74.0 |
| LWL | 54.0 | 73.5 | 73.3 | 52.0 | 73.6 | 73.6 |
| NaiveBayes | 55.9 | 44.3 | 45.0 | 54.5 | 42.2 | 43.1 |
| OneR | 52.5 | 71.2 | 71.2 | 58.4 | 72.2 | 72.2 |
| PART | 55.4 | 71.4 | 73.3 | 55.0 | 65.9 | 73.4 |
| RandomCommittee | 57.9 | 69.3 | 72.0 | 55.9 | 67.4 | 71.9 |
| RandomForest | 61.4 | 69.7 | 71.1 | 58.9 | 70.0 | 71.4 |
| SimpleLogistic | 56.9 | 71.0 | 71.6 | 52.5 | 70.9 | 71.6 |
| SMO | 53.0 | 63.8 | 67.4 | 47.0 | 60.9 | 67.7 |
| Average | 54.8 | 66.9 | 68.2 | 53.6 | 66.6 | 68.4 |

We observe that for 31 out of 32 cases, the inclusion of companies with higher quantities of accounting inconsistencies in the analysis led to an improvement in the insolvency prediction accuracy (considering 16 methods and two types of cross-validation). Similarly, the introduction of the variable DI improved the accuracy in 26 cases, gave the same result in four and worsened the accuracy only twice. These results clearly point out that disregarding accounting inconsistencies, by either filtering them out or smoothing/replacing them, leads to information loss and will have a negative impact on insolvency prediction.

### 4.2. False Solvency and False Insolvency Results

In addition to prediction accuracy, another important analysis is how the classification mistakes are distributed, commonly known in the literature as false positives and false negatives. Since our goal is to predict insolvency, a false positive would be a solvent company being mistakenly classified as insolvent (i.e. a false insolvent), and a false negative would be an insolvent company being mistakenly classified as solvent (i.e. a false solvent). The two situations should be avoided, of course, but in this study a false solvent is far worse than a false insolvent. The reason is that a false insolvent means that the ANS would subject an otherwise financially healthy company to more scrutiny. That is not a positive for the company, of course, but clients and investors are minimally harmed. The other situation is a false solvent. In this case a financially distressed company would potentially go undetected, suffer no scrutiny by the regulatory agency and eventually become insolvent, considerably damaging clients and investors.

This asymmetry in the consequences of false solvency and false insolvency classification errors requires that prediction methods should, at the same time, maximize accuracy and minimize false solvency. Towards that end, Table III we presents the results of false solvent and false insolvent classifications for the three datasets, all classification methods, and both 10-fold and leave-one-out cross-validations. In addition to false solvents and false insolvents, we also present the estimated overall error rate (EOER). The EOER tries to capture the impact of false solvents and false insolvents by multiplying their probabilities by the proportion of insolvent and solvent samples in the dataset respectively (Etheridge, Sriram, & Hsu, 2000; McKee, 2007). The proportions of insolvent and solvent samples were calculated using the

Table III. Results for false solvent and false insolvent classifications, and the EOER. Notice that, for both types of cross-validation, when data inconsistencies are largely ignored, the number of false solvents is very high (29.6% and 29.8% respectively), which represents a large risk to both clients and investors. When we add data inconsistencies to the analysis, false solvents are reduced considerably; and finally, when the attribute DI is used, false solvency drops even further on average, with a comparatively low increase in false insolvency numbers. Similarly, EOER is relatively high without the use of data inconsistencies – above 19%. EOER is considerably reduced when data inconsistencies are included; and reduced again by a small amount when the attribute DI is used

| Method/Instance | False solvent/false insolvent (EOER) (all %) | | |
|---|---|---|---|
| | $DI_0$ | $DI_{exc}$ | $DI_{inc}$ |
| *10-fold cross-validation* | | | |
| BayesNet | 40.0/9.4 (16.4) | 22.5/15.3 (17.0) | 22.1/15.3 (16.9) |
| ClassViaRegression | 25.7/23.3 (23.9) | 12.3/14.8 (14.2) | 12.2/14.7 (14.1) |
| FT | 23.8/20.3 (21.1) | 14.3/15.2 (15.0) | 13.9/13.6 (13.7) |
| IB1 | 28.2/16.3 (19.0) | 27.7/12.1 (15.7) | 27.6/11.6 (15.3) |
| J48 | 36.6/9.9 (16.0) | 14.7/14.8 (14.8) | 12.6/16.2 (15.4) |
| LADTree | 23.8/23.8 (23.8) | 40.6/4.8 (13.0) | 35.3/4.0 (11.2) |
| Logistic | 25.7/19.8 (21.2) | 16.4/13.0 (13.8) | 14.8/14.2 (14.3) |
| LogitBoost | 21.3/21.8 (21.7) | 13.1/12.9 (12.9) | 13.1/12.9 (12.9) |
| LWL | 41.6/4.5 (13.0) | 14.6/11.9 (12.5) | 14.8/11.9 (12.6) |
| NaiveBayes | 40.6/3.5 (12.0) | 13.3/42.4 (35.7) | 12.7/42.4 (35.6) |
| OneR | 20.8/26.7 (25.3) | 11.9/16.8 (15.7) | 11.9/16.8 (15.7) |
| PART | 36.1/8.4 (14.7) | 15.2/13.4 (13.8) | 12.2/14.5 (14.0) |
| RandomCommittee | 25.2/16.8 (18.7) | 10.9/19.7 (17.7) | 7.8/20.2 (17.4) |
| RandomForest | 22.3/20.3 (20.8) | 9.0/20.5 (17.9) | 7.0/21.7 (18.3) |
| SimpleLogistic | 23.8/19.3 (20.3) | 14.2/14.8 (14.7) | 13.8/14.6 (14.4) |
| SMO | 38.6/8.4 (15.3) | 21.0/15.1 (16.5) | 13.8/18.8 (17.7) |
| Average | 29.6/15.8 (19.0) | 17.0/16.1 (16.3) | 15.4/16.5 (16.2) |
| *Leave-one-out cross-validation* | | | |
| BayesNet | 44.6/4.0 (13.3) | 21.3/14.9 (16.4) | 21.5/14.7 (16.3) |
| ClassViaRegression | 21.8/22.8 (22.6) | 12.1/13.7 (13.3) | 12.1/13.6 (13.3) |
| FT | 27.2/22.3 (23.4) | 12.6/16.2 (15.4) | 13.9/13.8 (13.8) |
| IB1 | 29.7/17.8 (20.5) | 27.1/11.3 (14.9) | 26.8/10.9 (14.5) |
| J48 | 27.2/20.8 (22.3) | 16.1/14.7 (15.0) | 12.1/15.9 (15.0) |
| LADTree | 22.8/25.2 (24.6) | 38.3/4.5 (12.2) | 35.9/3.1 (10.6) |
| Logistic | 24.8/19.8 (20.9) | 16.1/12.6 (13.4) | 15.6/14.1 (14.4) |
| LogitBoost | 24.3/22.3 (22.8) | 13.0/12.9 (12.9) | 13.0/13.0 (13.0) |
| LWL | 43.1/5.0 (13.7) | 14.5/11.9 (12.5) | 14.5/11.9 (12.5) |
| NaiveBayes | 41.1/4.5 (12.9) | 11.6/46.2 (38.3) | 10.6/46.2 (38.0) |
| OneR | 20.8/20.8 (20.8) | 11.3/16.5 (15.3) | 11.3/16.5 (15.3) |
| PART | 27.2/17.8 (20.0) | 20.7/13.4 (15.1) | 11.7/14.9 (14.2) |
| RandomCommittee | 26.3/17.8 (19.7) | 11.2/21.5 (19.1) | 7.1/21.0 (17.8) |
| RandomForest | 24.3/16.8 (18.5) | 8.5/21.6 (18.6) | 7.3/21.4 (18.2) |
| SimpleLogistic | 29.2/18.3 (20.8) | 14.8/14.3 (14.4) | 13.6/14.8 (14.5) |
| SMO | 42.1/10.9 (18.1) | 23.8/15.2 (17.2) | 13.6/18.7 (17.5) |
| Average | 29.8/16.7 (19.7) | 17.1/16.3 (16.5) | 15.0/16.5 (16.2) |

DOI: 10.1002/isaf

original dataset. Therefore, the proportion of solvent samples is 1567/2033 = 0.77, and for insolvent samples it is 466/2033 = 0.23. The general formula of the EOER is

$$\text{EOER} = \text{false}_{\text{Solv}_{\text{rate}}} \times \text{proportion}_{\text{Insolv}_{\text{samples}}} + \text{false}_{\text{Insolv}_{\text{rate}}} \times \text{proportion}_{\text{Solv}_{\text{samples}}}$$

Starting with the 10-fold cross-validation, we observe for the dataset $DI_0$ that the number of false solvents is nearly twice the number of false insolvents (29.6% versus 15.8%, on average), which is exactly the worst scenario. Therefore, removing data inconsistencies from the analysis not only deteriorates classification accuracy, but also creates the worst type of misclassification. When the data inconsistencies are included (dataset $DI_{exc}$), the percentage of false solvents decreases considerably, from 29.6% to 17.0%, and false insolvents increase just marginally, by 0.3 percentage points – and in addition accuracy improves, as shown in the previous section. The inclusion of the attribute DI (dataset $DI_{inc}$) reduces false solvents by 1.6 percentage points and increases false insolvents by 0.4 percentage points, which again represents an improvement in the classification outcome. Regarding the EOER, the value starts at 19.0, when no inconsistency data are included. Once inconsistencies are included, EOER is reduced to 16.3; and finally, when those inconsistencies are captured into the attribute DI, there is a further, slight drop to 16.2. Therefore, the EOER also captures the improvement brought by the inclusion of data inconsistencies and their use as an attribute.

For the leave-one-out cross-validation we observe similar results. The number of false solvents and false insolvents for dataset $DI_0$ is 29.8% and 16.7% respectively. With dataset $DI_{exc}$, false solvents drop by 12.7 percentage points and false insolvents by 0.4 percentage points. The impact of including attribute DI is a further reduction of 2.1 percentage points in false solvents and a marginal increase of 0.2 percentage points in false insolvents. The EOER also presents a similar result, with 19.7 for $DI_0$, followed by a 3.2 percentage points reduction in $DI_{exc}$ and a further 0.3 percentage point reduction in $DI_{inc}$.

Given these results, it becomes clear that the more (and appropriately controlled) data inconsistency information that is used in the classification, the better the accuracy becomes; also, less false solvent misclassifications and lower EOER are observed – considering the average of the 16 classification methods.

## 5. CONCLUSION

This study has addressed the issue of inconsistencies in accounting information and how they could be used to improve the accuracy of insolvency prediction. We used an original dataset of 2033 Brazilian HMOs and created three datasets with different characteristics. Those datasets were built to reflect different ways of treating data inconsistencies: first by removing them (dataset $DI_0$); second by keeping them but not controlling for inconsistency (dataset $DI_{exc}$); and third by adding an attribute that aggregates inconsistency information, to be used in the insolvency prediction process (dataset $DI_{inc}$).

Sixteen general classification methods were used on the three datasets, and we employed both 10-fold and leave-one-out cross-validation procedures. Results indicate that removing inconsistencies from the data has a negative impact on insolvency prediction. Prediction accuracy is barely above 50% for both types of cross-validation. The use of data inconsistencies improves prediction by over 12 percentage points, and the use of the attribute that aggregates inconsistency information further improves accuracy by around 1.5 percentage points. That last improvement, although it might look marginal, is statistically significant, as shown by a paired *t*-test.

DOI: 10.1002/isaf

In addition to overall accuracy, we also analysed the results for false positives and false negatives –- in our case, false insolvency and false solvency respectively. It is shown that in using data inconsistencies we improve not only accuracy, but also reduce considerably the percentage of potentially harmful false solvents, at the cost of a marginal increase in false insolvent misclassifications. In addition, the EOER is also considerably reduced with the use of data inconsistencies.

This study has a few limitations, which we list next. First, it does not differentiate between inconsistencies caused by malicious data manipulation and those caused unintentionally (e.g. by a genuine clerical mistake). However, we argue that, independently of their nature, inconsistencies are representative of financial distress. Second, the classification methods used in this study are very traditional – and general in nature. There are specialized models for insolvency prediction that deliver better accuracy, as pointed out in Section 1. We chose to use traditional classification methods specifically to put the focus on the clever use of data inconsistencies, instead of removing them from the analysis. As a final remark, it is intuitive that the improvements observed in our tests might repeat themselves when more complex models are used; that hypothesis should be tested in the future.

This study has shown that the common practice of filtering out or smoothing/replacing data inconsistencies is not recommended when the goal is to predict insolvency or to analyse the financial health of companies. Therefore, data inconsistencies could, and should, be used to predict insolvency.

## REFERENCES

Aghion P, Hart O, Moore J. 1993. A proposal for bankruptcy reform in the U.K. Insolvency Law Practice. *Insolvency Law Practice* **9**: 103–108.

Altman EI. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance* **23**: 589–609.

Altman EI, Hotchkiss E. 2006. Corporate Financial Distress and Bankruptcy: Predict and Avoid Bankruptcy, Analyze and Invest in Distressed Debt. John Wiley & Sons, Inc.: New Jersey, USA.

Anandarajan M, Lee P, Anandarajan A. 2001. Bankruptcy prediction of financially stressed firms: an examination of the predictive accuracy of artificial neural networks. *International Journal of Intelligent Systems in Accounting, Finance and Management* **10**: 69–81.

Aziz MA, Dar HA. 2006. Predicting corporate bankruptcy: where we stand? *Corporate Governance* **6**(1): 18–33.

Balcaen S, Ooghe H. 2006. 35 years of studies on business failure: an overview of the classical statistical methodologies and their related problems. *British Accounting Review* **38**: 63–93.

Benham L. 2005. Licit and illicit firm responses to public regulation. In Handbook of New Institutional Economics, Menard C, Shirley MM (eds). Kluwer Press: New York, NY; Chapter 23.

Cardoso RL. 2005. Regulacao economica e escolhas de praticas contabeis: evidencias no mercado de saude suplementar brasileiro, PhD thesis in Accounting Sciences, Departamento de Contabilidade e Atuaria, Universidade de Sao Paulo, Sao Paulo, Brazil.

Charitou AC, Neophytou E, Charalambous C. 2004. Predicting corporate failure: empirical evidence for the UK. *European Accounting Review* **13**(3): 465–497.

Cormier D, Magnan M, Morard B. 2000. The contractual and value relevance of reported earnings in a dividend-focused environment. *European Accounting Review* **9**(3): 387–417.

Dechow PM, Sloan RG, Sweeney AP. 1995. Detecting earnings management. *The Accounting Review* **70**: 193–225.

Etheridge HL, Sriram RS, Hsu HYK. 2000. A comparison of selected artificial neural networks that help auditors evaluate client financial viability. *Decision Sciences* **31**(2): 531–550.

Fields TD, Lys TZ, Vincent L. 2001. Empirical research on accounting choice. *Journal of Accounting and Economics* **31**: 255–307.

Fuji AH. 2004. Gerenciamento de resultados contabeis no ambito das instituicoes financeiras atuantes no Brasil, Master thesis in Accounting Sciences, Departamento de Contabilidade e Atuaria, Universidade de Sao Paulo, Sao Paulo, Brazil.

DOI: 10.1002/isaf

Gunny K. 2010. The relation between earnings management using real activities manipulation and future performance: evidence from meeting earnings benchmarks. *Contemporary Accounting Research* **27**: 855–888.

Hart O. 2000. Different approaches to bankruptcy. Technical Report 7921, National Bureau of Economic Research, Massachusetts, USA, viewed 15 February 2014, http://www.nber.org/papers/w7921

Healy PM, Wahlen JM. 1999. A review of the earnings management literature and its implications for standard setting. *Accounting Horizons* **13**: 365–383.

Jones JJ. 1991. Earnings management during import relief investigations. *Journal of Accounting Research* **29**: 193–228.

Kang SH, Sivaramakrishnan K. 1995. Issues in testing earnings management and an instrumental variable approach. *Journal of Accounting Research* **33**: 353–367.

Kasanen E, Kinnunen J, Niskanen J. 1996. Dividend-based earnings management: empirical evidence from Finland. *Journal of Accounting and Economics* **22**: 283–312.

Kato K, Kunimura M, Yoshida Y. 2001. Banks' earnings management before violation of dividend regulation in Japan. Technical Report 238, Osaka University of Economics, Japan, viewed 14 February 2014, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=273945

Laughlin R. 2007. Critical reflections on research approaches, accounting regulation and the regulation of accounting. *The British Accounting Review* **39**: 271–289.

Leuz C, Wysocki P. 2008. Economic consequences of financial reporting and disclosure regulation: a review and suggestions for future research. Technical Report, University of Chicago and MIT Sloan School of Management, USA, viewed 14 February 2014, http://web.mit.edu/wysockip/www/papers/LW2008.pdf.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter* **11**: 10–18.

Martinez AL. 2001. Gerenciamento de resultados contabeis: estudo empirico das companhias abertas brasileiras, PhD thesis in Accounting Sciences, Departamento de Contabilidade e Atuaria, Universidade de Sao Paulo, Sao Paulo, Brazil.

Martinez AL, Cardoso RL. 2009. Gerenciamento de resultados no Brasil mediante decisoes operacionais. *REAd – Revista Eletronica de Administracao* **15**: 1–27.

McKee TE, Lensberg T. 2002. Genetic programming and rough sets: a hybrid approach to bankruptcy classification. *European Journal of Operational Research* **138**(2): 436–451.

McKee TE. 2005. Earnings Management: An Executive Perspective. Cengage Learning: Stamford, CT.

McKee TE. 2007. Altman's 1968 bankruptcy prediction model revisited via genetic programming: new wine from an old bottle or a better fermentation process? *Journal of Emerging Technologies in Accounting* **4**: 87–101.

Mensah YM, Considine JM, Oakes L. 1994. Statutory insolvency regulations and earnings management in the prepaid health-care industry. *Accounting Review* **69**: 70–95.

Min JH, Lee YC. 2005. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications* **28**(4): 603–614.

Navissi F. 1999. Earnings management under price regulation. *Contemporary Accounting Research* **16**: 281–304.

Newton GW. 2003. Corporate Bankruptcy: Tools, Strategies, and Alternatives. John Wiley & Sons, Inc.: New Jersey, USA

Ohlson JA. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research* **18**: 109–131.

Pompe PPM. 2005. Bankruptcy prediction: the influence of the year prior to failure selected for model building and the effects in a period of economic decline. *International Journal of Intelligent Systems in Accounting and Finance* **13**: 95–112.

Schipper K. 1989. Commentary on earnings management. *Accounting Horizons* **3**: 91–102.

Witten IH, Frank E. 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann: San Francisco, CA.

Zang AY. 2012. Evidences on the trade-off between real manipulation and accrual-based earnings management. *Accounting Review* **87**: 675–703.

Zhou L, Lai KK, Yen J. 2014. Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation. *International Journal of Systems Science* **45**(3): 241–253.

DOI: 10.1002/isaf