

# Log-Modulus for Knowledge Discovery in Databases of Financial Reports

Duarte Trigueiros

University of Macau, University Institute of Lisbon  
Lisbon, Portugal  
Email: dmt@iscte.pt

Carolina Sam

Master of European Studies Alumni Association  
Macau, China  
Email: kasm@customs.gov.mo

**Abstract**—An alternative is proposed to the use of ratios in financial predictive modelling. Such alternative, the “log-modulus”, overcomes limitations, which have hitherto thwarted most of the previous attempts to predict financial attributes from data. Moreover, the use of log-modulus opens-up the prospect of performing Knowledge Discovery in Databases (KDD) of financial reports. Using controlled experiments, the paper shows that models using log-modulus are accurate, robust and balanced in cases where ratios fail to deliver feasible results. The paper also provides a theoretical basis supporting the observed ability of log-modulus to allow knowledge discovery of financial statements.

**Keywords**—*Type of Information Mining; Knowledge Discovery in Databases; Predictive Modelling; Financial Reports.*

## I. INTRODUCTION

Business companies, namely those listed in stock markets, are required to prepare annual reports reflecting their financial activity and position at the end of each year. Large databases containing these reports are routinely scrutinised by investors, banks, regulators and other parties with the object of taking decisions regarding individual companies and industrial sectors. Such scrutiny, and the corresponding diagnostic, is known as “Financial Analysis”.

Financial Analysis aims to diagnose the financial outlook of a company. The major source of data for such diagnostic is the set of financial reports regularly made public by the company and by other companies in the same industrial sector. The diagnostic itself consists of identifying and in some cases measuring the state of financial attributes, such as Manipulation, Going Concern, Solvency, Profitability and others. The tool used by analysts to assess such attributes is the “ratio”, a quotient of two monetary amounts appropriately chosen. After being identified and measured, financial attributes convey a clear picture of a company’s future economic prospects and may support the taking of momentous decisions, such as to buy or not to buy shares, to lend money and others. Financial attributes, therefore, are the knowledge set where investing, lending and other decisions are based.

The paper is about the discovery and assessment of underlying attributes in databases of financial reports. It describes a methodology capable of reliably producing, from such databases, knowledge represented so as to allow inferencing.

Attempts to perform analytical modelling of financial attributes have largely failed except in one instance,

bankruptcy prediction [1]. Other, equally vital attributes, such as the trustworthiness as opposed to fraudulent reports, have resisted attempts to be reliably predicted [2]. Such failure is largely due to difficulties posed by ratios when used as predictors but, hitherto, no attempt has been made to find alternatives. The objective of the paper is to overcome the current stalemate by proposing a type of predictor, the log-modulus [3], which overcomes ratios’ limitations and is amenable to knowledge discovery. An effective KDD of financial reports would quicken and lighten the analysis process, freeing analysts to concentrate on specific cases thus improving their efficiency.

Section II describes the KDD challenge being tackled; Section III offers theoretical considerations supporting the use of log-modulus; Section IV presents results of controlled experiments where log-modulus are compared with ratios; Section V highlights expected benefits.

## II. FINANCIAL ANALYSIS, A KDD CHALLENGE

Financial reports are standardized data-sets prepared and published by business companies on a regular basis. They contain, besides non-numerical data, a collection of monetary amounts with an attached meaning: revenues of the period, different types of expenses, asset and liability values at the end of the period and others. Such amounts are obtained via a process involving the recognition and aggregation into “accounts”, of similar transactions relating to the period. The resulting “set of accounts” is made available to the public together with non-numerical information in the form of a financial report.

Amongst investors, regulators and banks, an extremely popular value-added product is the database containing current and past financial reports of companies listed in one or several regions. This database typically includes complementary information, such as an extended identification data, industrial and economic classifications, the rating of outstanding debts and the market value of shares. Financial services companies, such as Thomson-Reuters or Standard & Poor’s respectively sell “Datastream” and “Compustat” databases, two examples amongst others of such product. Analysts routinely access financial reports via databases.

Attributes examined by financial analysts are hierarchically linked: the meaning of one depends on the meaning of others higher up in a hierarchy (Figure 1). The top attribute, which allows all the others to be meaningful, is whether a report is trustworthy or not. If the report is free from

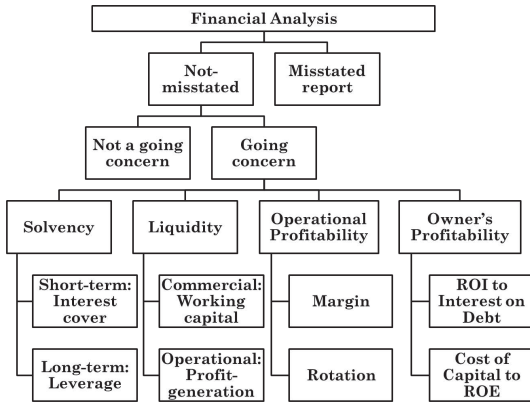


Figure 1. Uppermost dependencies in the hierarchy of financial attributes. manipulation then it may be asked whether the company is a going concern or not. Only in going concerns it makes sense to assess attributes, such as Liquidity, which also are at the root of hierarchies.

Knowledge discovery in databases of financial reports is the process of assigning each company in the database a set of logical classes/numerical values pertaining to attributes forming taxonomies similar to those of Figure 1. The assignment process is carried out using a corresponding set of models which, in turn, are built using “supervised learning” where algorithms learn to recognise classes from instances where diagnostics are already made; but unsupervised learning is also possible [4]. When completed, such process greatly facilitates the task of analysts, allowing them to concentrate on companies and conditions where algorithms may not be able to produce accurate diagnostics. If, for most of the attributes, the modelling is unreliable then knowledge discovery is of little use. Such is the present situation, where only one of the many attributes analysts work with is predicted accurately.

Financial analysis of a company is typically based on the comparison of two monetary amounts taken from published reports. For instance, when a company’s net income at the end of a given period is compared with assets required to generate such income, an indication of “Profitability” emerges. Pairs of items are often expressed in the form of a single value, their ratio. Since the size effect is similar for all items taken from the same company and period, size cancels out when a ratio is formed. Thus, ratios may be used to compare companies of different sizes [5]. Besides their size-removal ability, ratios directly measure attributes, which are implicit in reported statement numbers. Profitability, for instance, is identified as a specific ratio. Thus, the use of ratios has extended to cases where size-removal is not the major goal. Indeed, ratios are used because they embody the knowledge, which analysts possess [6].

Financial analysis is a rewarding albeit burdensome exercise. In the hands of an experienced analyst, a trustworthy financial report reveals the true condition of a company. Attempts to extract knowledge from such rich content did not succeed probably due to the very success of analysts. When trying to build automated, knowledge

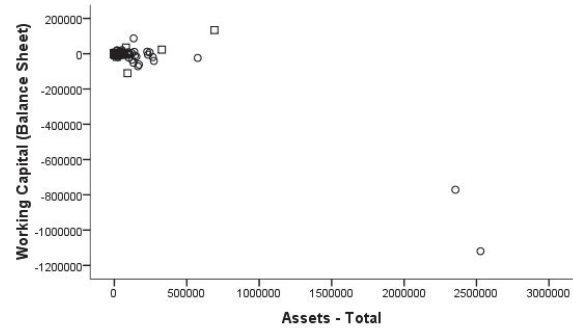


Figure 2. Influential cases in a scatter-plot of two ratio components, where some 3,000 cases are concentrated in a small region.

discovery algorithms applicable to databases, authors tend to imitate analysts namely in the use of ratios. But in spite of being the chief tool of analysts, ratios are inadequate to analytical knowledge discovery for two reasons: first, their statistical behaviour is atypical; second, they are themselves knowledge, focused pieces of knowledge, not just data.

Ratios are inadequate firstly because monetary amounts taken from financial reports, as well as ratios formed from them, obey a multiplicative law of probabilities, not an additive law. Ratio components, indeed any figure reported in a given set of accounts, are accumulations and, as such, they obey a specific generative mechanism where distributions are better described by the Lognormal and other similar functions with long tails (influential cases) and inherent heteroscedasticity [5]. Where the multiplicative character of financial statement data is ignored, any subsequent effort to model such data is fruitless, not so much because Ordinary Least Squares (OLS) or other assumptions are violated but due to the distorting effect of influential cases (Figure 2) and heteroscedasticity. And when predictive performance is the issue, the use of robust algorithms is not recommended because the cost of such robustness is lessened performance. Amongst the three basic types of measurement, Nominal, Ordinal and Scalar, the latter is the richest in content. When a scale is treated as an ordered sequence (as is the case of most robust algorithms), a great deal of content is lost.

In the second place, the use of ratios in knowledge-discovery entails a contradiction. When a ratio is chosen instead of other ratios, knowledge is involved. Each ratio embodies the analyst’s knowledge that, when two monetary amounts are set against each other, a hidden attribute is evidenced. Ratios, therefore, convey previously known knowledge.

Analysts use ratios because they can assess only one piece of information at a time. They are not able to jointly deal with collections of distributions, their moments and variance-covariance matrices as algorithms do. Analysts need focus, machines don’t. Predictive models can only lose by mimicking analysts’ requirements of separation of knowledge in small bits in order to rearrange it in a recognisable way. As explained in the coming section, algorithms are able to choose amongst a set of monetary amounts, those leading to optimal models. In doing so, algorithms build their own representations in a way similar

to analysts' task of selecting, amongst innumerable combinations of monetary amounts, the pair which highlights a desired attribute.

### III. THEORETICAL CONSIDERATIONS

Studies on the statistical characteristics of reported monetary amounts brought to light two facts. First, in cross-section the probability density function governing such amounts is nearly lognormal. Second, amounts taken from the same set of accounts share most of their variability as the size effect is prevalent [5]. Thus, variability of logarithm of account  $i$  from set of accounts  $j$ ,  $\log x_{ij}$ , is explained as the size effect  $s_j$ , which is present in all accounts from  $j$ , plus some residual variability  $\varepsilon_i$ :

$$\log x_{ij} = \mu_i + s_j + \varepsilon_i \quad (1)$$

$\mu_i$  is an account-specific expectation. Formulations such as (1), as well as the underlying random mechanism, apply to accumulations only. Accounts, such as Net Income, Retained Earnings and others, which can take on both positive- and negative-signed figures, are a subtraction of two accumulations. Net Income, for instance, is the subtraction of Total Costs from Revenue, two accumulations, not the direct result of a random mechanism.

Given two accounts  $i = 1$  and  $i = 2$  (Revenue and Expenses for instance) and the corresponding reported amounts  $x_1$  and  $x_2$  from the same set, the logarithm of the ratio of  $x_2$  to  $x_1$  is

$$\log \frac{x_2}{x_1} = (\mu_2 - \mu_1) + (\varepsilon_2 - \varepsilon_1) \quad (2)$$

It is clear why ratios formed with two accounts from the same set are effective in conveying information to analysts: the size effect,  $s_j$ , cancels out when a ratio is formed. In (2), the log-ratio has an expected value  $(\mu_2 - \mu_1)$ . The median ratio  $\exp(\mu_2 - \mu_1)$  is a suitable norm against which comparisons may be made while  $\exp(\varepsilon_2 - \varepsilon_1)$  indicates the deviation from such norm observed in  $j$ . Ratios thus reveal how well  $j$  is doing no matter its size. For instance, if the median of Net Income to Assets ratio is 0.15, any company with one such ratio above 0.15, no matter small or large, is doing better than the industry.

In (2), upward or downward deviations from the log of the industry norm are the result of subtracting two residuals, each of them size- and account type-free. The deviation  $\varepsilon_2 - \varepsilon_1$  from industry norms/benchmarks plays the crucial role of conveying to analysts the size-free, company-specific data they seek. It is clear, however, that  $\varepsilon_2 - \varepsilon_1$  is only part of the size-free, company-specific information available in  $x_1$  and  $x_2$ . When the ratio is formed, all variability common to  $x_1$  and  $x_2$  is removed. Residuals  $\varepsilon_1$  and  $\varepsilon_2$  are uncorrelated and the size-free, company-specific information contained in  $x_1$  and  $x_2$  but not conveyed by  $\varepsilon_2 - \varepsilon_1$  is the variable orthogonal to  $\varepsilon_2 - \varepsilon_1$ , which is  $\varepsilon_2 + \varepsilon_1$  [7]. Therefore,  $\varepsilon_2 + \varepsilon_1$  is size-free information not conveyed by the ratio.

It is thus demonstrated that the exclusive use of ratios as model predictors curbs the information offered to the algorithm. Only one dimension of the size-free information,

$\varepsilon_2 - \varepsilon_1$ , is made available while the other dimension,  $\varepsilon_2 + \varepsilon_1$ , is ignored.

Given this, it is worth asking whether amounts directly taken from reports would not do a better job than ratios as predictors in statistical models. Such possibility is attractive but raises questions. It is attractive because predictors obeying (1) behave exceedingly well: distributions are nearly Normal, relationships are homoscedastic and influential cases, when present, are true outliers. Indirectly, log-transformed numbers allow the use of powerful algorithms which make the most of existing content. In the downside, one obvious concern is how to deal with accounts, which can take on both positive- and negative-signed figures. Logarithms can only deal with positive values.

An equally pressing concern is how to interpret coefficients of such models. Consider the usual linear relationship where  $y$  is explained by a set of predictors  $x_1, x_2, \dots$

$$y = a + b_1x_1 + b_2x_2 + \dots \quad (3)$$

If, instead of  $x_1, x_2, \dots$  log-transformed predictors obeying (1) are included in (3), such relationship becomes

$$y = A + b_1\varepsilon_1 + b_2\varepsilon_2 + \dots + (b_1 + b_2 + \dots)s_j \quad (4)$$

where  $A = a + b_1\mu_1 + b_2\mu_2 + \dots$  is a constant value and residuals  $\varepsilon_1, \varepsilon_2, \dots$  now play the role of linear predictors. The term  $(b_1 + b_2 + \dots)s_j$  apportions the proportion of  $s_j$  (size) variability required by  $y$ . Coefficients  $b_1, b_2, \dots$  are under a constraint: their summation  $b_1 + b_2 + \dots$  must reflect the extent and sign of size-dependence in  $y$ ; and where  $y$  is size-independent,  $b_1 + b_2 + \dots$  must be zero so as to bar information conveyed by  $s_j$  from entering the relationship.

Suppose, for instance, that  $y$  is indeed size-independent. Moreover,  $y$  is being predicted by two accounts only,  $x_1$  and  $x_2$ . In this case  $b_2 = -b_1 = b$  and (4) becomes  $y = a + b(\mu_2 - \mu_1) + (\varepsilon_2 - \varepsilon_1)$  or

$$y = a + b \log \frac{x_2}{x_1} \quad (5)$$

In other words, a ratio is automatically formed so that size is removed from the relationship modelling  $y$ . Given the variety of companies' sizes found in cross-section relationships, the predictive power of  $s_j$  on  $y$  is, in most practical cases, small or non-existent. In such type of models  $b_1 + b_2 + \dots$  coefficients will indicate, not so much the strength and sign of the relationship between the  $\varepsilon$  and  $y$  but the amount of size-related variability, which is being allocated to a given predictor in order to counterbalance size-related variability from other predictors so that  $y$  is modelled by size-independent or nearly size-independent variability. When building an optimal model, the modelling algorithm assigns the role of denominator to some predictors (negative-signed coefficients) and that of numerator to others. Logarithmic representations similar to financial ratios are thus formed. In this way, financial attributes are modelled without the intervention of the analyst. This is a notable trait of the methodology.

The second concern, how to deal with accounts, which can take on both positive- and negative-signed amounts,

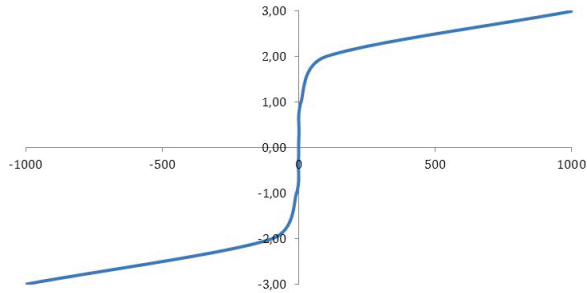


Figure 3. The x-axis represents  $x$  and the y-axis represents the log-modulus of  $x$ .

may be solved by using the “log-modulus” [3] or other similar transformation. Given variable  $x$ , the log-modulus consists of using

$$\text{sgn } x \log(|x| + 1) \quad (6)$$

instead of  $x$  (Figure 3). In this way, accumulations or subtractions of accumulations, no matter their sign, become statistically well-behaved.

The coming section shows that models using the log-modulus as predictors no longer need the support of analysts (who, when selecting appropriate ratios, apportion substantial knowledge into the model) and perform better than those using ratios. Internal representations tend to assume the form of ratios in log-space because instances used in the learning of the algorithm greatly differ in size while the attribute to be predicted is indeed predictable. Models thus tend to self-organize themselves into size-independent linear combinations of predictors, efficient in predicting classes of the attribute.

Another advantage of the log-modulus transformation is that it considerably reduces the frequency of missing values in random samples. Besides reducing the power of samples, missing values in predictors are a source of bias to models because the probability that a reported number be missing often is correlated to the attribute being predicted. For instance, it is frequent to find amounts of zero in dividends and other accounts. When ratios are formed with such values in the denominator, as is the case of the ratio “Changes in Dividends relative to Previous Year”, a missing case is created. Moreover, such missing case is correlated with the paying or not of dividends, an important predictor of Earnings’ increases.

The log-modulus transformation solves this problem. Changes in relation to the previous period, for instance, are expressed in log-modulus as

$$\delta \log x = \log x_{t-1} - \log x_t \quad (7)$$

where  $t$  and  $t-1$  express subsequent time periods and the operator  $\log$  refers to (6). Since the log-modulus transformation is continuous and monotonic, changes expressed as in (7) do not generate new missing values.

Incidentally, unlikely ratios, such changes are never ambiguous: assumed values cannot have two meanings. This is not the case with ratios where negative-valued

numerators and denominators lead to the same ratio sign as positive-valued numerators and denominators.

#### IV. CONTROLLED EXPERIMENTS

This section compares the predictive performance of ratios with that of log-modulus-transformed amounts taken from financial reports. Class proportions, period, industry, company size, the algorithm used and other characteristics, are similar for the two models being compared so as to equalise their influence on performance. The only differing characteristic is the type of predictor used. The modelling algorithm used throughout is the Binary Logistic Regression from the SPSS package.

Three experiments are performed respectively on the prediction of

- 1) bankruptcy, [8][1]
- 2) fraud [9][10][2]
- 3) and Earnings [11][12]

As depicted in Figure 1, bankruptcy and fraud are two basic attributes of financial analysis, directly influencing the way all other attributes are interpreted. As for Earnings, it is a good example of an attribute occupying a place further down in the hierarchy. Of the three, only bankruptcy prediction is reliable; in spite of the large research effort devoted to improving fraud detection, until today results are below the feasibility level, at 75% out-of-sample correct classification and highly unbalanced. All the previous literature, namely papers cited above, use ratios.

The first experiment replicates Altman’s bankruptcy predicting model [8]. A total of 2,997 cases of US bankruptcy filings is drawn from the UCLA-LoPucki Bankruptcy Research Database [13]. Bankruptcies but the first in each company are discarded as well as cases about which detailed financial figures are not available. Two random samples of nearly 900 different cases each are drawn from the remaining (nearly 2,200) bankruptcies. The two samples contain companies listed in US exchanges and present in the Standard & Poor’s “Compustat” database. They span the period 1979-2008. All sizes (Log-Total Assets deciles) and all the 24 “Global Industry Classification Standard” (GICS) groups are significantly represented in samples. Cases in the two samples are matched with an equal number of records from non-bankrupt companies. Pairing is based on the GICS group, on size decile and on year. Among financial statements fulfilling the pairing criteria, one case is randomly selected for matching and then such case is made unavailable for future matching. Although the same case is not used to match more than one bankruptcy case, cases from the same company in different years are allowed to be available for matching. The two matched samples have nearly 1,800 cases each. One of the two samples, always the same, is used as the learning-set and the other as the test-set. Due to missing observations, samples contain less than 1,800 cases:

Learning-set: non-bankrupt	845 (50.1%)
Learning-set: bankrupt	841 (49.9%)
Test-set: non-bankrupt (N)	837 (49.8%)
Test-set: bankrupt (P)	845 (50.2%)

TABLE I. BANKRUPTCY PREDICTION.

Bankruptcy predicting models	Ratios	Log-modulus
Non-bankrupt correct (TN)	782 (93.6%)	822 (98.2%)
Non-bankrupt incorrect (FP)	55 (6.4%)	15 (1.8%)
Bankrupt correct (TP)	819 (96.9%)	814 (96.3%)
Bankrupt incorrect (FN)	26 (3.1%)	31 (3.7%)
Precision: TP / (TP + FP)	93.71%	98.19%
Pseudo R-Square	0.595	0.693
Chi-Square	1526, 5 df	1993, 5 df

Two models are then built and tested. The first model uses Altman’s 5 ratios [8] as predictors while the second uses log-modulus of 5 accounts selected by the algorithm among the whole set. Test-set results for models using ratios and log-modulus are compared in Table I.

As mentioned, bankruptcy prediction is the sole case of successful modelling of financial attributes using ratios. This is due to the fact that the relationship is strong: along the last centuries, financial reports were perfected so as to highlight solvency problems. Also, Altman uses a small sample (thus limiting variability) and discarded the most notorious outliers. Even so, when the log-modulus methodology is used, performance improves and the proportion of explained variability (Cox and Snell Pseudo R-Square), as well as the overall significance of the model (Chi-Square), both increase markedly.

Log-modulus and coefficients in the model are:

Cash and Short Term Investments	+2,473
Total Liabilities	-3,532
Retained Earnings	+0,222
Tax Expense	+0,375
Cash-Flow from Operations	+0,269
Constant term	+7,129

Therefore, Total Liabilities plays the role of a denominator to the other four predictors in internally-generated linear combinations similar to ratios. Coefficients add to  $-0.193$ ; such variability models the size effect.

The second experiment replicates the fraud predicting model of Beneish [9]. The methodology is similar to the above bankruptcy-prediction case. Data used for learning and testing models consists of a collection of 3,403 “Accounting and Auditing Enforcement Releases” resulting from investigations made by the US Securities and Exchange Commission. The database is from the Centre for Financial Reporting and Management of the Haas School of Business (University of California) [14]. It contains enforcement releases issued between 1976 and 2012 against 1,297 companies which had manipulated financial reports. After removing cases for which no detailed financial data is available, the database contains 1,152 releases. Manipulated reports from the same company in different years are not removed from the sample. Enron, for instance, was the object of 6 releases and all of them are included. Two random samples of nearly 550 different cases each are then drawn. The two samples contain companies listed in US exchanges and which are present in the Standard and Poor’s “Compustat” database. They span the period 1976-2008. All sizes and all GICS groups are significantly represented. The two samples are matched with an equal number of reports from companies, which are neither the object of

TABLE II. FRAUDULENT REPORT PREDICTION.

Fraud predicting models	Ratios	Log-modulus
Non-fraud correct (TN)	244 (69.1%)	303 (85.8%)
Non-fraud incorrect (FP)	109 (30.9%)	50 (14.2%)
Fraud correct (TP)	328 (79.8%)	371 (90.5%)
Fraud incorrect (FN)	83 (20.2%)	39 (9.5%)
Precision: TP / (TP + FP)	75.1%	88.1%
Pseudo R-Square	0.305	0.569
Chi-Square	266, 8 df	617, 8 df

releases throughout the period nor bankrupt in the same year. Pairing is based on the GICS group, on size decile and on year. Amongst reports from companies fulfilling the pairing criteria, randomly selected cases for matching are made unavailable for future matching. Although the same case is not used to match more than one release case, cases from the same company in different years are allowed to remain as candidates to matching. Matched samples have nearly 1,100 cases each. One of the two samples, always the same, is used to build models and the other to test performance of models. Due to missing observations, the size of samples available for model-building and model-testing is less than 1,100 cases:

Learning set: non-fraud cases	335 (45.7%)
Learning set: fraud cases	398 (54.2%)
Test set: non-fraud cases (N)	353 (46.2%)
Test set: fraud cases (P)	411 (53.8%)

Two models are then built and tested. One of the models uses the 8 Beneish ratios [9] while the other uses 8 log-modulus selected by the algorithm. Since, in this case, some Beneish ratio components refer to the previous period, log-modulus are also allowed to express changes in relation to the previous period as in (6) and the algorithm has selected two such changes. Test-set results for models using ratios and log-modulus are compared in Table II.

Performance observed in the model using ratios agrees with that reported in the literature. The model using log-modulus shows a substantial increase in out-of-sample performance. Besides a clearly lower performance, the model using ratios introduces imbalance in the recognition of classes: misclassification in non-fraudulent cases is significantly higher than in fraudulent cases. It is also worth noting the proportion of explained variability (Cox and Snell Pseudo R-Square) and the Chi-Square of the model, which are less than half of that in the log-modulus model. Clearly, the latter fully uses the available variability while the former only uses a limited portion of it.

The third and last experiment involves the prediction of the sign of unexpected changes in Earnings per Share (EPS) one year ahead [11]. The characteristic features of this experiment are the large number of available cases (unexpected Earnings changes one year ahead can be estimated from the database), a weak relationship, indeed the weakest of the three relationships modelled and the absence of matching. The emphasis is placed on comparing the effect of unbalanced samples.

After estimating the classes to be predicted, a number of records is put aside, namely cases with missing values in the predicted dichotomous variable or in predictors. A total of nearly 140,000 cases remain, where some 90,000 are

TABLE III. INCREASE IN EPS PREDICTION.

EPS predicting models	Ratios	Log-modulus
EPS non-increases correct (TN)	42,006 (97.4%)	35,783 (85.7%)
EPS non-increases incorrect (FP)	1,101 (2.6%)	5,967 (14.3%)
EPS increases correct (TP)	4,725 (20.5%)	16,153 (70.8%)
EPS increases incorrect (FN)	18,378 (79.5%)	6,658 (29.2%)
Precision: TP / (TP + FP)	81.1%	73.0%
Pseudo R-Square	0.061	0.342
Chi-Square	4,191, 8 df	27,263, 8 df

non-increases and 50,000 are increases. Methodology has been detailed in previous experiments. The final number of cases in the learning- and test-set is:

Learning set: EPS non-increases 43,242 (64.7%)  
 Learning set: EPS increases 23,560 (35.3%)  
 Test set: EPS non-increases (N) 43,107 (65.1%)  
 Test set: EPS increases (P) 23,103 (34.9%)

Class proportions are significantly imbalanced; both the modelling process and the interpretation of results should reflect such imbalance [15].

From these samples, two models are built and tested. One of the models uses 8 ratios from previous authors [11] and the other uses a set of 8 log-modulus selected by the algorithm. Since, in this case too, some of the ratio components refer to the previous period, log-modulus are also allowed to express changes as in (6) and the algorithm has indeed selected two such changes. Test-set results for models using ratios and log-modulus are compared in Table III.

In this case, classification results should be interpreted in the light of the initial imbalance of classes in the training-set [15], which is 15.1%. For example, a classification accuracy of 70.6%, obtained from an initial imbalance of 15.1% means a gain, in relation to a classification made at random (without any previous information) of just  $5.5\% = 70.6\% - (50\% + 15.1\%)$ .

Ratios lead to an extremely small percentage of false-positives while the percentage of false-negatives is very high. The model is almost blind to unexpected increases in EPS while recognising decrease very sharply. Therefore, similarly to the previous experiment, the model based on ratios tends to amplify class imbalances.

The examination of the overall significance of the model and the proportion of explained variability shows conclusively that ratios fail to use much variability, which is clearly useful for the modelling of the relationship. This may also explain their notorious inability to produce balanced models: wherever there is neglected variability there is a bias.

## V. CONCLUSION

Till the present day, effective KDD of financial reports has proved to be an elusive goal except in the case of bankruptcy prediction, just one of the many attributes involved in Financial Analysis. Log-modulus, not requiring previous knowledge while apportioning all the available variability, may overcome this stalemate.

Predictive models based on ratios incorporate knowledge from the analyst, who is required to select appropriate

ratios capable of apportioning information needed to recognise specific attributes. Therefore, the modelling process is not fully automated. Log-modulus, by contrast, allow full KDD since the algorithm generates internal representations similar to ratios. It was shown that the modelling algorithm builds linear combinations of predictors able to unveil financial attributes, such as Solvency or Profitability.

It was also shown that the proposed methodology circumvents most of the difficulties associated with ratios when used as predictors in statistical models, namely the curtailing of variability apportioned by ratio components and the generation of missing cases. Finally, the use of controlled experiments has demonstrated that the log-modulus, agreeing with the statistical characteristics of data being modelled, perform better than ratios, delivering more accurate, robust and balanced models.

## ACKNOWLEDGMENT

This research is sponsored by the Foundation for the Development of Science and Technology of Macau (FDCT), China, under the project number 044/2014/A1.

## REFERENCES

- [1] D. Bellovary and M. Akers, "A Review of Bankruptcy Prediction Studies: 1930-present," *Journal of Financial Education*, vol. 33, pp. 1–42, 2007.
- [2] P. Dechow, C. Larson and R. Sloan, "Predicting Material Accounting Misstatements," *Contemporary Accounting Research*, vol. 28, no. 1, pp. 17–82, 2011.
- [3] J. John and N. Draper, "An Alternative Family of Transformations," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 29, no. 2, pp. 190–197, 1980.
- [4] S. Huang, R. Tsaih and F. Yu, "Topological Pattern Discovery and Feature Extraction for Fraudulent Financial Reporting," *Expert Systems with Applications*, vol. 41, no. 9, pp. 4360–4372, 2014.
- [5] S. McLeay and D. Trigueiros, "Proportionate Growth and the Theoretical Foundations of Financial Ratios," *Abacus*, vol. XXXVIII, no. 3, pp. 297–316, 2002.
- [6] W. Beaver, "Financial Ratios as Predictors of Failure," *Journal of Accounting Research, Supplement. Empirical Research in Accounting: Select Studies*, vol. 4, pp. 71–127, 1966.
- [7] D. Trigueiros, "Incorporating Complementary Ratios in the Analysis of Financial Statements," *Accounting, Management and Information Technologies*, vol. 4, no. 3, pp. 149–162, 1994.
- [8] E. Altman, *Corporate Financial Distress*. Wiley (New York), 1983.
- [9] M. Beneish, "The Detection of Earnings Manipulation," *Financial Analysts Journal*, vol. 55, no. 5, pp. 24–36, 1999.
- [10] A. Sharma and P. Panigrahi, "A Review of Financial Accounting Fraud Detection based on Data Mining Techniques," *International Journal of Computer Applications*, vol. 39, no. 1, pp. 37–47, 2012.
- [11] J. Ou and S. Penman, "Financial Statement Analysis and the Prediction of Stock Returns," *Journal of Accounting and Economics*, vol. 11, no. 4, pp. 295–329, 1989.
- [12] J. Ou, "The Information Content of Non-Earnings Accounting Numbers as Earnings Predictors," *Journal of Accounting Research*, vol. 28, no. 1, pp. 144–163, 1990.
- [13] url: <http://lopucki.law.ucla.edu/> Retrieved: Nov. 2015
- [14] url: <http://groups.haas.berkeley.edu/accounting/faculty/aaerdataset/> Retrieved: Nov. 2015
- [15] N. Chawla, "Data Mining for Imbalanced Datasets: an Overview," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2005, pp. 853–867.