DATA MINING IN FINANCE AND ACCOUNTING: A REVIEW OF CURRENT RESEARCH TRENDS

Efstathios Kirkos[1] Yannis Manolopoulos[2]

[1] Department of Accounting
Technological Educational Institution of Thessaloniki, Greece
tel 2310791209 email stkirk@acc.teithe.gr

[2] Department of Informatics
Aristotle University of Thessaloniki, Greece
tel 2310991912 email: manolopo@csd.auth.gr

ABSTRACT

Data mining tools become important in finance and accounting. Their classification and prediction abilities enable them to be used for the purposes of bankruptcy prediction, going concern status and financial distress prediction, management fraud detection, credit risk estimation, and corporate performance prediction. This study aims to provide a state-of-the-art review of the relative literature and to indicate relevant research opportunities.

Keywords: Data mining, finance, accounting, auditing

1. INTRODUCTION

Data Mining (DM) is a well honored field of Computer Science. It emerged in late 80's by using concepts and methods from the fields of Artificial Intelligence, Pattern Recognition, Database Systems and Statistics, DM aims to discover valid, complex and not obvious hidden information from large amounts of data. For this reason, another equivalent term for DM is Knowledge Discovery in Databases (KDD), which is equally often met in the literature.

Financial data are collected by many organizations like banks, stock exchange authorities, taxation authorities, big accounting and auditor offices specialized data bases, etc and in some cases are publicly available. The application of DM techniques on financial data can contribute to the solution of classification and prediction problems and facilitate the decision making process. Typical examples of financial classification problems are corporate bankruptcy, credit risk estimation, going concern reporting, financial distress and corporate performance prediction.

The importance of DM in finance and accounting has been recognized by many organizations. The American Institute of Chartered Public Accountants has identified data mining as one of the top ten technologies for tomorrow and the Institute of Internal Auditors has listed DM as one of the four research priorities (Koh, 2004).

Research on DM in finance and accounting and the application of its outcomes is a relatively new research field. The aim of the present study is to provide a state-of-the-art review about current research efforts on applying DM in finance and accounting. This review introduces the reader to specific topics concerning research objectives and methods employed. In particular this study tries to address the following questions:

- What are the specific financial application areas to which DM methods have been applied?
- What DM methods have been applied and to what extend. Do these methods outperform previous more traditional methods?
- Over what kind of data do the methods operate? Are sample sizes satisfactory large? What are the applied feature selection methods?
- What are the relative performance metrics considerations?

Such a study helps the researcher to avoid overlapping efforts and benchmark his/her practices against new developments. Another aim of this study is to indicate fertile areas for further research work in the area.

The remaining part of this work is organized as follows. Section 2 mentions the literature sources. Section 3 gives brief descriptions of the DM methods applied in the collected literature. Section 4 refers to specific applications and research studies. Finally, in Section 5 a critical appraisal and future research issues are reported. Conclusions can be found in Section 6.

2. THE LITERATURE SEARCH

For finding the studies concerning the application of DM techniques in finance and accounting we investigated the journals of four publishing houses: Elsevier, Emerald, Kluwer and Wiley. Relative articles have been found in the journals:

- Asia Pacific Financial Markets.
- Decision Support Systems,
- European Journal of Operational Research,
- Expert Systems with Applications,
- Intelligent Systems in Accounting, Finance & Management,
- International Journal of Accounting Information Systems,
- Journal of Forecasting,
- Knowledge Based Systems,
- Management Decision,
- Managerial Auditing Journal,
- Managerial Finance,
- Neural Networks, and
- Omega The International Journal of Management Science.

3. THE METHODS USED

The term Data Mining methods stands for a large number of algorithms, models and techniques derived from the osmosis of statistics, machine learning, databases and visualization. Several of these methods have been applied for examining financial data. Popular DM methods that will be mentioned in this study are Neural Networks, Genetic Algorithms, Decision Trees, Rough Set Theory, Case Base Reasoning and Mathematical Programming.

3.1 Neural Networks

Neural Networks (NN) is a mature technology with established theory and recognized applications areas. A NN consist of a number of neurons, i.e. interconnected processing units. Associated with each connection is a numerical value called “*weight*”. Each neuron receives signals from connected neurons. If the combined input signal strength exceeds a threshold, then the neuron fires. The input value is transformed by the transfer function of the neuron.

The neurons are arranged into layers. A layered network consists of at least an input (first) and an output (last) layer. Between the input and output layer there may exist one or more hidden layers. Different kinds of NNs have a different number of layers. Self-organizing maps (SOM) have only an input and an output layer, whereas a backpropagation NN has additionally one or more hidden layers.

After the network architecture is defined, the network must be trained. In backpropagation networks a pattern is applied to the input layer and a final output is calculated at the output layer. The output is compared with the desired result and the errors are propagated backwards in the NN by tuning the weights of the connections. This process iterates until an acceptable error rate is reached. The backpropagation NNs have become popular for prediction and classification problems.

SOM is a clustering and visualization method of unsupervised learning. For each input vector, only one output neuron will be activated. The winner’s weight vector is updated to correspond with the input vectors. Thus, similar inputs will be mapped to the same or neighboring output neurons forming clusters. Two commonly used SOM topologies are the rectangular lattice, where each neuron has four neighbors and the hexagonal lattice where each neuron has six neighbors.

An important disadvantage of NNs is that they act as black boxes as it is difficult to humans to interpret the way NNs reach their decisions. However, algorithms have been proposed to extract comprehendible rules from NNs. Another criticism on NNs is that a number of parameters like the network topology must be defined empirically.

It seems that NNs attract the interest of most researchers in the area of our concern. Their structure and working principles enable them to deal with problems where an efficient algorithm based solution is not applicable. Since they learn from examples and generalize to new observations they can classify previously unseen patterns. They have the ability to deal with incomplete, ambiguous and noisy data. Unlike traditional statistical techniques they do not assume a priori about the data distribution properties, neither they assume independent input variables.

3.2 Genetic Algorithms

Genetic Algorithms (GA) apply ideas from the natural evolution where the fittest individuals survive. Rules concerning a problem are encoded as a set of strings each of which is composed of bits. These strings form a population. GA allows the strings with the highest fitness value to survive and proliferate renewing the population.

A chromosome is a particular string representing a point in the solution space. Population is a set of chromosomes. After the random creation of the initial population each chromosome is evaluated using a user-defined fitness function. The role of the fitness function is to evaluate the performance of the chromosome.

Three operators are applied to chromosomes.

- *Reproduction,* where the individuals self proliferate by replicating themselves with a probability analogous to their fitness value
- *Crossover,* where two chromosomes mutually exchange some bits creating new chromosomes
- *Mutation,* which operates on a single chromosome by changing one or more bits. The probability of mutation is very low.

3.3 Decision Trees

Decision Trees is a classification and prediction method, which successively divides observations into mutually exclusive subgroups. The method searches for the attribute that best separates the samples into individual classes. Subgroups are successively divided until the subgroups are too small or no significant statistical difference exists between candidate subsets. If the decision tree becomes too large it is finally pruned.

3.4 Rough Set Theory

Rough Set Theory (RST) was introduced by Pawlak (1982). RST extents set theory with the notion of an element's possible membership in a set. Given a class C, the lower approximation of C consists of the samples that certainly belong to C. The upper approximation of C consists of the samples that can not be defined as not belonging to C. RST may be used to describe dependencies between attributes, to evaluate significance of attributes, to deal with inconsistent data and to handle uncertainty (Dimitras et al.1999).

3.5 Case Base Reasoning

Case Base Reasoning (CBR) is a reasoning problem solving method. For solving a problem, CBR tries to retrieve a similar case from a case base. Key issues in CBR are the similarity measure and the retrieval of a similar case. Popular matching techniques are k-nearest neighbor (k-NN), inductive learning and knowledge guide. In its simplest version, k-NN assesses the similarity of two cases by calculating their Euclidean distance. This approach assumes that all features are equally relevant. Since this is not always the case, improved algorithms introducing weighted features have been proposed.

4. APPLICATION AREAS AND SPECIFIC RESEARCH STUDIES

Due to their classification and prediction capabilities, DM techniques have been employed to facilitate the auditing process, to predict corporate performance, and to facilitate credit risk estimation.

In the field of auditing, DM techniques evolve as a promising contribution. Recent events indicate considerable problems in the auditing process. The collapse of Enron and Arthur Andersen among other and the seemingly extensively applied “book cooking” accounting practices, provide evidence for changing demands in the audit process (Koskivaara, 2004).

According to the Statement of Auditing Standards 56 (SAS 56) issued by AICPA, the auditor develops his/her own expectations and compares these expectations to recorded amounts or ratios. In accomplishing this task, the auditor employs analytical procedures which compare expected relationships among data items with actual observed relationships. Analytical procedures allow the examination of the accuracy of an account’s balance without examining the relevant individual transactions. Fraser classifies the Analytical Review Techniques in Non-Quantitative (NQT) such as scanning, Simple Quantitative (SQT) such as trend, ratio and reasonableness test and Advance Quantitative (AQT) such as regression analysis and NNs (Fraser et al.1997) (Koskivaara, 2004).

A modern trend in auditing is to embrace the concept of business risk, which emphasizes the strategic objectives of a business enterprise. In this top down approach the auditor understands the strategic objectives and works downwards to business process. DM techniques such as NNs, GA, CBR and fuzzy logic may facilitate this new risk-based auditing approach (Calderon et al., 2002).

Papers related to specific application areas included in the field of auditing refer to Bankruptcy Prediction, Going Concern Prediction and Financial Distress and Management Fraud.

4.1 Bankruptcy Prediction

Bankruptcy prediction seems to be most popular topic of the application of DM techniques on financial data. Corporate bankruptcy causes economic damages for management, investors, creditors and employees together along with social cost. For these reasons bankruptcy prediction is an important issue in finance. Bankruptcy prediction by using financial statements data attracts its origin from the work of Altman in 1968. Altman argues that corporate failure is a long period process and that the financial statement data should include warning signals for the imminent bankruptcy. By applying Multiple Discriminant Analysis techniques he developed a model for bankruptcy prediction. Since the work of Altman many researchers developed alternative models by using statistical techniques (Ohlson 1980 used Logit, Zmijewski 1984 used Probit). In the last years research effort has been made to build models which use DM techniques.

Lin and McClean (2001) tried to predict corporate failure by using four different methods. Two of the methods are statistical (Discriminant Analysis and Logistic Regression), whereas the remaining two methods are Machine Learning techniques (Decision Trees – C5.0 and Neural Networks). Additionally they proposes a hybrid algorithm. Their sample included data about 1133 UK companies. 690 non failed companies and 106 failed companies were used as training set, where 289 non failed companies and 48 failed companies were used as testing set. No attempt was made to match failed and not failed companies. 37 financial ratios originated from balance sheet and income statements were selected as input variables. Two feature selection methods have been employed reducing the input variables to 4 by using human judgment and to 15 by using

ANOVA. The authors report better results for the NNs and decision trees models for both the human judgment based and the ANOVA feature selection. Finally, the authors propose a hybrid algorithm employing weighted voting of different classifiers. Marginally better performance is reported for the hybrid model.

Tung et al. (2004) employed a hybrid model integrating NNs and fuzzy systems. The model called “Generic Self-organizing Fuzzy Neural Network” was a rule-base consisting of IF–THEN fuzzy rules that can self-adjust the parameters of the fuzzy rules using learning algorithms derived from the NN paradigm. The main advantage of the fuzzy NN was mentioned to be its ability to model a problem by using easily comprehendible high level linguistic model instead of complex mathematical expressions.

The model was applied to predict bank failures. Input variables are 9 financial variables, which have been found to be significant in previous studies. The sample contained data about 2555 non failed and 548 failed banks. 20% of the data were used as training set and 80% as testing set. To reduce Type 1 errors the sample was balanced to include equal number of failed and not failed banks.

Authors report a performance of 93% when using data from the last available financial statement, 85% when using statements obtained one year prior to the last record and 75% for statements two years prior to the last available record. The model produces a set of around 50 IF-THEN fuzzy rules, which describe the interactions between the 9 selected input variables and their impact on the financial health of the observed banks.

Shin and Lee (2002) proposed a model based on GAs. The authors stressed the fact that as opposed to the NNs, GAs can produce comprehendible rules. GAs were applied to find thresholds for one or more variables above or below which a company is considered dangerous. The model used a rule structure that contains 5 conditions each of which referred to a variable out of 9 financial ratios. The conditions were combined with the logical AND operator. The data set contained 264 failed and 264 non failed firms, whereas 9 financial ratios have been selected as input variables. 90% of the sample was used for training and 10% for validation. The general reported performance was about 80%.

Kim and Han (2003) built a qualitative model based on experts problem solving knowledge. Experts work with their subjective knowledge evaluating qualitative and quantitative facts. The model used a GA method to extract decision rules from experts qualitative bankruptcy predictions. The model followed the method of the experts of a Korean commercial bank. In order to predict bankruptcy the experts evaluated 6 major risk factors. In the model a chromosome contained 6 segments representing a categorization of a firm according to the 6 risk factors. A 7th segment in the chromosome classified the firm as bankrupt or non bankrupt. The data sample contained 772 companies, half of which were bankrupt. The experts evaluated the 6 risk factors for these companies. The genetic evolution process extracted 11 bankruptcy rules. Additionally rules have been extracted by using a backpropagation NN and Inductive Learning. Rules extracted with GA are reported to have better predicting accuracy than NN and inductive learning.

Dimitras et al. (1998) applied RST for the aim of bankruptcy prediction. The training set contained data for 40 failed and 40 matched non failed Greek firms covering a period of five years. The testing set contained 19 failed and 19 non failed firms. A credit manager of a Greek bank selected 12 financial ratios to enter the information table and discretized the continuous values. The rough set analysis produced 54 reducts, each containing 5-7 attributes, the bank manager selected the one reduct and thus the remaining attributes were eliminated. Finally, the decision rules were derived. The results of the method have been compared with the results of discriminant analysis and logit analysis and have been found to prevail.

McKee (2003) compared results obtained by using RST with actual auditors' opinions for the purpose of bankruptcy prediction. The data sample included 146 bankrupt and 145 matched non bankrupt US companies. 11 predictive factors were chosen, 10 of which were financial ratios and 1 was a prior audit opinion. The rough set produces 87 reducts, each employing 4-6 variables and 2 reducts are selected. Two models of decision rules were developed. The results of the models were compared with actual auditors' signaling rates and have been found almost equal. The author concludes that the models developed in this research offered no significant comparative predicting advantage over auditors' current methodologies.

Beynon and Peel (2001) employed a development of RST: the Variable Precision RST. VPRST incorporated probabilistic Decision Rules and allowed partial classification by introducing a degree of confidence in classification. In contrast to previous research efforts where the discretization of the values had been made by humans, the author employed the FUSINTER method for the discretization purpose. The data sample contained 45 failed and 45 non failed UK industrial firms. 30 failed and 30 non failed firms form the training sample, whereas the remaining formed the holdout sample. 12 variables, 8 financial and 4 qualitative variables have been selected for the rule generation. After the reducts production and the selection of one of them, a set of 12 rules have been obtained. The results of VPRST were compared with results of Multiple Discriminant Analysis, Logit Analysis, Recursive Partitioning Algorithms Decision Trees and Elysee ordinal discriminant method. In the training and holdout sample VPRST outperformed some of these methods.

Park and Han (2002) in a CBR study developed a model to predict bankruptcy. The distance measure used weighted features. The weights were calculated by using the Analytic Hierarchy Process method (AHP). The sample included 1072 failed and 1072 non failed firms. 13 financial and 15 non financial variables were chosen for input. The authors argued that AHP/CBR performed better than pure CBR, regression CBR and logit CBR.

4.2 Going Concern and Financial Distress

According to SAS 59, the auditor has to evaluate the ability of his/her client to continue as a GC for at least one year beyond the balance sheet data. If there are indications that the client company will face financial difficulties, which may lead to failure, the auditor has to issue a going concern report. The assessment of the going concern status is not an easy task. Studies report that only a relative small proportion of failed firms have been qualified on a going concern basis (Koh 2004). To facilitate the auditors on the going concern report issuing task, statistical and machine learning techniques have been proposed.

Koh (2004) compared backpropagation NN, Decision Trees and logistic regression methods in a going concern prediction study. The data sample contained 165 going concern firms and 165 matched non going concern firms. 6 selected financial ratios have been used as input variables. The author reported that Decision Trees outperformed the other two methods.

Tan and Dihardjo (2001) built upon a previous study of Tan, which tried to predict financial distress for Australian credit unions by using NNs. In his previous study Tan used quarterly financial data and tried to predict distress in a quarter base. Tan and Dihardjo improved the method by introducing the notion of "early detector". When the model predicts that a credit union will go distressed in a particular quarter and the union actually goes distressed in a next quarter, in a maximum of four quarters, the quarter is labeled as "Early Detector". This improved method performed better than the previous one in terms of Type II errors rate. 13 financial ratios were used as input variables and a sample of 2144 observations was used. The results were compared with those of a Probit model and were found marginally better especially for the Type 1 error rate.

Konno and Kobayashi (2000) proposed a method for enterprise rating by using Mathematical Programming techniques. The method made no distribution assumptions about the data. Three alternatives based on discrimination by hyperplane, discrimination by quadratic surface and discrimination by elliptic surface were employed. 6 financial ratios derived from financial statements were used as input variables. The data sample contained 455 enterprises. The method calculated a score for each enterprise.

4.3 Management Fraud

Management fraud is the deliberated fraud committed by managers through falsified financial statements. Management fraud injures tax authorities, shareholders and creditors.

Spathis (2002) developed two models for identifying falsified financial statement from publicly available data. Input variables for the first model contain 9 financial ratios. For the second model z-score is added as input variable to accommodate the relationship between financial distress and financial statement manipulation. The method used is logistic regression and the data sample contained 38 FFS and 38 non FFS firms. For both models the results show that 3 variables with significant coefficients entered the model.

4.4 Corporate Performance Prediction

Lam (2003) developed a model to predict the return rate on common shareholders equity. She used backpropagation NNs and inferred rules from the weights of the connections by applying the GLARE algorithm. The input vector included 15 financial statement ratios and 1 technical analysis variable. In an additional experiment 11 macroeconomic variables were also included. The data sample contained 364 firms.

Back et al. (2001) developed two models to cluster companies according to their performance. Both models used SOMs. The first model operated over financial data of 160 companies. By employing text mining techniques, the second model analyzed the CEOs' annual report of the companies. The authors concluded that there are differences between the clustering results of the two methods.

Kloptchenko et al. 2004 built on the previously mentioned research effort. Two models were developed, one analyzing financial ratios and the other analyzing the CEOs' reports. In this study a different method, the Prototype-Matching Text Clustering, was used for analyzing the reports. By comparing the results of the qualitative and the quantitative methods the authors concluded that the text reports tend to foresee changes in the financial state before these changes explicitly influence the financial ratios.

4.5 Credit Risk Estimation

The task of credit risk analysis becomes more demanding due to the increased number of bankruptcies and the competitive offers of creditors. DM techniques have been applied to facilitate the estimation of credit risk.

Huang et al. (2003) performed credit rating analysis by using Support Vector Machines (SVMs), a machine learning technique. Two data sets were used; one containing 74 Korean firms and the other containing 265 US firms. For both data sets 5 rating categories were defined. Two models for Korean data set and two models for US data set, each one having a different input vector were built. SVMs and a backpropagation NNs were used to predict credit rating. SVMs performed better in the three of the four models. Another consideration of the study was to interpret the NN. The Garson method was used to measure the relative importance of the input values.

Mues et al. (2004) used decision diagrams to visualize credit risk evaluation rules. Decision diagrams have the theoretical advantage over decision trees that they avoid the repetition of isomorphic subtrees. Two data sets, one containing German data and two containing Benelux data were used. A NN was employed to perform the classification. The rule extraction methods Neurorule and Trepan were applied to extract rules from the network. Additionally C4.5, C4.5 rules and Entropy-based Oblivious Decision Graphs methods were used to produce decision trees and rules. The performance of Neurorule and Trepan was comparable with the performance of the NNs and superior of the performance of the other methods. Finally the rules were visualized in the form of decision diagrams.

5. EVALUATION AND FUTURE RESEARCH ISSUES

Finance and accounting are popular application fields for DM. The classification and prediction abilities of DM methods enables them to be used for the purposes of bankruptcy prediction, going concern status and financial distress prediction, management fraud detection, credit risk estimation, and corporate performance prediction. Auditors, credit scoring experts and investors can be facilitated in their work and gain time and cost in their decision making process.

Bankruptcy prediction seems to attract the interest of most researchers since almost half of the papers refer to this topic. The application areas of the examined literature are depicted in Table 1. The examination of the collected literature gives rise for discussion in terms of methods employed, data used and performance metrics topics.

APPLICATION AREAS	PAPERS
Bankruptcy	8
Going Concern & Financial Distress	3
Corporate Performance Clustering / Prediction	3
Credit Risk Estimation	2
Management Fraud	1

Table 1. Application areas

5.1 Methods and Models

The term DM methods includes a wide range of methods derived from Statistics, Artificial Intelligence and Databases. In the collected literature Neural Networks are the most used model. Table 2 shows the models employed.

MODEL	PAPERS
Neural Networks	8
Rough Set	3
Decision Trees	2
Genetic Algorithms	2
Hybrid	2
Case Base Reasoning	1
Mathematical Programming	1
Logistic Regression	1
Support Vector Machines	1

Table 2. Models employed.

Although many researchers stress the fact that hybrid models which combine characteristics and advantages of particular models may improve performance or interpretability, hybrid models are used in only two cases. A future research direction may be the development and application of hybrid models.

Another direction for model improvement is the enhancement of existing models with advanced algorithms. Variable Precision RST, Analytic Hierarchy Process CBR and GA that uses a niching method are examples of this case.

The design of the NN architecture is still a matter of art. The number of neurons, the number of layers and the transformation function are arbitrarily and subjectively defined. Methods that propose an optimal NN architecture for a particular case can be developed.

Although the main criticism over NN is that they act as black boxes, in only two cases effort has been made to interpret the model (algorithms GLARE, Neurorule, Trepan). Research effort can be directed towards the interpretation of the decision making pattern of NNs.

In four cases AI models are benchmarked against statistical models. AI methods have the theoretical advantage that they don't impose arbitrary assumptions on the input variables. However, the reported results of AI methods only slightly outperform the results of statistical methods. In some cases statistical models are reported to perform better. Additional research effort is required to materialize the theoretical advantages of AI models.

Visualization methods which communicate the extracted knowledge in a comprehendible way are another future research focus.

Data mining tools are isolated applications or are parts of statistical analysis software suites. Embedding DM tools in commercial databases or ERP systems may facilitate the dissemination and usage of DM tools to business professionals.

5.2 Data

The data used in the collected literature are mainly financial ratios derived from financial statements. In eight cases the input vector consists solely from financial ratios. In only one case financial ratios are not used in the input vector. Many authors refer to the need to enrich the input vector with more information. Macroeconomic variables can be included. Qualitative information such as the achievement of corporate strategic objectives, previous auditor opinion, the management experience, market information and many others can be used to capture economic, political, social and technological factors. In two papers text mining techniques are used to classify and predict corporate performance.

As recognized by the authors, in some of the examined papers the sample size is not satisfactory large. Small samples may bias the results. Furthermore there are important differences in the size of the training, testing and validation samples. Table 3 depicts sample sizes.

Sample size	Papers
> 1000	4
>500	2
>200	5
<=200	5

Table 3. Sample sizes

Financial databases in many cases contain a significant number of financial ratios. Many of these ratios contain overlapping information. Furthermore research has shown that a relative small number of ratios is adequate for the classification and prediction purposes. For these reasons feature selection is required. In seven cases researchers rely on previous studies to select the input variables. In four cases the selection is based on human judgement. The introduction of formal methods like ANOVA may improve the feature selection practices

The existence of missing values is common in financial data. Strategies for handling missing financial data like using the mean value of the given class or using the most probable value may be evaluated and proposed.

Data discretization is another issue for consideration. In some cases humans have been employed to discretize the data where in other cases discretization methods are used.

5.3 Performance Metrics

Another important consideration is the performance metrics. The performance is assessed by testing the model against a testing and possibly a validation sample.

In many cases there are no validation samples and the testing sample is used to measure the model performance. Some algorithms use the testing sample to stop the training of the model. Since this may introduce a bias (called "overfitting") it is important to measure the performance on a validation sample.

The basic accuracy is computed as the proportion of correct classifications or predictions. However, there is an additional consideration concerning the Type1 and Type 2 errors. A Type 1 error occurs when the model predicts no bankruptcy for a firm and the firm actually goes bankrupt. A Type 2 error occurs when the model predicts bankruptcy for a healthy firm. Type 1 and Type 2 errors have different costs. Type 1 errors may lead to wrong decisions that may cause financial injuries. Type 2 errors may cause just additional investigations. Thus Type 1 errors have bigger cost than Type 2 errors. Relative cost of Type I and Type II errors must be considered in performance metrics.

6. Conclusions

DM techniques have classification and prediction capabilities which can facilitate the decision making process in financial problems. The financial and prediction tasks in the collected literature address the topics of bankruptcy prediction, credit risk estimation, going concern reporting, financial distress, corporate performance prediction and management fraud. Bankruptcy prediction seems to be the most popular application area.

The data mining methods employed in the collected literature include Neural Networks, Genetic Algorithms, Decision Trees, Rough Set Theory, Case Base Reasoning and Mathematical Programming. Most of the researches seem to prefer the Neural Network model.

Although a considerable amount of research effort has address the application of DM techniques in finance there are many fertile areas for further research.

The introduction of hybrid models, the improvement of existing models, the extraction of comprehendible rules from Neural Networks, the improvement of performance and the integration of ERP systems with DM tools are some possible future research directions.

In terms of the data used the enrichment of the input vector with qualitative information and the usage and evaluation of formal methods for feature selection and data discretisation are open research possibilities.

Another consideration which requires further research is the evaluation of the relative cost of Type I and Type II errors.

The future is open. Further research effort will improve models and methods making DM an even more valuable tool in finance and accounting.

References

1. B. Back, J. Toivonen, H. Vanhatanta and A. Visa: “Comparing Numerical Data and Text Information from Annual Reports Using Self-organizing Maps”, *International Journal of Accounting Information Systems*, Volume 2, Issue 4, December , 2001, pp. 249-269 .
2. M.J. Beynon and M.J. Peel: “Variable Precision Rough Set Theory and Data Discretisation: an Application to Corporate Failure Prediction”, *Omega The International Journal of Management Science*, Volume 29, Issue 6, December, 2001, pp. 561-576.
3. T.G. Calderon and J.J. Cheh: “A Roadmap for Future Neural Networks Research in Auditing and Risk Assessment”, *International Journal of Accounting Information Systems*, Volume 3, Issue 4, December, 2002, pp. 203-236.
4. A.I. Dimitras, R. Slowinski, R. Susmaga and C. Zopounidis: “Business Failure Prediction using Rough Sets”, *European Journal of Operational Research*, Volume 114, Issue 2, April, 1998, pp 263-280.
5. I.A.M. Fraser, D.J. Hatherly, K.Z. Lin: “An empirical investigation of the use of analytical review by external auditors”, The British Accounting Review, Volume 29, Issue 1, March, 1997, pp.35-47.
6. Z. Huang, H. Chen, C.J. Hsu, W.H. Chen and S. Wu: “Credit Rating Analysis with Support Vector Machines and Neural Networks: a Market Comparative Study”, *Decision support Systems*, In Press, 2003.
7. M.J. Kim and I. Han: “The Discovery of Experts’ Decision Rules from Qualitative Bankruptcy Data using Genetic Algorithms”, *Expert Systems with Applications*, Volume 15, Issue 4, November, 2003, pp.637-646.
8. A. Kloptchenko, T. Eklud, J. Karlsson, B. Back, H. Vanharanta and A. Visa: “Combining Data and Text Mining Techniques for Analyzing Financial Reports”, *Intelligent Systems in Accounting, Finance and Management*, Volume 12, Issue 1, January/March, 2004, pp. 29-41..
9. H.C. Koh, “Going Concern Prediction using Data Mining Techniques”, *Managerial Auditing Journal*, Volume 19, No 3, 2004, pp. 462-476.
10. H. Konno and H. Kobayashi: “Failure Discrimination and Rating of Enterprises by Semi-Definite Programming”, *Asia-Pacific Financial Markets,* Volume 7, Issue 3, September, 2000, pp.261-273.
11. E. Koskivaara, “Artificial Neural Networks in Analytical Review Procedures”, *Managerial Auditing Journal*, Volume 19, No 2, 2004, pp. 191-223
12. M. Lam: “Neural Network Techniques for Financial Performance Prediction: Integrating Fundamental and Technical Analysis”, *Decision Support Systems*, In Press, 2003.
13. F.Y. Lin and S. McClean: “A Data Mining Approach to the Prediction of Corporate Failure”, *Knowledge-Based Systems*, Volume 14, Issues 3-4, June, 2001, pp. 189-195.
14. T. McKee: “Rough Sets Bankruptcy Prediction Models vs. Auditor Signalling Rates”, *Journal of Forecasting*, Volume 22. Issue 8, December, 2003, pp.569-586.
15. C. Mues, B. Baesens, C.M. Files and J. Vanthienen: “Decision Diagrams in Machine Learning: an Empirical Study on Real-life Credit-risk Data”, *Expert Systems with Applications*, In Press, 2004.

16. C.S. Park and I. Han: “A Case-base Reasoning with the Feature Weights Derived by Analytic Hierarchy Process for Bankruptcy Prediction”, *Expert Systems with Applications*, Volume 23, Issue3, October, 2002, pp.255-264.
17. K.S. Shin and Y.J. Lee: “A Genetic Algorithm Application in Bankruptcy Prediction Modeling”, *Expert Systems with Applications*, Volume 23, Issue 3, October, 2002, pp.321-328.
18. C. Spathis: “Detecting False Financial Statements Using Published Data: some Evidence from Greece”, *Managerial Auditing Journal*, Volume 17, No 4, 2002, pp.179-191.
19. C.N.W. Tan and H. Dihardjo: “A Study on Using Artificial Neural Networks to Develop an Early Warning Predictor for Credit Union Financial Distress with Comparison to the Probit Model”, *Managerial Finance*, Volume 27, No 4, 2001, pp.56-77.
20. W.L. Tung, C. Quek and P. Cheng: “GenSO-EWS: a Novel Neural Fuzzy Based Early Warning System for Predicting Bank Failures”, *Neural Networks*, Volume 17, Issue 4, May, 2004, pp. 567-587.