Bankruptcy Prediction with Missing Data

Q. Yu[1], **Y. Miche**[1], **A. Lendasse**[1] **and E. Séverin**[2]
[1]Department of Information and Computer Science, Aalto University, Espoo, Finland
[2]Department GEA, Université Lille 1, Lille, France

Abstract— *Bankruptcy prediction have been widely studied as a binary classification problem using financial ratios methodologies. When calculating the ratios, it is common to confront missing data problem. Thus, this paper proposes a classification method Ensemble Nearest Neighbors (ENN) to solve it. ENN uses different nearest neighbors as ensemble classifiers, then make a linear combination of them. Instead of choosing k in original k-Nearest Neighbors algorithm, ENN provides weights to each classifier which makes the method more robust. Moreover, using a adapted distance metric, ENN can be used directly for incomplete data. In a word, ENN is a robust and a comparatively simple model while providing good performance with missing data. In the experiment section, four financial datasets which are publicly available are used to prove this conclusion.*

Keywords: Missing data, Ensemble model, Nearest Neighbors, Bankruptcy Prediction

1. Introduction

The business failure has been widely studied, trying to identify the various determinants that can affect the existence of firms. Especially due to the recent changes in the world economy and as more firms, large and small, seem to fail now more than ever. The prediction of the bankruptcy, is then of increasing importance.

In most of the studies, bankruptcy prediction is treated as a binary classification problem. The target (output) variable of the models is commonly a dichotomous variable where 'firm filed for bankruptcy' is set to 1 and 'firm remains solvent' is set to 0. The reference (input) variables are often financial ratios drawn from financial statements and include measures of profitability, liquidity, and leverage. The pioneer study using univariate statistic of financial ratios originated from Beaver (1966) [4] and Altman's work (1968) [5]. Using multivariate discriminate analysis to assess predictive power of ratio analysis, financial ratios methodologies are becoming indispensable tools for modeling, analysis and prediction. The other main steam is employing Artificial Intelligence (AI) methods, which have been applied to bankruptcy prediction problem from 1990âĂŹs, including decision tree [24], [25], fuzzy set theory [26], case-based reasoning [27], [28], genetic algorithm [29], support vector machine [30], several kinds of neural networks such as BPNN (back propagation trained neural network) [32], [31], [34], PNN (probabilistic neural networks) [33], SOM (self-organizing map) [35], [36].

However, when calculating the financial ratios, for example, from companies' annual reports, it is very common to encounter the problem of missing value [1]. Some classification methods choose to remove all the ratios (variables) and the observations (samples) which contain missing values to train the model. The drawback is that it loses data, especially when the quantity of the observations is originally small. Furthermore, the new observations with missing values are no longer predictable. On the other hand, a great number of methods have been already developed for solving the problem by filling this missing values (also named imputation), for example, Kriging [6] and several other Optimal Interpolation methods, such as Objective Analysis [7].

In this paper, a third approach is proposed: a classification model which is directly applied to datasets with missing values. In order to predict whether the target company is healthy or not, this method provides an Ensemble model of nearest-neighbors (ENN) aimed at solving the classification task. Since it is impossible to calculate a standard Euclidean distance in NN algorithm when the sample have missing data, this method uses a new distance metric to measure the closeness between incomplete samples[10]. Thus the observations with missing values can be used to train the model, and moreover, the incomplete new observations can also be predicted. Besides, how to choose the suitable 'k' is always an issue when using KNN methods. In this paper, an ensemble method [9] is used. Instead of choosing a specific k, different nearest neighbors are treated as several different classifiers. The ensemble strategy assigns different weights to each classifier and then a linear combination of the nearest neighbors is used as the global output of the ensembles. This method is robust as the ensemble of classifiers has a smaller variance than each single classifier and then will reach better prediction performances [23].

The following section introduces the ensemble concept in general and particular strategy used in this paper with k Nearest Neighbors. It is followed by a presentation of incomplete data problem and a feature-weighted distance metric measurement in Section 3. In Section 4, four data sets are performed using random interpolation of missing

[1]Missing data, or missing values, occur when no data value is stored for the variable in the current observation. If a input data has N observations (samples) with d dimensions (variables). Then, when we say a missing data in this data, it implies one missing point among the original $(N * d)$ points.

data and the Monte-Carlo cross test. Finally, Section 5 shows more discussion about the experiment results and some conclusion.

2. Ensemble Nearest neighbors (ENN)

In machine learning, ensemble methods use multiple models to obtain better predictive performance than could be obtained from any of the constituent models [1], [2], [3]. It is a supervised learning algorithm, because it can be trained and then used to make predictions. Empirically, ensembles tend to yield better results when there is a significant diversity among the models [11], [12]. Many ensemble methods, therefore, seek to promote diversity among the models they combine [13]. On the other hand, ensembles can be shown to have more flexibility in the functions they can represent. This flexibility can, in theory, enable them to over-fit the training data more than a single model would, but in practice, some ensemble techniques (for example bagging) tend to reduce problems related to over-fitting of the training data. In the following from this section, more details are presented about ensemble of different k nearest neighbors.

2.1 The classifiers used for ensembles

An effective way to improve a classification method's performance is to create ensembles of classifiers. Two elements are believed to be important in constructing an ensemble: (a) the performance of each individual classifier; and (b) diversity among the classifiers. Therefore, different k nearest neighbors are chosen to perform such tasks.

In the original k-NN algorithm, the main difficulty is how to choose k properly. To solve this problem, we use Nearest Neighbors with each specific k as classifiers in this method. Therefore, the method will choose or weight each k automatically, using the ensemble technique. Besides, k-NN algorithm itself is proved to be an efficient classifier [14]. Another advantage of using different k NN is that NN is a distance-based algorithm, which provides us the opportunity to solve missing data problem simultaneously with the corresponding distance metric.

This part is shown as step 1 in Fig 1.

2.2 Linear optimization strategy

There exist some common types of ensembles:

- Bayes optimal classifier. The Bayes Optimal Classifier is an optimal classification technique. It is an ensemble of all the hypotheses in the hypothesis space. On average, no other ensemble can outperform it, so it is the ideal ensemble [17]. Unfortunately, Bayes Optimal Classifier cannot be practically implemented for any but the most simple of problem.
- Bootstrap aggregating (bagging). It involves having each model in the ensemble vote with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly-drawn subset of the training set. As an example, the random forest algorithm combines random decision trees with bagging to achieve very high classification accuracy [18].
- Boosting. Boosting involves incrementally building an ensemble by training each new model instance to emphasize the training instances that previous models misclassified. In some cases, boosting has been shown to yield better accuracy than bagging, but it also tends to be more likely to over-fit the training data. By far, the most common implementation of Boosting is Adaboost, although some newer algorithms are reported to achieve better results.
- Linear combination. The reason for linear combination is that taking a weighted average over several models reduces the error by decreasing the variance around the target. On the other hand, linear ensemble makes the final model relatively simple and easier to interpret. Therefore, linear combination is used in this paper.

However, it is not easy to determine the weights in practice. In this paper, Non-Negative Least Square (NNLS) algorithm is used. According to Miche et al. [19], the advantage of NNLS is that it is efficient and fast. The square of the difference between the actual output and the weighted leave-one-outputs of the classifiers is minimized such that the weights ω_j are positive, as seen in Equation 1.

$$\min_{\omega_j} \|y - \sum_j y_{loo}^j \omega_j\|^2, \qquad s.t. \quad \omega_j \geq 0 \tag{1}$$

This linear combination using non-negative constraints of the weights also helps to avoid over-fitting. This part is illustrated on step 2 of Fig 1.

2.3 Leave-One-Out

LOO is an special case of k-fold cross-validation where k equals to the number of observations. In this paper, LOO method is used in the training set to get even better prediction and meantime, to reduce the risk of over-fitting. One problem with the LOO method is that it can get very time consuming, especially if the dataset tends to have a high number of observations. Fortunately, the PRESS (or PREdiction Sum of Squares) statistics provide a direct and exact formula for the calculation of the LOO error for linear models.

$$\epsilon^{\text{PRESS}} = \frac{y_i - \hat{y}_i \omega}{1 - \hat{y}_i \mathbf{P} \hat{y}_i^T}, \tag{2}$$

where $\mathbf{P}$ is defined as $\mathbf{P} = (\mathbf{Y}^T \hat{\mathbf{Y}})^{-1}$ and $\hat{\mathbf{Y}}$ is the estimated output matrix, and ω is the ensemble weight for each model. Read from [15] and [16] for details on this formula and implementations.

This LOO part can be found on step 3 of Fig 1.

3. Distance metric with missing data

In this section, a distance measurement is introduced using as much the existing data as possible.

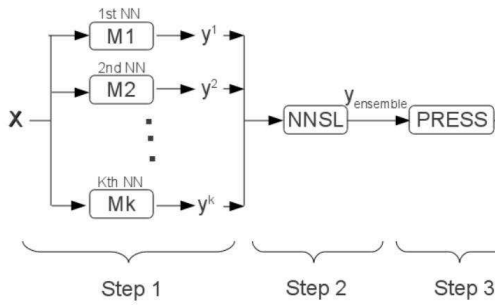

Fig. 1: Ensemble Nearest Neighbors (ENN) framewoFrk. X represents the Input data, $y^i, i = 1, ..., k$ represents the estimated output of each classifiers.

3.1 Incomplete data

We have already defined 'missing data' previously in footnote. Besides, the work in this paper assumes that the missing data is missing completely at random (MCAR) or missing at random (MAR) [20], meaning that the values of the data have no affect on whether the data is missing or not. MCAR occurs when the probability that a variable is missing is independent of the variable itself and any other external influence.

3.2 Measuring distance

Euclidean distance is normally used to measure closeness in NN series algorithms. But when confronting the incomplete data, some changes should be made to handle the missing data. Instead of making use of all the features between two observations, the adaptation of Euclidean distance is calculated by taking into account only the features with no missing values in both observations [10]. The distance is then normalized with respect to the number of features used to compute. The normalization is important to reduce the effect of the missing data. Otherwise, more features used, larger distance computed.

It may be more clear to use an example to explain. Suppose we have two observations $[2, ?, 4, 6, 8]$ and $[3, 5, 7, ?, 2]$. '?' represents a missing data. According to our new distance metric, the distances between these two observations is computed by using only the first, third, and fifth features. The second and fourth features are ignored because they contain missing values. Thus, the distance would be computed like this: $\sqrt{\frac{5}{3}((2-3)^2 + (4-7)^2 + (8-2)^2)}$. If there is no missing data in both observations, then the distance calculated is exactly the Euclidean distance in between.

4. Experiments with four datasets

In order to test the proposed method for bankruptcy prediction, four datasets are chosen in this paper. I would like to thank Dr du Jardin, Dr. Pietruszkiewicz, Dr. Atiya and Laura Kainulainen again for sharing these dataset which I know are expensive to obtain. The other reason for using these dataset is they have been used in some published articles which the results can be easily compared.

On the other hand, how to get a more general performance of the model remains to be a problematic issue. A common solution is to split the whole dataset into training, validating and testing sets, which is a good practice. In this paper, we only need to separate training and testing set because Leave-One-Out validation is used with the training set, i.e. the error we get from the trainning set is actually the LOO error. Furthermore, Monte-Carlo method is performed to split the data in order to reduce the effect of limited data size.

4.1 Monte-Carlo preprocessing

Monte-Carlo methods refer to various techniques. In this paper, Monte-Carlo methods are used to preprocessing the data, aiming to two tasks. Firstly, training set are drawn randomly about 75% of the whole data sets, the rest leaves for test set. Meantime, the proportion on the two class (healthy or bankruptcy) of both the training and testing set remain the same as the original one. Secondly, this Monte-Carlo preprocessing are repeated for many times for each dataset independently. Therefore, after these rounds of training and testing, a average test error is computed to represent the more general performance of the method.

4.2 Generating the missing date

There is no missing value originally in these four dataset. Therefore, missing data is artificially added in each datasets, in order to test the performance on incomplete data of the method. More precisely, the missing data is added one by one at randomly position till each observation has only one feature left. For example, if we have training set with N observations and d features ($N \times d$ data point totally), missing data is added till there is only N data points left (each sample has one variable). So that the model is trained and tested $N \times (d-1)$ times.

Moreover, in the following experiments, missing data is also added to the test set. The goal is to evaluate if the model trained with incomplete data can fits on the the incomplete new observations.

4.3 Pietruszkiewicz dataset

Wiesław Pietruszkiewicz has developed a data set of 240 cases of which 112 are bankrupted companies and 128 healthy. In total there are 120 companies, because the data comes from two years in a row. The possible bankruptcy occurred from two up to five years after the observations [38], [39]. The 29 variables consist of ratios of different financial variables. These variables are presented in Table 1.

Since this dataset is relatively small, Monte-Carlo cross test is used in order to present more general performance. In each round of Monte-Carlo test, the same size of samples (180 out of 240) are randomly chosen to train the model

Table 1: The variables used in the Pietruszkiewicz dataset

X1	cash/current liabilities
X2	cash/total assets
X3	current assets/current liabilities
X4	current assets/total assets
X5	working capital/total assets
X6	working capital/sales
X7	sales/inventory
X8	sales/receivables
X9	net profit/total assets
X10	net profit/current assets
X11	net profit/sales
X12	gross profit/sales
X13	net profit/liabilities
X14	net profit/equity
X15	net profit/(equity + long term liabilities)
X16	sales/receivables
X17	sales/current assets
X18	(365*receivables)/sales
X19	sales/total assets
X20	liabilities/total income
X21	current liabilities/total income
X22	receivables/liabilities
X23	net profit/sales
X24	liabilities/total assets
X25	liabilities/equity
X26	long term liabilities/equity
X27	current liabilities/equity
X28	EBIT (Earnings Before Interests and Taxes)/total assets
X29	current assets/sales

and the rest of samples are used to be test, keep the same proportion of each classes. In this experiment, 1000 times of Monte-Carlo tests are performed, and the average accuracy is calculated and shown in Fig 2.

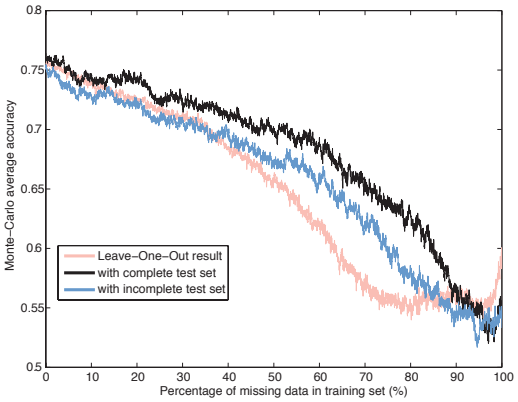

Fig. 2: Results of Pietruszkiewicz dataset

In general, this data set is very challenging to predict. It has been used and tested in many papers, for example, in Kainulainen's work [23], the best accuracy it achieved is around 75% without any missing data. Fig 2 shows the Leave-One-Out accuracy in red (the lowest curve), and test accuracies are presented in blue (curve in middle) and black (curve on the top). Incomplete test set contains one third of the missing values for each observitions. Both complete and incomplete test set performances start to decrease drastically for more than 60% missing data in the training set.

4.4 Philippe du Jardin datasets

The second and third data sets are somewhat similar. They were both used in the thesis of Philippe du Jardin. The dataset of 2002 comprises companies that have accounting data from the year 2002 and net equity data from the year 2001. The bankruptcy decisions, or more accurately, decisions of reorganization or liquidation, are from the year 2003. The dataset of 2003 was constructed similarly. In both datasets, the proportion of healthy and bankrupted companies is 50:50. In total, there were 500 and 520 samples, respectively. The companies are all from the trade sector and they have a similar structure, juridically and from the point of view of the assets. In addition, the healthy companies were still running in 2005, and had activities at least during four years. The ages of the companies were also considered, in order to obtain a good partition of companies of different ages [40]. Both of the datasets have 41 variables. The labels of the variables are presented in Table 2.

Jardin dataset and the Pietruszkiewicz dataset are fairly similar in terms of the variables. Both of them use financial ratios. The ratios are not exactly the same in all the cases, but very similar.

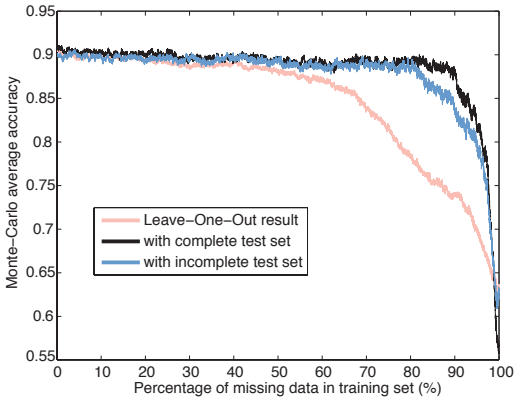

Fig. 3: Results of Philippe du Jardin datasets 2002

For the Jardin dataset, year 2003 is surely more difficult to predict than year 2002. There are some published results for this [40], [41]. Same as the previous dataset, in each Monte-Carlo round, about 75% of the observations are selected for training and the rest for testing, keep the same proportion of each classes in both training and testing set. Fig 3 shows the average classification results of 30 times Monte-Carlo processes. In general, the results remain on a high level (around 90% accuracy) and being relatively stable. More precisely, blue and black curve are interweaved together from 0% of missing data till about 80%, i.e. even the model built using only 20% of the training data, is still trustable.

Table 2: The variables used in the du Jardin datasets. EBITDA = Earnings Before Interest, Taxes, Depreciation and Amortization.

X1	Profit before Tax/Shareholders' Funds
X2	Net Income/Shareholders' Funds
X3	EBITDA/Total Assets
X4	EBITDA/Permanent Assets
X5	EBIT/Total Assets
X6	Net Income/Total Assets
X7	Value Added/Total Sales
X8	Total Sales/Shareholders' Funds
X9	EBIT/Total Sales
X10	Total Sales/Total Assets
X11	Gross Trading Profit/Total Sales
X12	Operating Cash Flow/Total Assets
X13	Operating Cash Flow/Total Sales
X14	Financial Expenses/Total Sales
X15	Labor Expenses/Total Sales
X16	Shareholders' Funds/Total Assets
X17	Total Debt/Shareholders' Funds
X18	Total Debt/Total Assets
X19	Net Operating Working Capital/Total Assets
X20	Long Term Debt/Total Assets
X21	Long Term Debt/Shareholders' Funds
X22	(Cash + Marketable Securities)/Total Assets
X23	Cash/Total Assets
X24	(Cash + Marketable Securiti es)/Total Sales
X25	Quick Ratio
X26	Cash/Current Liabilities
X27	Current Assets/Current Liabilities
X28	Quick Assets/Total Assets
X29	Current Liabilities/Total Assets
X30	Quick Assets/Total Assets
X31	EBITDA/Total Sales
X32	Financial Debt/Cash Flow
X33	Cash/Total Debt
X34	Cash/Total Sales
X35	Inventory/Total Sales
X36	Net Operating Working Capital/Total Sales
X37	Accounts Receivable/Total Sales
X38	Accounts Payable/Total Sales
X39	Current Assets/Total Sales
X40	Change in Equity Position
X41	Change in Other Debts

Moreover, there is no significant differences to predict a complete new observation or a observation with one third data missing.

Result from year 2003 is similar as year 2002. 30 times of Monte-Carlo process is done so far and shown in Fig 4. Since the size of Jardin data is larger, compared to Pietruszkiewicz dataset, it takes more time to compute in each round. More rounds of test will be done in order to further reduce the effect of randomness when adding missing data. Results will be updated later on.

4.5 Atiya dataset

The data set developed by Amir Atiya consists of 983 firms. 607 of them were solvent and 376 defaulted, but the prediction for the defaulted firms was performed at two or four instants before default. The observations of the defaulted firms come from a time period of 1 month to 36

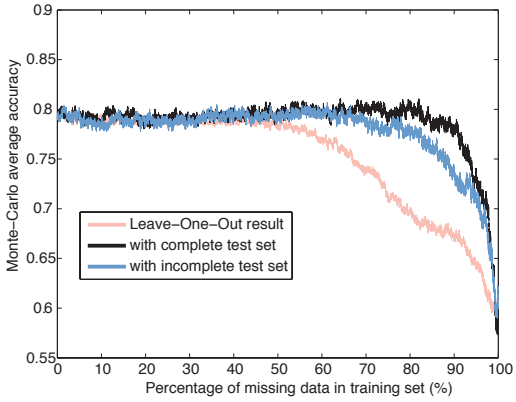

Fig. 4: Results of Philippe du Jardin datasets 2003

months before the bankruptcy, the median time being 13 months [37]. In total, there were 63 variables. The data was standardized to 0 mean and variance 1 before performing classification task. The values of the Atiya dataset are presented in Tables 3 and 4

Since the Atiya dataset is unbalanced with regards to the number of healthy companies and number of bankrupted companies, a different measure for mean accuracy is used. That measure is defined in Equation 3.

$$\frac{\frac{True\ positive}{Total\ positive} + \frac{True\ negative}{Total\ negative}}{2} \tag{3}$$

This Atiya dataset (983 observations and 63 variables) is relatively larger than previous three datasets. Thus, after each round of Monte-Carlo split, there are 737 samples (about one third) using for training. Missing data is added from 1 to 45694 $(737 \times (63-1))$ to the training set, i.e., the model have to be trained and tested 45694 times for each Monte-Carlo round. It is very time consuming. Therefore, 3 rounds is done so far, more rounds of experiments are still running and more results will be updated.

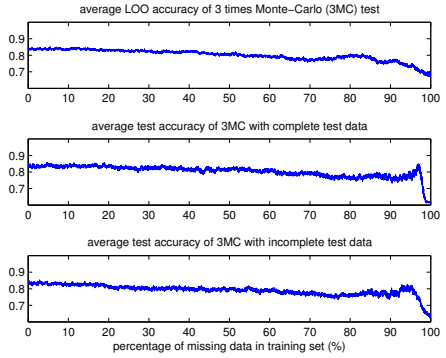

Fig. 5: Results of Atiya dataset

Fig 5 illustrates the results using three separate figures

Table 3: The variables used in the Atiya dataset, part 1. ROC=rate of change (usually over 4 year period), CFPS=cashflow per share, EPS=earning per share, GOI=gross operating income (i.e.before taxes, interest and other deductions), profit mgn=profit margin, TA=total assets, gross profit mgn=profit margin as related to GOI, EQ=shareholders equity (also called book value), NOI=net operating income (after taxes, etc), P/CF=price cashflow ratio, PE = price earnings ratio.

X1	cash/tot assets
X2	working capital/tot assets (TA)
X3	working capital/curr assets
X4	equity (EQ)/TA
X5	1-(long term debt/TA)
X6	rate of chg of cash flow per share (CFPS)
X7	rate of chg (ROC) of earnings per share (EPS)
X8	ROC(EPS from cont. operations)
X9	ROC(gross operating income GOI)
X10	ROC(net oper. Inc NOI)
X11	ROC(sales)
X12	ROC(gross profit margin)
X13	ROC(net profit margin)
X14	a measure of share price chg
X15	a measure of chg of gross oper mgn
X16	one year chg in net profit mgn
X17	ROC(TA)
X18	one year chg in EQ
X19	other ROC(CFPS) (other measure of chg)
X20	other ROC(EPS)
X21	other ROC(EPS cont oper)
X22	other ROC(GOI)
X23	other ROC(NOI)
X24	other ROC(sales)
X25	gross profit mgn
X26	net profit mgn
X27	a measure of dividend incr/decr
X28	cash flow (CF)/TA
X29	earnings/TA
X30	earnings cont oper/TA
X31	GOI/TA
X32	NOI/TA
X33	sales/TA
X34	PE ratio
X35	P/CF ratio
X36	price sales ratio
X37	price book value ratio
X38	return on assets ROA
X39	return on equity
X40	current ratio

(one curve each). The reason is because these three curves are interweaved together which is impossible to see clearly in White-Black print. The curve is not as smooth as previous ones because only three Monte-Carlo are used. However, the tendency is similar as previous results. The models built with up to at least 50% of missing data keep stable at a high level, and test with complete data or incomplete data (one third missing) doesn't make obvious differences.

Table 4: The variables used in the Atiya dataset, part 2. ROC=rate of change (usually over 4 year period), CFPS=cashflow per share, EPS=earning per share, GOI=gross operating income (i.e.before taxes, interest and other deductions), profit mgn=profit margin, TA=total assets, gross profit mgn=profit margin as related to GOI, EQ=shareholders equity (also called book value), NOI=net operating income (after taxes, etc), P/CF=price cashflow ratio.

X41	Quick ratio
X42	market capitalization/(long term debt LTD)
X43	relative strength indicator
X44	gross profit mgn
X45	net profit mgn
X46	one-year rel chg of CF
X47	one-year rel chg of GOI
X48	one-year rel chg og NOI
X49	4 yr ROC(CF)
X50	4 yr ROC(GOI)
X51	4 yr ROC(NOI)
X52	3 yr ROC(CF)
X53	3 yr ROC(GOI)
X54	3 yr ROC(NOI)
X55	TA
X56	sector default prob
X57	one year ROC(price)
X58	4 yr ROC(price)
X59	3 yr ROC(price)
X60	price
X61	a measure of ROC(price)
X62	volatility
X63	3 yr ROC(EQ)

5. Conclusions

In this paper, a new methodology to achieve classification for bankruptcy prediction with incomplete data is introduced. The approach ENN assembles k different Nearest Neighbor classifiers, and makes a linear combination of them. The most significant advantage is that ENN uses a modified Eulidean distance metric to solve the missing data problem while keeping the comparable performance.

In the experiments, Monte Carlo test is used in order to reduce variability of the performances casued by limited data size. Results on the four financial datasets illustrate that the performances of the proposed methodology are not deteriorating significantly with missing data from a percentage going from 0 to at least 50% of missing data in both the training and the testing data. The test results remain on the same level with both complete testing observations and incomplete testing ones (one third of the data are missing for each observations).

The results confirm the advantages of this method: being robust while providing good performance with missing data and a comparatively simple model.

References

[1] D. Opitz, R. Maclin, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[2] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, pp. 21–45, 2006.

[3] L. Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review*, vol. 33, 1–39,2010.

[4] W.H. Beaver, " Financial ratios as predictors of failure," *Journal of Accounting Research*, vol. 4, pp. 71–111, 1966.

[5] E.I. Altman, " Financial ratios, discriminant analysisi and the prediction of corporation bankruptcy," *The Journal of Finance*, vol. 23, pp. 589–609, 1968.

[6] H.Wackernagel, "Multivariate Geostatistics - An introduction with applications," Springer, Berlin, 1995.

[7] L.S. Gandin, "Objective analysis of meteorological fields," *Israel Program for Scientific Translations*, Jerusalem, pp. 242, 1969.

[8] G.E. Batista, M.C. Monard, "A study of k-nearest neighbour as an imputation method, *Second International Conference on Hybrid Intelligent Systems*, vol. 87, pp 251–260, Santiago, Chile, 2002.

[9] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and Systems Magazine*, vol. 6, pp. 21–45, 2006.

[10] G. Doquire, M. Verleysen, "Mutual information for feature selection with missing data," *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges (Belgium), 27–29 April 2011, to appear.

[11] L. Kuncheva and C. Whitaker, "Measures of diversity in classifier ensembles," *Machine Learning*, vol. 51, pp. 181–207, 2003.

[12] P. Sollich and A. Krogh, "Learning with ensembles: How overfitting can be useful," *Advances in Neural Information Processing Systems*, vol. 8, pp. 190–196, 1996.

[13] G. Brown, J. Wyatt, R. Harris and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion*, vol. 61, pp. 5–20, 2005.

[14] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, 21–27, 1967.

[15] R. H. Myers, "Classical and Modern Regression with Applications," Duxbury, Pacific Grove, CA, 1990. USA

[16] G. Bontempi, M. Birattari, and H. Bersini, "Recursive lazy learning for modeling and control," *Euro- pean Conference on Machine learning*, pp. 292–303, 1998.

[17] T. M. Mitchell, *Machine Learning*, pp. 175, 1997.

[18] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[19] Y. Miche, E. Eirola, P. Bas, O. Simula, C. Jutten, A. Lendasse and M. Verleysen, "Ensemble modeling with a constrained linear system of LOO outputs," *ESANN2010: 18th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine learning*, bruges, Belgium, pp. 19–24, Apr 2010.

[20] R. J. A. Little and D. B. Rubin, "Statistical Analysis with Missing Data (second ed.)," Wiley, NJ, USA, 2002.

[21] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[22] http://www.pietruszkiewicz.com/datasets

[23] L. Kainulainen, Q. Yu, Y. Miche, E. Eirola, E. Séverin and A. Lendasse, " Ensembles of Locally Linear Models: Application to Bankruptcy Prediction," *In Proceedings of the 2010 International Conference on Data Mining*, pp. 280–286, July, 2010.

[24] H. Frydman, E.I. Altman and D. Kao, "Introducing recursive partitioning for financial classification: The case of financial distress, Journal of Finance," vol 40 , pp. 269–291, 1985.

[25] M.L. Marais, J. Patel and M. Wolfson, "The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications," *Journal of Accounting Research*, vol. 22, pp. 87–114, 1984.

[26] J. Zimmermann, "Fuzzy set theory and its applications," *Kluwer Academic Publishers*, London, 1996.

[27] S.M. Bryant, "A case-based reasoning approach to bankruptcy prediction modeling," *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 6, pp. 195–214, 1997.

[28] Park and I. Han, "A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction," *Expert Systems with Applications*, vol. 23, pp. 255–264, 2002.

[29] K.S. Shin and Y.J. Lee, "A genetic algorithm application in bankruptcy prediction modeling," *Expert Systems with Applications*, vol. 23, pp. 321–322, 2002.

[30] J.H. Min and Y.C. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters", *Expert Systems with Applications* vol. 28, pp. 603–614, 2005.

[31] M. Lam, "Neural networks techniques for financial performance prediction: Integrating fundamental and technical analysis," *Decision Support Systems*, vol. 34, pp. 567–581, 2004.

[32] A.F. Atiya, "Bankruptcy prediction for credit risk using neural networks: A survey and new results," *IEEE Transactions on Neural Networks*, vol. 12, pp. 929–935, 2001.

[33] Z.R. Yang, M.B. Platt and H.D. Platt, "Probability neural network in bankruptcy prediction," *Journal of Business Research*, vol. 44, pp. 67–74, 1999.

[34] P. Swicegood and J.A. Clark, "Off-site monitoring systems for predicting bank underperformance: A comparison of neural networks, discriminant analysis and professional human judgment," *International Journal of Intelligent Systems in Accounting, Finance and Management*, vol. 10, pp. 169–186, 2001.

[35] S. Kaski, J. Sinkkonen and J. Peltonen, "Bankruptcy analysis with self-organizing maps in learning metrics," *IEEE Transaction on Neural Networks*, vol. 12, pp. 936–947, 2001.

[36] K. Lee, D. Booth and P. Alam, "A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms," *Expert Systems with Applications*, vol. 29, pp. 1–16, 2005.

[37] A. F. Atiya, "Bankruptcy prediction for credit risk using neural networks: A survey and new results," *Neural Networks, IEEE Transactions*, vol. 12, pp. 929–935, Jul, 2001.

[38] W. Pietruszkiewicz, "Application of Discrete Predicting Structures in an Early Warning Expert System for Financial Distress," Ph.D. thesis, Fac- ulty of Computer Science and Information Technology, Szczecin University of Technology, Dec 2004.

[39] W. Pietruszkiewicz, "Dynamical systems and nonlinear kalman ïňĄlter- ing applied in classiïňĄcation," *In Proceedings of 2008 7th IEEE International Conference on Cybernetic Intelligent Systems*, pp. 263–268, 2008.

[40] P. D. Jardin, "Prévision de la défaillance et réseaux de neurones: léapport des méthodes numériques de sélection de variables," PhD thesis, Université de Nice-Sophia-Antipolis, 2007.

[41] D. Sovilj, A. Sorjamaa, Q. Yu, Y. Miche and E. Séverin, "OPELM and OPKNN in long-term prediction of time series using projected input data," *Neurocomputing*, vol. 73, pp. 1976–1986, June, 2010.