DE GRUYTER OPEN

International Conference KNOWLEDGE-BASED ORGANIZATION
Vol. XXI No 2 2015

A CLUSTERING OF LISTED COMPANIES CONSIDERING CORPORATE GOVERNANCE AND FINANCIAL VARIABLES

Darie MOLDOVAN, Mircea MOCA

"Babeş-Bolyai" University, Cluj-Napoca, Romania
darie.moldovan@econ.ubbcluj.ro mircea.moca@econ.ubbcluj.ro

***Abstract**: The corporate governance quality has always been a decision criterion for investments, many recent studies trying to define metrics in order to help investors in their decision process. In this paper we investigate whether the clustering of companies' information concerning their corporate governance politics and financial information could be mapped with the help of clustering. Our approach is to build clusters using machine learning techniques, based on corporate governance and financial variables from a number of 1400 listed companies. We evaluate the obtained clusters by matching them with the classes of two well-known indicators (Tobin's Q and Altman Z-score), used to estimate the companies' performance. We obtain partial matches of the benchmark variables and we compare the performances of the used algorithms.*

Keywords: corporate governance, clustering, Tobin's Q, Altman Z-Score, machine learning

1. Introduction

In the decision process of investing in a listed company one can include a myriad of factors, beginning with company's financial statements, technical analysis for its price evolution or risk assessment.

In order to facilitate the decision, synthetic indicators are considered, such as Tobin's Q ratio [1], for performance evaluation, or Altman Z-score [2], an indicator for the bankruptcy risk.

In this work we are focused on using the above mentioned indicators as a benchmark for clustering companies from three different wide stock indexes, according to their financial information and corporate governance variables. The goal is to verify if the clusters obtained map both, one or none of the widely used indicators. In this way we can observe if the selected variables explain the two indicators and also the affinities between the corporate governance and financial variables.

The three stock indexes chosen are S&P500, containing 500 companies from the United States of America, STOXX Europe 600, from Western Europe and STOXX Eastern Europe 300.

We used several clustering algorithms in order to obtain stable clusters after multiple iterations and validate our approach.

The remainder of the paper is structured as follows. In section 2 we give related work, offering insights on the utility of the two indicators for evaluating the companies' performance and a literature review on the clustering techniques used in finance. Section 3 describes the employed clustering methodology, section 4 presents the experiments and obtained results and section 5 concludes the paper.

2. Related work

The Tobin's Q ratio was introduced by Tobin [1] to compare the physical assets value of a company with its replacement value, including in this way the intangible value of the company, such as the intellectual capital.

DOI: 10.1515/kbo-2015-0056

Altman Z-Score [2] is an estimator for the bankruptcy risk of a company in the next two years, being widely proposed in literature [3,4,5] as a reliable solution.
The relation between the corporate governance variables and the companies' performances has been studied in different ways in order to identify patterns concerning the size of the board, the presence of women on board, the proportion of independent directors or the leadership style. These patterns can lead to a classification of the companies, otherwise difficult to observe [6].
Clustering methods are used in the financial area, as collected data volumes and the need to obtain accurate analysis on the natural structure of it are increasing.
Classic clustering methods such as k-means [7], Expectation Maximization [8], but also density based methods like DBSCAN [9] and OPTICS [10] are nowadays used in practice. Also, methods for data stream clustering are used, especially created for rapidly changing environments (e.g. market data).
The related literature concerning clustering in finance includes the works of Aitken [11], who studied the relation between stock price increase and company evaluation. Lux [12] tried to explain the effects of volatility on the market efficiency, by creating volatility clusters. Kumar [13] was interested in creating volatility clusters on the indifference pricing of options. The effects of political instability on price clustering were studied by Narayan [14] on the Asian markets.
Financial times series clustering was performed by Bastos [15], introducing a new distance measure, based on variance tests. In a paper by Cameron [16] another variance estimator is proposed for logit, probit or GMM methods in order to obtain robust clusters.
Aghabozorgi [17] proposed a three step clustering method for companies categorization based on their market similarities, using sub-clusters, in order to increase the efficiency and effectiveness. Along clustering, classification and neural networks were employed by Enke [18] with the aim of obtaining accurate results in forecasting the stock market prices.
A review of the clustering methods used in finance was proposed by Cai [19], showing practical examples to emphasize the differences between these methods from different angles.
When it comes to classes via clustering, several works contribute by proposing different approaches. For instance, Vilalta [20] proposes a method to improve the estimates of Naive Bayes, Lopez [21] uses the technique for a prediction problem, while Verma [22] is using an ensemble classifier based on clustering methods, aiming to facilitate efficient learning.

3. Methodology

In order to achieve our objective for clustering the companies' information from the three market indexes, we used the Weka software.
Regarding the data mining framework, we considered the life cycle of knowledge discovery according to the KDD model (Knowledge Discovery in Databases), consisting of several phases. Next, we describe each phase with its specific tasks.

3.1. Data selection

In our study we used data from Bloomberg, containing 50 variables from both corporate governance area and financial information. In order to cover three large areas we chose three representative stock indexes: S&P500 (SPX - 500 listed companies from USA), STOXX Europe 600 (SXXP - 600 companies from the Euro zone) and STOXX Eastern Europe 300 (EEBP - 300 companies from Eastern Europe). Data were collected in November 2014, being divided into three sub-sets, mapping the three indexes.

3.2. Pre-processing phase

In order to prepare the data for the data mining phase we need to perform a cleaning task. In order to obtain robust results from clustering, the outliers must be

eliminated for each involved variable.

In the same time, clustering is also sensitive to missing data. Several methods are available to tackle this issue, including the removal of records containing missing information, creating a separate class for the missing data or replacing it with the mean value of the variable. For our research, in order to preserve the available information in a certain record, we replaced the missing data using the average.

3.3. Data Mining phase

For the Data Mining phase we adopted three algorithms in order to perform the clusters learning.

We used an implementation of k- means algorithm [7], one of the most popular algorithms for clustering. It partitions the data into k clusters, starting from centroids, and computing the fitness with the help of the Euclidian distance.

The second algorithm employed is the Expectation Maximization (EM) [8], another well known method. It functions by maximizing the log likelihood of a function by multiple iterations.

The last algorithm used is an artificial intelligence method for reducing the highly dimensional data, named Self Organizing Maps (SOM) [22]. It reduces the vectors by creating a network to store the collected information and makes sure to maintain the learned relations.

3.4. Results evaluation

We evaluate the results obtained from two different perspectives. One is the quality of the clusters obtained, in terms of homogeneity inside the cluster and cluster stability between several iterations. For measuring the distance between the components of the cluster, the Euclidian Distance was used, being largely adopted in literature.

The second perspective is related to the class mapping and concerns a business evaluation. A good mapping of the two benchmark indicators (Tobin's Q and Altman Z-Score) would show a strong connection between the corporate governance variable, financial information and the companies' performance.

4. Experiments and results

Our experiments consist of building clusters by using three different methods and map them on the two benchmark indicators.

In order to prepare the data for learning, we pre-processed it by eliminating the outliers from each of the datasets. In this way, the three datasets were resized, as some of the instances were removed. After pre-processing, the datasets have the following dimensions: SPX is 419 instances, SXXP is 470 instances and EEBP is 293 instances.

As clustering is sensitive to missing values, we replaced the missing values in the three datasets with the mean value of the variable, avoiding in this way to reduce the data even more. Once the data was prepared for learning, we can proceed to apply the algorithms. Applying the algorithms require specific settings. One of the key parameters in algorithms tuning is the number of clusters. We set the minimum number of clusters to 2 for mapping Tobin's Q ratio and 3 for Altman Z-score, corresponding to the number of classes of each variable.

We begin by training the SOM algorithm, which computes the best number of clusters for each given dataset. In this way we obtain another benchmark for the other two algorithms used (for which we need to manually select the desired number of clusters). We run a total of 36 scenarios.

The EM and k-means algorithms are applied considering different numbers of clusters, between the minimum value already set and the value offered by the SOM algorithm.

In Tables 1, 2 and 3 we present the clustering results, for each dataset, showing the algorithm, the number of clusters used and the mapping ratio for each class in the two benchmark variables. When the resulting clusters are more than 2 for Tobin's Q and three for Altman Z-score, the mapping refers only to the clusters that resemble most to the classes, ignoring the other clusters.

Table no. 1: Clustering results for the SPX dataset

Algorithm	*Number of clusters*	*Tobin's Q mapping*				
		0	*1*	*0*	*1*	*2*
k- means	2	51%	52.5%	N/A	N/A	N/A
	3	52.5%	52%	19%	15%	65%
	4	55%	53%	19%	19%	64%
EM	2	50%	51%	N/A	N/A	N/A
	3	51%	51%	20%	30%	67%
	4	52%	57%	15%	17%	61%
SOM	2	N/A	N/A	N/A	N/A	N/A
	3	N/A	N/A	N/A	N/A	N/A
	4	67%	55%	28%	27%	67%

Table no. 2: Clustering results for the SXXP dataset

Algorithm	*Number of clusters*	*Tobin's Q mapping*				
		0	*1*	*0*	*1*	*2*
k- means	2	51%	42%	N/A	N/A	N/A
	3	57%	60.4%	25.6%	29%	64%
	4	57.5%	60.2%	25.9%	28.1%	52%
EM	2	59%	66.5%	N/A	N/A	N/A
	3	70.5%	51%	20%	34%	61%
	4	73%	51%	48%	24%	55%
SOM	2	N/A	N/A	N/A	N/A	N/A
	3	N/A	N/A	N/A	N/A	N/A
	4	59%	62%	29%	32%	66%

Table no. 3: Clustering results for the EEBP dataset

Algorithm	*Number of clusters*	*Tobin's Q mapping*				
		0	*1*	*0*	*1*	*2*
k- means	2	51.7%	50.5%	N/A	N/A	N/A
	3	51.1%	50.5%	8%	51%	64.5%
	4	51.2%	53.1%	8.5%	53%	64.6%
EM	2	50.7%	50.3%	N/A	N/A	N/A
	3	50.7%	51%	32.5%	52.3%	62.4%
	4	47.3%	54%	11%	57%	62.9%
SOM	2	N/A	N/A	N/A	N/A	N/A
	3	N/A	N/A	N/A	N/A	N/A
	4	55%	51%	11%	50%	61%

The obtained results show a partial match between the obtained clusters and the benchmark classes. We can observe a tendency for a better match when increasing the number of clusters, due to the fact that the other clusters are absorbing the noise, in

this way the creating the opportunity for a more precise differentiation between the members of each cluster.

The Self Organizing Map algorithm has better performed in general, not only concerning the match between its clusters and the benchmark classes, but also in terms of fitness. The optimal number of clusters chosen was 4. For every considered scenario, this value suggests that classifying the companies only in two or three categories, according to Tobin's Q ratio or Altman Z-score is not sufficient in order to identify the differences between the companies.

The best accuracies in mapping were obtained for the SPX dataset when using the SOM algorithm. There is a clear difference between the mapping of Tobin's Q ratio and Altman Z-score. In this sense, the results show the first as being easier to map by the corporate governance and financial variables. The high risk of bankruptcy, which is represented by class 0 of the Altman Z-score, was poorly mapped by the clustering algorithms, emphasizing the difficulty of this task. This can also be explained by the imbalances in the dataset, the examples with high risk being in a much smaller proportion than those in the safe category.

5. Conclusions

In this paper we investigated whether the clustering of companies' information concerning their corporate governance politics and financial information could be mapped with the help of clustering on the categories of two well known performance indicators: Tobin's Q ratio and Altman Z-score.

The results show a partial map of the obtained clusters on the classes of the two benchmark variables, suggesting the need of a more precise differentiation of the companies considering the available information.

As future developments, we will consider building a more precise model based on the corporate governance and financial data, in order to show more differences between the companies' performances than obtained with the two indicators.

Acknowledgements

This work was co financed from the European Social Fund through Sectorial Operational Program Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/134197 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain".

References

[1] Tobin, J: A General Equilibrium Approach To Monetary Theory. Journal of Money, Credit and Banking (1) pp.15–29 (1969)

[2] Altman, E. I., Saunders, A.: Credit risk measurement: Developments over the last 20 years. Journal of banking & finance, 21(11), pp. 1721-1742 (1997).

[3] Weimin Chen, Guocheng Xiang, Youjin Liu, Kexi Wang, Credit risk Evaluation by hybrid data mining technique, Systems Engineering Procedia, 3 (2012)

[4] Kambal, E.; Osman, I. ; Taha, M. ; Mohammed, N. ; Mohammed, S.Credit scoring using data mining techniques, Computing, Electrical and Electronics Engineering (ICCEEE), IEEE (2013)

[5] Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos. "Data mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32(4) pp.995-1003 (2007)

[6] Moldovan, D., and Mutu, S., Learning the Relationship between Corporate Governance and Company Performance using Data Mining, Proceedings of the 11th International Conference on Machine Learning and Data Mining (MLDM'15), Hamburg, Germany, July 2015, In press.

[7] Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding."

Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2007.

[8] Moon, Todd K. "The expectation-maximization algorithm." *Signal processing magazine, IEEE* 13.6 (1996): 47-60.

[9] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

[10] Ankerst, Mihael, et al. "OPTICS: ordering points to identify the clustering structure." *ACM Sigmod Record*. Vol. 28. No. 2. ACM, 1999.

[11] Aitken, Michael, et al. "Price clustering on the Australian stock exchange." *Pacific-Basin Finance Journal* 4.2 (1996): 297-314.

[12] Lux, Thomas, and Michele Marchesi. "Volatility clustering in financial markets: a microsimulation of interacting agents." *International Journal of Theoretical and Applied Finance* 3.04 (2000): 675-702.

[13] Kumar, Rohini. "Risk indifference price of options under fast mean-reverting stochastic volatility." *Conference on Stochastic Asymptotics and Applications*. 2014.

[14] Narayan, Paresh Kumar, and Russell Smyth. "Has political instability contributed to price clustering on Fiji's stock market?." *Journal of Asian Economics* 28 (2013): 125-130.

[15] Bastos, João A., and Jorge Caiado. "Clustering financial time series with variance ratio statistics." *Quantitative Finance* 14.12 (2014): 2121-2133.

[16] Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. "Robust inference with multiway clustering." *Journal of Business & Economic Statistics* 29.2 (2011).

[17] Aghabozorgi, Saeed, and Ying Wah Teh. "Stock market co-movement assessment using a three-phase clustering method." *Expert Systems with Applications* 41.4 (2014): 1301-1314.

[18] Enke, David, and Suraphan Thawornwong. "The use of data mining and neural networks for forecasting stock market returns." *Expert Systems with applications* 29.4 (2005): 927-940.

[19] Cai, Fan, Nhien-An Le-Khac, and M-Tahar Kechadi. "Clustering approaches for financial data analysis: a survey." *Proceedings of the 8th International Conference on Data Mining,(DM'12), Las Vegas, Nevada, USA*. 2012.

[20] Vilalta, Ricardo, and Irina Rish. "A decomposition of classes via clustering to explain and improve naive Bayes." *Machine Learning: ECML 2003*. Springer Berlin Heidelberg, 2003. 444-455.

[21] Lopez, Manuel Ignacio, et al. "Classification via Clustering for Predicting Final Marks Based on Student Participation in Forums." *International Educational Data Mining Society* (2012).

[22] Kohonen, Teuvo. "The self-organizing map." *Proceedings of the IEEE* 78.9 (1990): 1464-1480.