ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Knowledge Discovery in Quarterly Financial Data of Stocks Based on the Prime Standard using a Hybrid of a Swarm with SOM

Michael C. Thrun[1]

1- University of Marburg, Mathematics and Computer Science
Hans-Meerwein Str., 35032 Marburg, Germany
thrun@deepbionics.de

Abstract. Stocks of the German Prime standard have to publish financial reports every three months which were not used fully for fundamental analysis so far. Through web scrapping, an up-to-date high-dimensional dataset of 45 features of 269 companies was extracted, but finding meaningful cluster structures in a high-dimensional dataset with a low number of cases is still a challenge in data science. A hybrid of a swarm with a SOM called Databionic swarm (DBS) found meaningful structures in the financial reports. Using the Chord distance the DBS algorithm results in a topographic map of high-dimensional structures and a clustering. Knowledge from the clustering is acquired using CART. The cluster structures can be explained by simple rules that allow predicting which future stock courses will fall with a 70% probability.

1 Introduction

Some of the companies listed in the Frankfurt stock exchange comply with a rigorously defined transparency standard higher than the general standard of stocks traded on the world market. This market segment of stocks is called German Prime standard and requires companies to publish their accounting information [1]. Currently, the Prime standard contains 324 companies [2]. Usually, fundamental analysis has the purpose of predicting the market value of a stock. Without applying a forecasting method, the "fundamental" value of a stock is determined [3] and compared to the current value. Some empirical studies have used accounting information to predict the future performance of a firm [3]. Often fundamental analysis is a part of a larger system with the goal to select the right stock (e.g., CANSLIM system [4]) using a low amount of variables and a larger number of stocks [5-7]. Thus, "research does not fully exploit the wealth of information contained in general purpose financial reports but is outside of the primary financial statements" [8]. In this work, high-dimensional structures are investigated which are characterized by 45 variables describing the quarterly financial statement, income sheet, and cash flow defined by the German prime standard. The information is extracted directly from the web [9] by applying a self-made web scrapping algorithm allowing the extraction of data for 269 companies. The dataset was extracted for the first quarter of 2018, and the stock courses were extracted for the second quarter of 2019.

2 Methods

After preprocessing (i.e., handling of missing values, normalization, and decorrelation) of data the Databionic swarm (DBS) algorithm can be applied [10] on

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

269 companies where 43 variables are used. The algorithm consists of three parts. First, the high-dimensional data is projected into a two-dimensional space using a swarm which utilizes game theory, self-organization, and emergence as well as swarm intelligence [10]. Besides the number of clusters and a single Boolean parameter describing the type of structure, the DBS does not require any parameters to be set. Here, the Chord distance (c.f. [11] p.49) is chosen for cluster analysis because distributions analysis of distances indicates a bimodality (Fig. 1). Statistical testing with Hartigan's dip test [12] agrees that the distribution is not unimodal ($p(N=36046,D= 0.027821)<2.2e-16$). Bimodality in the distance distributions serves as an indicator that there are smaller intra-cluster distances and larger inter-cluster distances resulting in the assumption that a cluster structure exists. Note, that in general, all financial reports are quite dissimilar because there are no small distances below 0.5 in Fig.1.

In the second part, the visualization of a topographic map is generated using a simplified emergent self-organizing map [13] in order to visualize high-dimensional structures. The topographic map accounts for projection errors because two-dimensional similarities in the scatter plot cannot coercively represent high-dimensional distances [14] and common quality measures of dimensionality reduction methods require prior assumptions about the underlying structures [15]. The topographic map combines a 3D landscape with hypsometric tints. Hypsometric tints are surface colors which depict ranges of elevation [16] intersected with contour lines. "Blue colors indicate small distances (sea level), green and brown colors indicate middle distances (small hilly country) and white colors indicate high distances (snow and ice of high mountains). The valleys and basins indicate clusters and the watersheds of hills and mountains indicate borderlines of clusters" [17]. In sum, this visualization is consistent with a 3D landscape for the human eye enabling 3D printing of a representation of high-dimensional structures; therefore one has a haptic grasp and sees data structures intuitively enabling layman to interpret them [17].

In [18] it was shown that the visualized elevation between two projected points is an approximation of the input-space distance *D(l,j)* between the two high-dimensional points (here two companies). Voronoi cells around each projected point define the abstract U-matrix (AU-matrix) and generate a Delaunay graph $\mathcal{D}$. For every projected point all direct connections are weighted using the input-space distances *D(l,j)* because each border between two Voronoi cells defines a height. All possible weighted Delaunay paths $p_{l,j}$ between all points are calculated. Then, the minimum of all possible path distances between a pair of points $\{l,j\} \in O$ in the output space O is calculated as the shortest path $G(l,j,\mathcal{D})$ using the algorithm of [19] resulting in a new high-dimensional distance $D^*(l,j)$ which defines the distance of each two companies based on their financial accounting. In this case, the connected structure type of DBS clustering is chosen where the similarity between two subsets of data points is defined as the minimum distance between data points in these subsets and the clustering process is agglomerative (c.f. [10]): Let $\widetilde{D}$ be the distance between two clusters $c_1 \subset I$ and $c_2 \subset I$, and let *D(l,j)* be the distance between two data points in the input space I; then the connected approach is defined with $\tilde{D}(c_1,c_2) = \min_{l \in c_1, j \in c_2} D(l,j)$.

Through inspecting the topographic map the central problem of estimation of the number of clusters is solved by counting the number of valleys and the structure type

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

can be set. A dendrogram can be shown additionally. The clustering is valid if mountains do not partition clusters indicated by colored points of the same color [10].

3 Results

The high-dimensional structures of the financial reports of companies are visible in the topographic map in Fig .3 where three valleys can be identified. The number of valleys seen in the topographic map lead to the choice of three clusters. Additionally, large changes in fusion levels of the ultrametric portion of the Chord distance indicate the best cut (Fig.2, left, y-axis). After the clustering process, the heatmap in Fig. 2 indicates that more similar points are inside a cluster (yellow) and more dissimilar points are outside a cluster (red). The three main clusters computed by the connected approach of DBS are separated by mountain ranges (Fig. 3) and have an average intra-cluster distance of c_1=0.97, c_2=1.01 and c_3=0.91 being located in the first mode of Fig. 1. The points in the topographic map symbolize the companies and are colored by the clustering. Two outliers can be identified. The heatmap agrees with the topographic map that the clusters consist of more similar points inside a cluster than outside the cluster. Applying the CART algorithm [20] to the clustering yields a simple set of rules (Fig. 5). If "net income from continuing ops" NIFCO < -264 and "operating income or loss" OIL < -58.3 then class 2 is defined. NIFCO defines the after-tax earnings that business has generated from its operational activities. NIFCO "is considered to be a prime indicator of the financial health of a firm's core activities" [21]. OIL is the difference between revenues and costs generated by ordinary operations and before deducting interest, taxes et cetera [22]. The extracted rules for class 2 lead to the hypothesis that stock prices of companies in the second cluster are overvalued and will fall in the next quarter. Using stock prices of Q2/2018 the hypothesis is verified under the assumption that the data of all companies is available on the same date at the beginning of the respective quarter of a year. In Fig. 4 the class-dependent MD-plot [28] of the rate of return of stocks is shown. The courses were compared at the last trading day of the first quarter against the last trading day of the second quarter in 2018 with relative differences [24]. The rate of return in Class 2 is significantly lesser than in Class 1: Class 1 has a shift equal to zero with a Wilcoxon rank sum test (shift not equal to zero p(N=196, V=9512)=0.86), class 2 a shift not equal to zero with a Wilcoxon rank sum test p(N=54,W=6792) <0.001. Stock courses of eight companies (3% of class 1 and 2 and outliers) could not be extracted from the web.

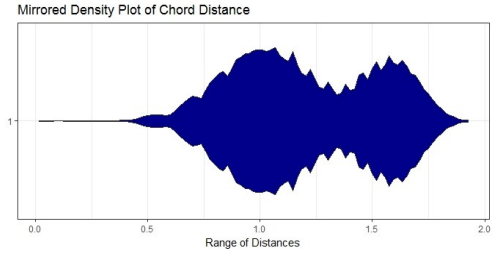

Fig. 1: MirroredDensity plot (MD-plot [28]) of the R package 'DataVisualizations' on CRAN [23] shows a bimodal distribution of Chord distances with the first mode around 1 and the second mode around 1.6.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

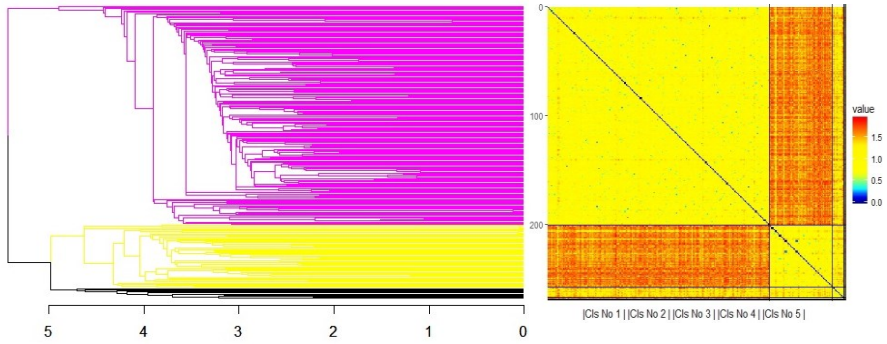

Fig. 2: Dendrogram (left), and heatmap (right) of the distances sorted by the clustering using the R package 'DataVisualizations' on CRAN [23]. In the heatmap, the smaller distances are in yellow belonging to a clusters and the larger distances in red in-between different clusters (c.f. Fig. 1). The branches of the dendrogram are colored by the first three clusters.

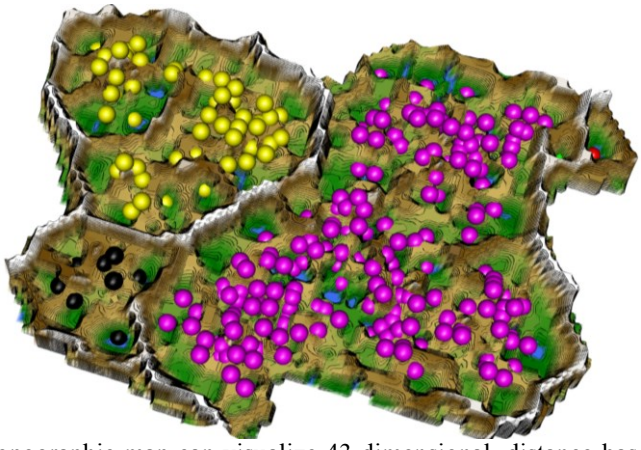

Fig. 3: The topographic map can visualize 43 dimensional, distance-based structures. It shows three valleys - one major cluster with companies represented by magenta points (N=199), one smaller cluster with companies in yellow (N=57) and black (N=10) and two outliers (green and red).

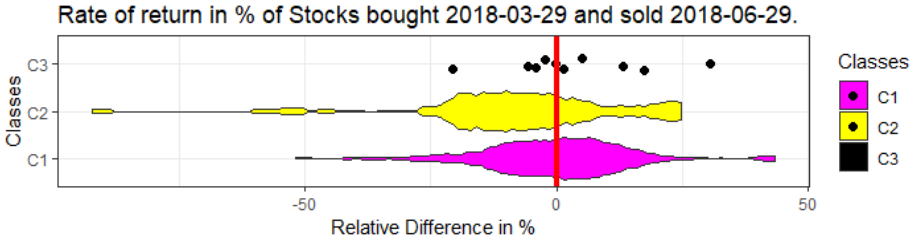

Fig. 4: MDplot [28] of the rate of return in % calculated with relative differences [24]. The red line marks a rate of return of zero. Class 2 has 30% of stocks with a rate of return above zero and significantly differs from Class 1.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

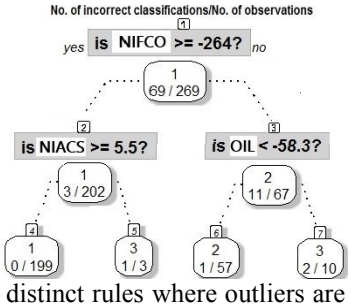

Fig. 5: CART shows three distinct rules where outliers are incorrectly classified: "net income from continuing ops" (NIFCO>=-264), "operating income or loss" (OIL<-58.3) and "net income applicable to common shares" (NIACS>= 5.5).

4 Discussion

The topographic map visualizes a 43 dimensional, distance-based structures in a 3D landscape in Fig. 3. It showed three distinctive valleys leading to the hypothesis that the dataset has three clusters. The topographic map in Fig. 3 is noisy because the sample of data is small but the feature space is large, and there are no small distances (see Fig. 1). However, the distributions analysis (Fig. 1) and the heatmap (Fig. 2) indicate high-dimensional and distance-based structures of the data which can be visualized by the topographic map (Fig. 3). These structures are reproduced in the cluster analysis and a dendrogram. The heatmap could also indicate that cluster three is a sub cluster of cluster one, yet the clustering is valid because mountains do not partition clusters and the intra-cluster distances are smaller than the inter-cluster distances (Fig. 2). The primary cluster (N=199) does not yield interesting insights about the data which is typical for cluster analysis [25]. However, the extracted rules (Fig. 5) are interesting for the second cluster (N=57) where NIFCO has to be lower than -264 and (OIL) lower than -58.3. By understanding these two variables through the two rules, it can be concluded that the stock prices of these companies will fall in the next quarter. This prediction can be verified by the significant higher probability of a negative rate of return of stock prices of companies in class 2 in Q2/2018 compared to class 1. In sum, 7 out of 10 companies in class 2 lost value on the stock market during the second quarter. In comparison, the success rate at stock picking by a hybrid AI system was reported with on average 55.19 to 60.69% [26], and experts had a success rate worse than chance [27]. Thus, the clustering allows a data-driven stock picking with a high chance of success for a short position.

5 Conclusion

This work presents an example in which the DBS algorithm made it possible to apply cluster analysis to high-dimensional data where only a low number of cases exist. Further examples and a comparison to common clustering algorithms as well as to dimension reduction techniques are presented in [10]. The clusters can be explained by simple rules allowing to select stocks for the next quarter by simple thresholds in the data (selling first and buying later). Besides the choice of the number clusters and a Boolean parameter describing the type of structure, DBS is parameter-free and can be downloaded as the R package "DatabionicSwarm" on CRAN.

Acknowledgments
Gratitude goes to Hamza Tayyab for programming the web scrapping algorithm which extracted the quarterly data.

References

[1] Prime-Standard. *Teilbereich des Amtlichen Marktes und des Geregelten Marktes der Deutschen Börse für Unternehmen, die besonders hohe Transparenzstandards* [18.09.2018]; from: http://deutsche-boerse.com/dbg-de/ueber-uns/services/know-how/boersenlexikon/boersenlexikon-article/Prime-Standard/2561178.

[2] Gelistete Unternehmen in Prime Standard, G.S.u.S., *http://www.deutsche-boerse-cash-market.com/dbcm-de/instrumente-statistiken/statistiken/gelistete-unternehmen*. 2018, Deutsche Börse: Frankfurt.

[3] Abad, C., S.A. Thore, and J. Laffarga, *Fundamental analysis of stocks by two‐stage DEA.* Managerial and Decision Economics, 2004. **25**(5): p. 231-241.

[4] O'Neil, W.J., *How to make money in stocks*. Vol. 10. 1988: McGraw-Hill New York.

[5] Deboeck, G.J. and A. Ultsch, *Picking stocks with emergent self-organizing value maps.* Neural Network World, 2000. **10**(1): p. 203-216.

[6] Ou, J.A. and S.H. Penman, *Financial statement analysis and the prediction of stock returns.* Journal of accounting and economics, 1989. **11**(4): p. 295-329.

[7] Mohanram, P.S., *Separating winners from losers among lowbook-to-market stocks using financial statement analysis.* Review of accounting studies, 2005. **10**(2-3): p. 133-170.

[8] Richardson, S., I. Tuna, and P. Wysocki, *Accounting anomalies and fundamental analysis: A review of recent research advances.* Journal of Accounting and Economics, 2010. **50**(2-3): p. 410-454.

[9] Yahoo! Finance. *Income statement, Balance Sheet and Cash Flow*. 2018 [cited 2018 29.09.2018]; Available from: https://finance.yahoo.com/quote/SAP/financials?p=SAP (Exemplary).

[10] Thrun, M.C., *Projection Based Clustering through Self-Organization and Swarm Intelligence*. 2018, Heidelberg: Springer.

[11] McCune, B., J.B. Grace, and D.L. Urban, *Analysis of ecological communities, chapter 6*. Vol. 28. 2002: MjM software design Gleneden Beach.

[12] Hartigan, J.A. and P.M. Hartigan, *The dip test of unimodality.* The annals of Statistics, 1985. **13**(1): p. 70-84.

[13] Ultsch, A. and M.C. Thrun, *Credible Visualizations for Planar Projections*, in *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization (WSOM)*. 2017, IEEE: Nany, France. p. 1-5.

[14] Dasgupta, S. and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss.* Random Structures & Algorithms, 2003. **22**(1): p. 60-65.

[15] Thrun, M.C. and A. Ultsch. *Investigating Quality measurements of projections for the Evaluation of Distance and Density-based Structures of High-Dimensional Data*. in *European Conference on Data Analysis (ECDA)*. 2018. Paderborn, Germany.

[16] Patterson, T. and N.V. Kelso, *Hal Shelton revisited: Designing and producing natural-color maps with satellite land cover data.* Cartographic Perspectives, 2004(47): p. 28-55.

[17] Thrun, M.C., et al., *Visualization and 3D Printing of Multivariate Data of Biomarkers*, in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2016: Plzen. p. 7-16.

[18] Lötsch, J. and A. Ultsch. *Exploiting the Structures of the U-Matrix*. in *Advances in Self-Organizing Maps and Learning Vector Quantization*. 2014. Mittweida, Germany: Springer International Publishing.

[19] Dijkstra, E.W., *A note on two problems in connexion with graphs.* Numerische mathematik, 1959. **1**(1): p. 269-271.

[20] Breiman, L., et al., *Classification and regression trees*. 1984: CRC press.

[21] Bragg, S. *Net income from continuing operations*. 2018 [cited 2018 03.11.2018]; Available from: https://www.accountingtools.com/articles/2017/5/12/net-income-from-continuing-operations.

[22] Silver, C., et al. *Operating Income*. 2014 [cited 2018 03.11.2018]; Available from: https://www.investopedia.com/terms/o/operatingincome.asp.

[23] Thrun, M.C. and A. Ultsch, *Effects of the payout system of income taxes to municipalities in Germany*, in *12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, 2018, Foundation of the Cracow University of Economics: Cracow, Poland. p. 533-542.

[24] Ultsch, A., *Is Log Ratio a Good Value for Measuring Return in Stock Investments?*, in *Advances in Data Analysis, Data Handling and Business Intelligence*. 2009, Springer. p. 505-511.

[25] Behnisch, M. and A. Ultsch, *Knowledge Discovery in Spatial Planning Data: A Concept for Cluster Understanding*, in *Computational Approaches for Urban Environments*. 2015, Springer. p. 49-75.

[26] Tsaih, R., Y. Hsu, and C.C. Lai, *Forecasting S&P 500 stock index futures with a hybrid AI system.* Decision Support Systems, 1998. **23**(2): p. 161-174.

[27] Torngren, G. and H. Montgomery, *Worse than chance? Performance and confidence among professionals and laypeople in the stock market.* The Journal of Behavioral Finance, 2004. **5**(3): p. 148-153.

[28] Thrun, M.C. and A. Ultsch, *Analyzing the Fine Structure of Distributions*. Technical Report being submitted, Dept. of Mathematics and Computer Science, Philipps-University of Marburg, 2019: Marburg: p.1-22.