ELM Clustering – Application to Bankruptcy Prediction

ELM Clustering
– Application to Bankruptcy Prediction –

Anton Akusok[1], David Veganzones[7,8], Kaj-Mikael Björk[3], Eric Séverin[6], Philippe du Jardin[5], Amaury Lendasse[1,2,3,4,8], Yoan Miche[1,9]

[1] Department of Information and Computer Science, Aalto University School of Science, FI-00076, Finland
[2] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
[3] Arcada University of Applied Sciences, 00550 Helsinki, Finland
[4] Department of Mechanical and Industrial Engineering, The University of Iowa, Iowa City, IA 52242-1527, USA
[5] EDHEC Business School, BP3116, 06202 Nice cedex 3 - France
[6] University of Lille 1, IAE, 104 avenue du peuple Belge, 59043 Lille, France
[7] Departamento de Engenharia de Teleinformática (DETI), Universidade Federal do Ceará (UFC) - Campus do Pici, CP 6005, CEP 60455-970, Fortaleza, Ceará, Brazil
[8] Computer Science Faculty, University Of The Basque Country, Paseo Manuel Lardizabal 1, Donostia/San Sebastián, Spain
[9] Ecole Centrale de Lille, Cité Scientifique, 59651 Villeneuve-d'Ascq, France

Abstract. This paper presents a new clustering technique based on Extreme Learning Machine (ELM). This clustering technique can incorporate *a priori* knowledge (of an expert) to define the optimal structure of the clustering; for example, the number of points in each cluster. Using ELM, the clustering can be rewritten as a Traveling Salesman Problem and solved by a Genetic Algorithm. This efficient and fast clustering technique is used in order to analyze and predict financial distress of French companies.

1 Introduction

In the field of corporate finance, bankruptcy forecasting is one of the most studied problems [1–5]. Prediction models are usually designed using financial ratios and rely on regression or classification techniques. As there is no general theory of business failure, all these models are empirical. As a consequence, they do not rely on any *a priori* knowledge and are therefore strongly data-dependent. This is the reason why we have designed a method that allows to take into account prior knowledge so as to build a prediction model and that is based on a clustering technique. In this paper, clustering is used in order to analyze the ratios and predict this unknown financial state.

Clustering is the general task of grouping *similar* objects together [6]. A unique solution does not always exist for any given input dataset [7], apart from the trivial assignment of all data samples to one cluster, or assigning one cluster per each sample. Clustering algorithms utilize different assumptions about the data: hierarchical clustering [8, 9] groups *nearby* objects together, k-means [10, 11] clustering finds dense clusters in a less dense space, and Expectation-Maximization (EM) algorithm [12] assumes that distribution of samples in clusters can be approximated by multivariate Gaussian distribution(s) [13, 14].

The number of clusters can be estimated automatically by some methods (like a density-based DBSCAN [15]), and in other methods like k-means it is a hyper-parameter, optimized using a cost function. Not all the existing clusters in a dataset are easily separable, for instance EM algorithm poorly distinguishes density-based clusters, and k-means algorithm tends to find clusters of similar size [16]; therefore the exact number of clusters may be complex to find.

There is a special class of clustering problems, where the desired cluster configuration is given in advance. Number of clusters and number of samples in each cluster are known *a priori* (expert *a priori* knowledge), and the cluster assignment of each sample is to be found. This configuration helps solving tasks with highly uneven sizes of clusters. The solution is a mapping between data samples and the desired clusters, which can be generalized to new samples and used for prediction.

The problem of finding a mapping between a set of data samples and a set of "sample slots" in clusters (*i.e.* find which sample should be in which cluster) is an NP-hard set ordering problem. Heuristic methods are a suitable approach [17]. One requirement is a fast general cost function, which estimates how well the samples are mapped to the clusters. It should give a high value for distant samples mapped to the same cluster, and a low value for similar samples mapped together. An Extreme Learning Machine (ELM) provides such fast nonlinear cost function and is used in the methodology described in Section 2. In section 3, the proposed methodology is illustrated on a toy example and on a real financial dataset.

2 Methodology

The methodology of the proposed clustering task is based on an Extreme Learning Machine (ELM) model [18, 19]. With a fixed set of inputs (data samples in the inputs space), ELM is adapted to calculate training error to arbitrary outputs (cluster assignments) extremely fast. Using ELM for error estimation is possible because it approximates a smooth function - similar inputs lead to similar outputs, so a low error indicates that samples inside a cluster are similar, while samples in different clusters are dissimilar. In the following methodology, no distance-measuring is needed, but only a limited number of matrix multiplications are used. Compared to traditional clustering techniques, this methodology allows the *a priori* insertion of expert knowledge in the form of the number of samples for each cluster.

The next Subsection describes the original ELM, the following one, an adaptation of ELM for fast error estimation in the context of this methodology, and the final Subsection summarizes the whole method.

2.1 Original Extreme Learning Machine

The Extreme Learning Machine algorithm is originally proposed by Guang-Bin Huang *et al.* in [18] and improved in [20–22] and analyzed in [23], and it uses the structure of a Single Layer Feed-forward Neural Network (SLFN). The main concept behind ELM is the replacement of a computationally costly procedure of training the hidden layer, by its random initialization. Then an output weights matrix between the hidden representation of inputs and the outputs remains to be found. The ELM is proven to be a universal approximator given enough hidden neurons [19]. It works as following:

Consider a set of N distinct samples $(\mathbf{x}_i, \mathbf{y}_i)$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathbf{y}_i \in \mathbb{R}^c$. Then a SLFN with n hidden neurons is modeled as $\sum_{j=1}^{n} \beta_j \phi(\mathbf{w}_j \mathbf{x}_i + b_j),\ i \in [1, N]$, with $\phi : \mathbb{R} \to \mathbb{R}$ being the activation function, $\mathbf{w}_j$ the input weights, b_j the biases and β_j the output weights.

In case the SLFN would perfectly approximate the data, the errors between the estimated outputs $\hat{\mathbf{y}}_i$ and the actual outputs $\mathbf{y}_i$ are zero, and the relation between inputs, weights and outputs is then $\sum_{j=1}^{n} \beta_j \phi(\mathbf{w}_j \mathbf{x}_i + b_j) = \mathbf{y}_i, i \in [1, N]$ which can be written compactly as $\mathbf{H}\boldsymbol{\beta} = \mathbf{Y}$, with $\boldsymbol{\beta} = (\beta_1^T \dots \beta_n^T)^T$, $\mathbf{Y} = (\mathbf{y}_1^T \dots \mathbf{y}_N^T)^T$.

Solving the output weights $\boldsymbol{\beta}$ from the hidden layer representation of inputs $\mathbf{H}$ and true outputs $\mathbf{Y}$ is achieved using the Moore-Penrose generalized inverse of the matrix $\mathbf{H}$, denoted as $\mathbf{H}^\dagger$ [24]. The training of ELM requires no iterations, and the most computationally costly part is the calculation of a pseudo-inverse of the matrix $\mathbf{H}$, which makes ELM an extremely fast Machine Learning method.

2.2 Fast training error estimation with ELM

Calculation of a cost function requires having cluster centers and data points in the same space. Clusters are typically given by a 1-in-all code, which produces independent and equally spaced cluster centers. However projecting input data to that space introduces errors into distances between data samples, which may hamper the clustering. A better solution is to perform the clustering in the original data space (with original between-sample distances) and project cluster centers there.

Cluster centers can be projected into input data space with an ELM. The ELM is used in a "reversed" way — cluster indexes of the samples are the inputs, and the original samples are the desired outputs. Sample indexes are projected to the hidden layer output matrix $\mathbf{H}$, the output weights $\boldsymbol{\beta}$ are calculated with a pseudo-inverse, and the training error is obtained as $E_{\text{train}} = \frac{1}{N}||\mathbf{H}\boldsymbol{\beta} - \mathbf{X}||_2^2$.

Denoting by $\boldsymbol{\beta}^{\text{old}} = \mathbf{H}^\dagger \mathbf{X}^{\text{old}}$ the output coefficients, the methodology "swaps" samples in $\mathbf{X}$ so as to find a better — regarding the error criterion — assignment of samples to the respective clusters. This modification of $\mathbf{X}$ requires a re-computation of the $\boldsymbol{\beta}$ coefficients, denoted then $\boldsymbol{\beta}^{\text{new}}$ as

$$\boldsymbol{\beta}^{\text{new}} = \mathbf{H}^{\dagger}\mathbf{X}^{\text{new}},\ \mathbf{X}^{\text{new}} \text{ s.t. } \begin{cases} \mathbf{X}^{\text{new}}_{(i)} & \leftarrow \mathbf{X}^{\text{old}}_{(j)} \\ \mathbf{X}^{\text{new}}_{(j)} & \leftarrow \mathbf{X}^{\text{old}}_{(i)} \end{cases}, \tag{1}$$

where $\mathbf{X}^{\text{new}}$ is the matrix $\mathbf{X}^{\text{old}}$ for which lines i and j are swapped as in Eq. 1.

While $\mathbf{H}^{\dagger}$ remains fixed, the matrix product $\mathbf{H}^{\dagger}\mathbf{X}^{\text{new}}$ is computationally costly when repeated for many different "swaps". Thus, we propose to use the update formula for the output weights $\boldsymbol{\beta}^{\text{new}}$ as

$$\boldsymbol{\beta}^{\text{new}} = \boldsymbol{\beta}^{\text{old}} - \mathbf{H}^{\dagger(i)}\mathbf{X}_{(i)} - \mathbf{H}^{\dagger(j)}\mathbf{X}_{(j)} + \mathbf{H}^{\dagger(i)}\mathbf{X}_{(j)} + \mathbf{H}^{\dagger(j)}\mathbf{X}_{(i)}, \tag{2}$$

where $\mathbf{H}^{\dagger(i)}$ represents the matrix $\mathbf{H}^{\dagger}$ where all entries are put to zero except for the *column* i, and $\mathbf{X}_{(i)}$ represents the matrix $\mathbf{X}$ where all entries are put to zero except for the *line* i.

The error for the updated data matrix is thus calculated as

$$E^{\text{new}}_{\text{train}} = \frac{1}{N}||\mathbf{H}\boldsymbol{\beta} - \mathbf{X}^{\text{new}}||_2^2, \tag{3}$$

which is the criterion that is optimized in the methodology.

Keeping in mind that $\mathbf{H}^{\dagger}$ is already computed, this update formula only recomputes the elements of the product $\mathbf{H}^{\dagger}\mathbf{X}^{\text{new}}$ that are affected by the swap of two samples in $\mathbf{X}$, and reduces the computational effort. Fig. 1 summarizes the idea of the above mentioned methodology.

2.3 Clustering with ELM

The clustering method starts with assigning random data samples from $\mathbf{X}$ to "sample slots" in clusters, and estimating the training error (see Eq. 3) with the resulting ELM. At this initial step, the number n of hidden neurons in ELM is adjusted for the lowest possible error, and used throughout the rest of the method.

The problem of assigning data samples to the clusters in the best way — w.r.t. the error criterion —, is equivalent to an optimization problem known as the Travelling Salesman Problem. The proposed methodology searches for a solution to this problem by random permutations of data samples, which is equivalent to a search using a Genetic Algorithm where only mutations are used [25–28].

In this methodology, the number of samples swapped in one iteration (*i.e.* when building one $\mathbf{X}^{\text{new}}$) is chosen randomly, as are the samples that are swapped. The method has thus virtually no hyper-parameters (apart from the selection of the number of neurons for the ELM structure).

The following Section presents the results of the clustering on a toy data set, as well as on real financial data for bankruptcy prediction.

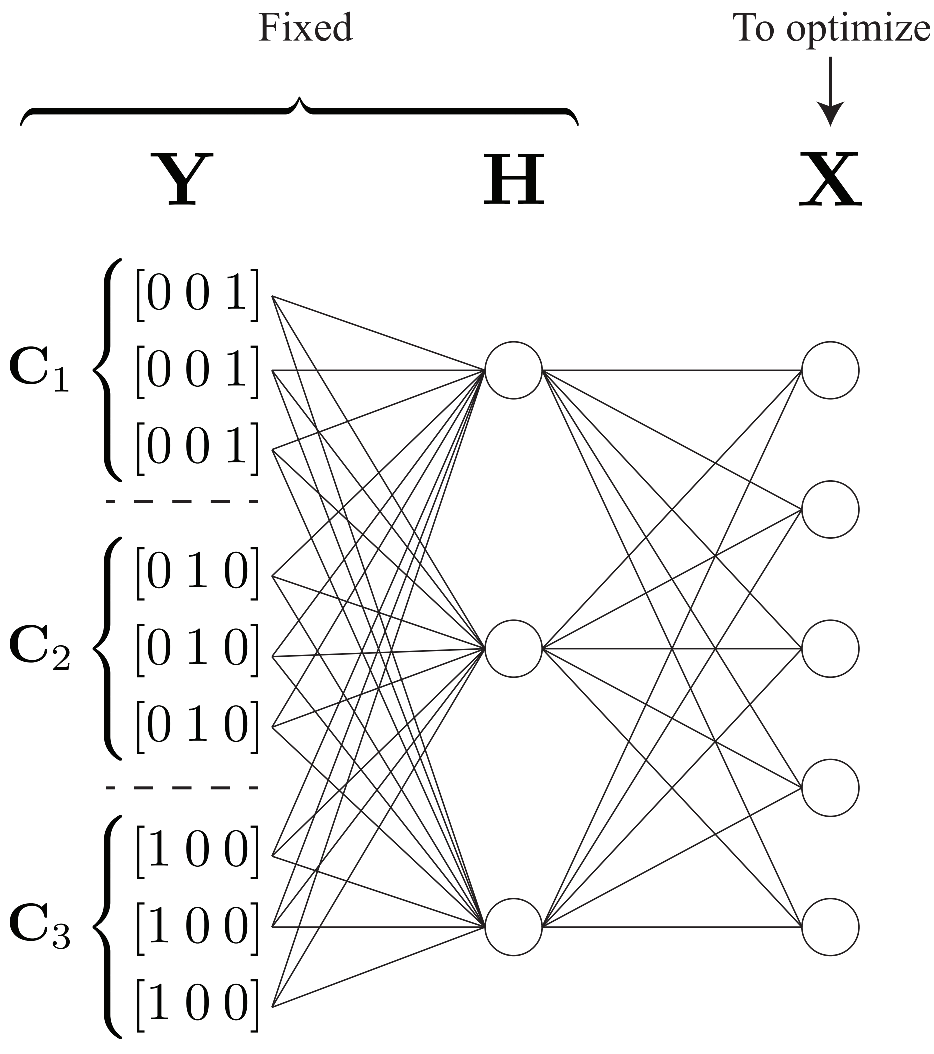

Fig. 1. Illustration of the methodology for a 3-clusters case ($\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$) and 3 hidden neurons. Note that the traditional ELM structure is "reversed" in this methodology, so as to allow the swapping of data samples from $\mathbf{X}$. With this structure, most of the matrices remain unchanged in the ELM structure, as $\mathbf{H}$ need not be re-computed when samples are swapped in $\mathbf{X}$. See subsection 2.2 for details.

3 Experimental results

3.1 Toy Data set

The toy data set displayed on Fig. 2 is made of 1000 samples lying in $\mathbb{R}^2$ in a circular arrangement.

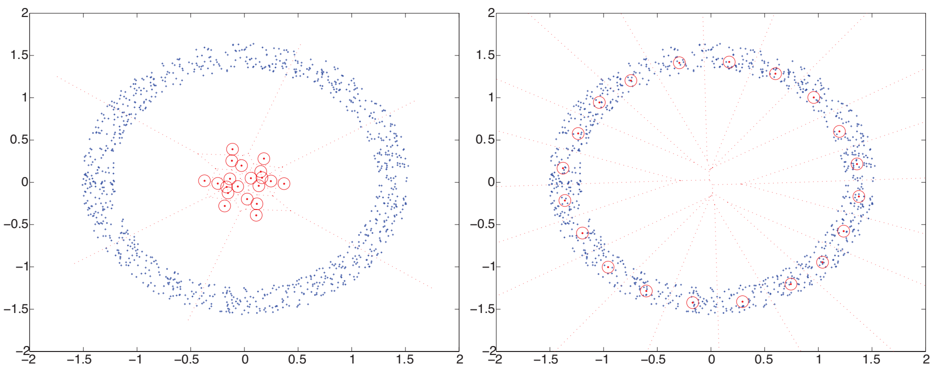

Fig. 2. Convergence of clustering with ELM: cluster centroids are initialized randomly in the middle, and get distributed evenly across the data points.

The methodology converges to a solution in which the cluster centers — depicted by red circles — converge from the initial randomly centered arrangement in the left figure of Fig. 2, to the even distribution on the right figure.

3.2 Financial Data set

This data set includes 500 firms from year 2002 [29]. In the data set, the proportion of healthy and bankrupt firms is 50:50 and the firms are all from the trade sector. Before clustering the data, variable selection is applied with 7 variables selected for the clustering. 4 out of these 7 selected variables measure the profitability (Profit before tax / Shareholders funds), (Net income / Shareholders funds), (EBITDA / Total assets) and (Net income / Total Assets). One of these selected variables measures the financial structure (Shareholders funds / Total assets). The last 2 out of these 7 selected variables measure the liquidity (Cash / Total assets) and (Cash / Total debts). The variables ("ratios" in the following) and their structure are summarized in Tab. 1.

In the methodology, each firm is clustered to be associated with similar firms. The firms are clustered into 5 balanced clusters, so there are 100 firms in each cluster. Means and standard deviations of the selected variables are computed for each cluster to get an overview of the cluster characteristics. These are used for the detection of mislabeled samples as performed in [30].

The following paragraphs provide an analysis of the contents of the 5 clusters after convergence.

Ratios	Structure of the ratios
Profit1	Profit before tax / Shareholders funds
Profit2	Net income / Shareholders funds
Profit3	EBITDA / Total assets
Profit4	Net income / Total assets
Fin.Str.	Shareholders funds / Total assets
Liq.1	Cash / Total assets
Liq.2	Cash / Total debts

Table 1. Structure of the ratios.

Ratios	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2
	-0.0489	0.3228	0.0567	0.0066	0.1385	-0.1448	-0.1688

Table 2. Means of the liquidity problem firms cluster. Structure of the ratios is given in Tab. 1.

Liquidity problem firms cluster The first cluster includes 68 bankrupt firms and 32 healthy ones. Analyzing means (Tab. 2) and standard deviations is observed that the main characteristic of this cluster is the liquidity problem of the firms. The mean of the cash ratios that measure liquidity (cash/total assets; and cash/total debts) are both negative and significant. Cash is computed as (cash - short term debts). The negative signs of the cash ratios suggest that the firms do not have enough liquidity to pay back current debts, so the firms could be bankrupt or close to it. Thus, in this cluster those 32 firms labeled as healthy could be close to bankruptcy due to liquidity problem or might be already bankrupt and be an outlier/mislabeled. Hence, the healthy firms are analyzed in order to find possible outliers/mislabeled. The most of the healthy firms are profitable so they are able to reimburse its debt but, maybe due to not be paid by the clients yet, the cash ratios are negative. However, 5 samples out of these 32 healthy firms are identified as possible outliers/mislabeled. In fact, the MD-ELM method used in [30] has identified 3 out of these 5 as mislabeled. Those samples are: #41, #160 and #301. The sample #41 is surely mislabeled but there are some disagreement for sample #160 and #301. The samples #61 and #357 are also identified as possible outliers/mislabeled. The identified samples have liquidity problems and they are low profitable or nonprofitable firms, so they could be bankrupt (the means variable of the outlier/mislabeled samples in the appendix, Tab. 7.)

Ratios	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2
	0.8209	0.3883	0.1545	0.1057	0.3421	0.1586	0.2465

Table 3. Means of the profitable but high level of debt firms cluster. Structure of the ratios is given in Tab. 1.

Profitable but high level of debt firms cluster The second cluster includes 12 bankrupt firms and 88 healthy ones. The firms included in this cluster are profitable but they have high level of debt, on average (Tab. 3). The profitability ratios are positive and quite good but, the financial structure ratio (shareholder's fund/total assets) is less than 0.5, on average. It means that firms are using debts to increase the profits. Thus, firms that use debts to be more profitable take high risk strategy, because the firms should be enough profitable and have high liquidity in order to reimbursed its debts. Hence, 12 firms are labeled as bankrupt because it seems that they are not able to reimbursed its debts but, it could be that even though they have high level of debt they could afford its payments, so they are outliers/mislabeled. In the analysis of the bankrupt firms, 3 out of these 12 are identified as possible outliers/mislabeled. Those samples are: #448, #482 and #490. In fact, the MD-ELM method [30] identified those samples as mislabeled. The sample #448 is surely mislabeled. However, there is a disagreement for sample #482 and #490, because those samples are quite profitable but maybe not enough to ensure the payback of the debts (the means variable of the outlier/mislabeled samples in the appendix, Tab. 8.)

Ratios	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2
	0.0398	-2.4548	0.0424	-0.0108	0.0033	0.0695	0.0765

Table 4. Means of the profitability problems with high level of debt firms cluster. Structure of the ratios is given in Tab. 1.

Profitability problems with high level of debt firms cluster The third cluster includes 68 bankrupt firms and 32 healthy ones. The main characteristics of this cluster are negative net of income (profitability problems) and high level of debt (Tab. 4). The profitability ratios are negative or really low and the financial structure ratio (shareholders fund/total assets) is extremely bad, 0.0033 on average. The most of the firms are using debts to finance its investment but they are not enough profitable, so they are already bankrupt or on way of being bankrupt. 32 firms are labeled as healthy, even though they have low profitability and high level of debt. In the analysis, 5 out of these 32 are identified as possible outliers/mislabeled. The MD-ELM method [30] also identifies the sample #168 as mislabeled but there is a disagreement. The others identified sample have very low profitability or negative profitability and high level of debt, so they could be bankrupt. Those samples are: #139, #184, #370 and #383 (the means variable of the outlier/mislabeled samples in the appendix, Tab. 9.)

Profitable with low debt and high liquidity firms cluster The fourth cluster includes 2 bankrupt firms and 98 healthy ones. The firms selected in this cluster are very profitable, the financial structure of the firms is mainly compound by shareholder's funds, so the level of debt is quite low and they

Ratios	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2
	2.4821	0.4885	0.2849	0.2300	0.4669	0.4338	0.9906

Table 5. Means of the profitable with low debt and high liquidity firms cluster. Structure of the ratios is given in Tab. 1.

have high liquidity (Tab. 5). In fact, those most of the firms are absolutely healthy and they are able to pay back its debts. However, two bankrupt firms are selected in this cluster, so they are analyzed in order to know whether they are correctly labeled. One out of these 2 bankrupt firms is identified as possible outlier/mislabeled. This firm, the sample #465, is quite profitable and the cash ratios are positive and high so it is able to pay back its debt but, its financial structure is based on debts. In fact, it is also identified as mislabeled using MD-ELM method [30] so it could be healthy firm, but there is a disagreement (the means variable of the outlier/mislabeled samples in the appendix, Tab. 10.)

Ratios	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2
	-13.5254	2.5444	-0.3227	-0.4980	-0.6540	-0.0292	-0.0240

Table 6. Means of the Nonprofitable with debt and liquidity problem firms cluster. Structure of the ratios is given in Tab. 1.

Nonprofitable with debt and liquidity problem firms cluster The fifth and last cluster includes 100 firms, all of them bankrupt. On average, the firms are nonprofitable, the financial structure of the firms is negative so the debts are higher than the shareholder's funds, and the cash ratios are also negative, firms do not have enough cash in order to pay back its current debts (Tab. 6). Hence, it seems that all firms are bankrupt taking into account the means of the variables. However, the selected samples are analyzed in order to find possible outliers/mislabeled. No outlier/mislabeled is found so it means that all firms are correctly labeled as bankrupt.

4 Conclusions and Further Work

The proposed methodology enables fast and accurate clustering of large high-dimensional data. No distance-measuring is needed, instead only matrix multiplications are used. Compared to traditional clustering techniques, the proposed methodology allows the *a priori* insertion of expert knowledge such as the number of samples for each cluster.

For example, in the toy example and the real financial dataset, the inclusion of *a priori* knowledge provides an accurate and coherent clustering that can be straightforwardly analyzed by experts in the field.

In further work, the authors will include an *a priori* structure for the clustering. As a result, it will be possible to obtain self-organization of the clusters. Furthermore, the new method will combine all the benefits of Self-Organizing Maps [31] and ELM Clustering.

In the meantime, an extended financial data set will be studied. It will include several consecutive years of financial ratios; temporal trajectories [32] will be visible in the self-organized clustering. It will be then possible to predict the financial distress of companies and dynamical/temporal factors that lead to bankruptcy will be identified.

Appendix

Samples	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2	Class
#41	-0.2354	-0.3072	-0.2304	-0.1514	0.4926	-0.1858	-0.3662	H
#61	0.0658	-0.0058	0.0751	-0.0017	0.2857	-0.0427	-0.0598	H
#160	1.3897	-0.0050	0.0712	-0.0007	0.1437	-0.1279	-0.1500	H
#301	0.1181	0.0071	0.0714	0.0012	0.1734	-0.0486	-0.0589	H
#357	0.0459	0.0587	0.0992	0.0288	0.4909	-0.2138	-0.4221	H

Table 7. Value of the variables of outliers/mislabeled samples of the cluster1. Class: H is healthy and B is bankrupt.

Samples	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2	Class
#448	0.9278	0.8285	0.3785	0.2945	0.3555	0.1485	0.2559	B
#482	0.2881	0.2631	0.1750	0.1044	0.3966	0.1934	0.3198	B
#490	0.2148	0.2286	0.1111	0.0743	0.3248	0.2163	0.3204	B

Table 8. Value of the variables of outliers/mislabeled samples of the cluster2. Class: H is healthy and B is bankrupt.

Samples	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2	Class
#139	0.0740	0.0547	0.0638	0.0161	0.2941	0.0641	0.0908	H
#168	0.0397	-0.0200	0.0424	-0.0049	0.2452	0.0005	0.0007	H
#184	0.3617	0.4909	0.0700	0.0806	0.1642	0.0764	0.0914	H
#370	0.0558	0.0991	0.0651	0.0356	0.3595	0.0538	0.0841	H
#383	0.0509	0.1075	0.0656	0.0219	0.2038	0.0967	0.1214	H

Table 9. Value of the variables of outliers/mislabeled samples of the cluster3. Class: H is healthy and B is bankrupt.

Samples	Profit1	Profit2	Profit3	Profit4	Fin.Str.	Liq.1	Liq.2	Class
#465	0.0201	0.5558	0.1749	0.1471	0.2646	0.4669	0.7219	B

Table 10. Value of the variables of outliers/mislabeled samples of the cluster4. Class: H is healthy and B is bankrupt.

References

1. Q. Yu, Y. Miche, E. Séverin, and A. Lendasse, “Bankruptcy prediction using extreme learning machine and financial expertise,” *Neurocomputing*, vol. 128, no. 0, pp. 296 – 302, 2014.
2. E. I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy,” *The Journal of Finance*, vol. 23, no. 4, pp. pp. 589–609, 1968.
3. L. Kainulainen, Y. Miche, E. Eirola, Q. Yu, B. Frénay, E. Séverin, and A. Lendasse, “Ensembles of local linear models for bankruptcy analysis and prediction,” *Case Studies in Business, Industry and Government Statistics (CSBIGS)*, vol. 4, no. 2, November 2011. [Online]. Available: http://www.bentley.edu/centers/csbigs/volume-4-number-2
4. P. du Jardin and E. Séverin, “Forecasting financial failure using a kohonen map: A comparative study to improve model stability over time,” *European Journal of Operational Research*, vol. 221, no. 2, pp. 378 – 396, 2012.
5. P. du Jardin, “Predicting bankruptcy using neural networks and other classification methods: The influence of variable selection techniques on model accuracy,” *Neurocomputing*, vol. 73, no. 10–12, pp. 2047 – 2060, 2010.
6. R. B. Cattell, “The description of personality: basic traits resolved into clusters.” *The journal of abnormal and social psychology*, vol. 38, no. 4, p. 476, 1943.
7. V. Estivill-Castro, “Why so many clustering algorithms: a position paper,” *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 1, pp. 65–75, 2002.
8. T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.
9. R. Sibson, “SLINK: an optimally efficient algorithm for the single-link cluster method,” *The Computer Journal*, vol. 16, no. 1, pp. 30–34, 1973.
10. J. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 233, pp. 281–297, 1967.
11. J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *Applied statistics*, pp. 100–108, 1979.
12. A. P. Dempster, N. M. Laird, D. B. Rubin, and Others, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.
13. E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki, “Mixture of gaussians for distance estimation with missing data,” *Neurocomputing*, vol. 131, no. 0, pp. 32 – 42, 2014.
14. A. Lourme and C. Biernacki, “Simultaneous gaussian model-based clustering for samples of multiple origins,” *Computational Statistics*, vol. 28, no. 1, pp. 371–391, 2013.
15. M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise.” in *KDD*, vol. 96, 1996, pp. 226–231.
16. C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information science and statistics, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. Springer, 2006, vol. 4, no. 4.
17. K. Helsgaun, “General k-opt submoves for the Lin–Kernighan TSP heuristic,” *Mathematical Programming Computation*, vol. 1, no. 2-3, pp. 119–163, 2009.
18. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, no. 1, pp. 489–501, Dec. 2006.

19. G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *Neural Networks, IEEE Transactions*, vol. 17, no. 4, pp. 879–892, 2006.
20. Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, "OP-ELM: Optimally-pruned extreme learning machine," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 158–162, January 2010.
21. Y. Miche, M. van Heeswijk, P. Bas, O. Simula, and A. Lendasse, "Trop-elm: A double-regularized ELM using LARS and tikhonov regularization," *Neurocomputing*, vol. 74, no. 16, pp. 2413 – 2421, 2011.
22. M. van Heeswijk, Y. Miche, E. Oja, and A. Lendasse, "GPU-accelerated and parallelized ELM ensembles for large-scale regression," *Neurocomputing*, vol. 74, no. 16, pp. 2430–2437, September 2011.
23. E. Cambria, G.-B. Huang, L. Kasun, H. Zhou, C. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li, V. Leung, L. Feng, Y.-S. Ong, M.-H. Lim, A. Akusok, A. Lendasse, F. Corona, R. Nian, Y. Miche, P. Gastaldo, R. Zunino, S. Decherchi, X. Yang, K. Mao, B.-S. Oh, J. Jeon, K.-A. Toh, A. Teoh, J. Kim, H. Yu, Y. Chen, and J. Liu, "Extreme learning machines," *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 30–59, 2013.
24. C. R. Rao and S. S. K. Mitra, "Generalized inverse of a matrix and its applications," *Wiley Series in Probability and Mathematical Statistics (New York)*, 1971.
25. D. Simon, R. Rarick, M. Ergezer, and D. Du, "Analytical and numerical comparisons of biogeography-based optimization and genetic algorithms," *Information Sciences*, vol. 181, no. 7, pp. 1224–1248, 2011.
26. A. Guillén, M. Arenas, M. van Heeswijk, D. Sovilj, A. Lendasse, L. Herrera, H. Pomares, and I. Rojas, "Fast feature selection in a gpu cluster using the delta test," *Entropy*, vol. 16, no. 2, pp. 854–869, 2014.
27. K.-M. Björk and R. Nordman, "Solving large-scale retrofit heat exchanger network synthesis problems with mathematical optimization methods," *Chemical Engineering and Processing: Process Intensification*, vol. 44, no. 8, pp. 869–876, 2005.
28. D. E. Goldberg and R. Lingle, Jr., "Alleles, loci, and the traveling salesman problem," in *Proceedings of the 1st International Conference on Genetic Algorithms*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1985, pp. 154–159.
29. P. du Jardin, *Prévision de la défaillance et réseaux de neurones: l'apport des méthodes numériques de sélection de variables.* Université de Nice-Sophia-Antipolis, 2007.
30. A. Akusok, D. Veganzones, Y. Miche, E. Séverin, and A. Lendasse, "Finding originally mislabels with md-elm," in *ESANN2014: 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2014.
31. A. Sorjamaa, F. Corona, Y. Miche, P. Merlin, B. Maillet, E. Séverin, and A. Lendasse, "Sparse linear combination of SOMs for data imputation: Application to financial database," in *Advances in Self-Organizing Maps - Proceedings of WSOM 2009 International Workshop on Self-Organizing Maps, Saint Augustine (Florida)*, ser. Lecture Notes in Computer Science, vol. 5629/2009. Springer Berlin / Heidelberg, June 8-10 2009, pp. 290–297.
32. E. Séverin, "Self organizing maps in corporate finance: Quantitative and qualitative analysis of debt and leasing," *Neurocomputing*, vol. 73, no. 10-12, pp. 2061–2067, 2010.