A Two-Level Approach to Making Class Predictions

Adrian Costea
Turku Centre for Computer Science and IAMSR / Åbo Akademi University, Turku, Finland, Adrian.Costea@abo.fi

Tomas Eklund
Turku Centre for Computer Science and IAMSR / Åbo Akademi University, Turku, Finland, Tomas.Eklund@abo.fi

Abstract

In this paper we propose a new two-level methodology for assessing countries'/companies' economic/financial performance. The methodology is based on two major techniques of grouping data: cluster analysis and predictive classification models. First we use cluster analysis in terms of self-organizing maps to find possible clusters in data in terms of economic/financial performance. We then interpret the maps and define outcome values (classes) for each data row. Lastly we build classifiers using two different predictive models (multinomial logistic regression and decision trees) and compare the accuracy of these models. Our findings claim that the results of the two classification techniques are similar in terms of accuracy rate and class predictions. Furthermore, we focus our efforts on understanding the decision process corresponding to the two predictive models. Moreover, we claim that our methodology, if correctly implemented, extends the applicability of the self-organizing map for clustering of financial data, and thereby, for financial analysis.

1. Introduction

In this study, we are interested in the relationship between a number of macro/microeconomic indicators of countries/companies and different economic/financial performance classifications. We have based our research on two previous studies [2] and [3]. In [2] we compared two different methods of clustering central-east European countries economic data (self-organizing maps and statistical clustering) and presented the advantages and disadvantages of each method. In [3], the self-organizing map (SOM) was used for benchmarking international pulp and paper companies. In both previous studies we were mainly concerned with finding patterns in economic/financial data and presenting this multi-dimensional data in an easy-to-read format (using SOM maps). However, we have not addressed the problem of class prediction as new cases are added to our datasets. From our previous results we cannot directly infer a procedure with which a new data row could be fit into our maps. As we obtain new data, depending upon the standardization technique used, we may be forced to retrain the maps, and repeat the entire clustering process. This is very time consuming, and requires the effort of an experienced SOM user. As Witten & Frank say in their book on data mining: "The success of clustering is measured subjectively in terms of how useful the result appears to be to a human user. It may be followed by a second step of classification learning where rules are learned that give an intelligible description of how new instances should be placed into the clusters." [17, p.39]

Here we propose a methodology that enables us to model the relationship between economic/financial variables and different classifications of countries/companies in terms of their performances. Defining the model permits us to predict the class (cluster) to which a new case belongs. In other words, we insert new data into our model and identify where they fit in the previously constructed map. Choosing the best technique for these two phases of our analysis (clustering/benchmarking/visualization and class prediction) is not a trivial task. In the literature there is a large number of techniques for both clustering and class prediction.

In this study, we use SOM as the clustering technique due to the advantages of good visualization and reduced computational cost. Even with a relatively small number of samples, many clustering algorithms – especially hierarchical ones (for example, Unweighted Pair Group Method with Arithmetic Mean (UPGMA), Ward's, or other bottom-up hierarchical clustering methods) – become intractably heavy [16].

Descriptive techniques, such as clustering, simply summarize data in convenient ways, or in ways that we hope will lead to increased understanding. In contrast, predictive techniques, such as multinomial logistic regression and decision trees, allow us to predict the probability that data rows will be clustered in a specific class in the trained SOM model. In order to find the predictive technique that is most suitable in our particular case, we conduct two experiments using multinomial logistic regression and decision tree techniques. When building real classifiers one can use three different

Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)
0-7695-1874-5/03 $17.00 © 2002 IEEE

fundamental approaches: the *discriminative approach*, the *regression approach*, and the *class-conditional approach* [6, p.335]. We chose to compare two regression approach methods: *multinomial logistic regression* and *decision trees*.

The rest of the paper is structured as follows. In Section two we present our methodology. In Section three, the datasets are presented and SOM clustering is performed. In Sections four and five, the multinomial regression and decision tree models are built and validated, and in Section six the models are compared. Finally, in Section seven, we present our conclusions.

2. Methodology

In our two-level approach we add another level (class prediction phase) to SOM clustering, as is depicted in Figure 1 (the arrows are the levels):

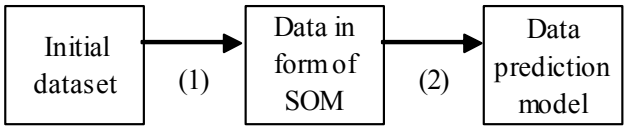

Figure 1. Two-level methodology

(1) – consists of several stages: preprocessing of initial data, training using the SOM algorithm, choosing the best maps, identifying the clusters, and attaching outcome values to each data row; [1]

(2) – depending on the technique that we apply, there can be different stages for this methodology level. When applying statistical techniques, such as multinomial logistic regression, we follow these steps: developing the analysis plan, estimation of logistic regression, assessing model fit (accuracy), interpreting the results, and validating the model. When applying the decision tree algorithm: constructing a decision tree step by step including one attribute at a time in the model, assessing model accuracy, interpreting the results, and validating the model.

After the predictive models for classification were constructed we compared them, based on their accuracy measures. Quinlan [10] states that there are different ways of comparing models besides their accuracy, e.g. the insight provided by the predictive model. However, we will use the accuracy measure since the example above is a subjective measure.

3. Clustering Using SOM

The SOM algorithm stands for self-organizing map algorithm, and is based on a two-layer neural network using the unsupervised learning method. The self-organizing map technique creates a two-dimensional map from n-dimensional input data. This map resembles a landscape in which it is possible to identify borders that define different clusters [8]. These clusters consist of input variables with similar characteristics, i.e. in this report, of countries/companies with similar economic/financial performance. The methodology used when applying the self-organizing map is as follows [1]. First, we choose the data material. It is often advisable to standardize the input data so that the learning task of the network becomes easier [8]. After this, we choose the *network topology*, *learning rate*, and *neighborhood radius*. Then, the network is constructed. The construction process takes place by showing the input data to the network iteratively using the same input vector many times, the so-called *training length*. The process ends when the *average quantization error* is small enough. The best map is chosen for further analysis. Finally, we identify the clusters using the *U-matrix* and interpret the clusters (assign labels to them) using the *feature planes*. From the feature planes we can read per input variable per neuron the value of the variable associated with each neuron.

The network topology refers to the form of the lattice. There are two commonly used lattices, *rectangular* and *hexagonal*. The hexagonal lattice is preferable for visualization purposes as it has six neighbors, as opposed to four for the rectangular lattice [8]. The learning rate refers to how much the winning input data vector affects the surrounding network. The neighborhood radius refers to how much of the surrounding network is affected. The average quantization error indicates the average distance between the best matching units and the input data vectors. Generally speaking, a lower quantization error indicates a better-trained map.

The sample data size is not of a major concern when using SOM algorithm. In [15] the author claims that SOM is easily applicable to small data sets (less than 10000 records) but can also be applied in case of medium sized data sets.

To visualize the final self-organizing map we use the unified distance matrix method (U-matrix). The U-matrix method can be used to discover otherwise invisible relationships in a high-dimensional data space. It also makes it possible to classify data sets into clusters of similar values. The simplest U-matrix method is to calculate the distances between neighboring neurons, and store them in a matrix, i.e. the output map, which then can be interpreted. If there are "walls" between the neurons, the neighboring weights are distant, i.e. the values differ significantly. The distance values can also be displayed in color when the U-matrix is visualized. Hence, dark colors represent great distances while brighter colors indicate similarities amongst the neurons. [14]

3.1. Datasets

In this study we have used two datasets from our previous papers: one dataset on the general economic performance (EconomicPerf) of the central-east-European countries [2] and another (FinancialPerf) on the financial

performance of international pulp and paper companies [3]. The variables for the first dataset are:

- Currency Value, or how much money one can buy with 1000 USD, depicts the purchasing power of each country's currency (the greater the better),
- Domestic Prime Rate (Refinancing Rate), which shows financial performance and level of investment opportunities (the smaller the better),
- Industrial Output in percentages to the previous periods, to depict industrial economical development (the greater the better),
- Unemployment Rate, which characterizes the social situation in the country (the smaller the better), and
- Foreign Trade in millions of US dollars, to reveal the deficit/surplus of the trade budget (the greater the better).

In [2] there were two more variables in the dataset: import and export in million USD, as intermediary measures to calculate the foreign trade. We did not take them into account here, since they are strongly correlated with the foreign trade variable. Also, we have replaced the first variable (Foreign Exchange Rate) from the previous study [2] with Currency Value, which is calculated from the Foreign Exchange Rate variable by reversing it and multiplying the result with 1000. We have changed this variable to ensure the comparability among different countries' currencies.

Our dataset contains monthly/annual data for six countries (Russia, Ukraine, Romania, Poland, Slovenia and Latvia) during 1993-2000, in total 225 cases with five variables each. We have in some cases encountered lack of data, which we have completed using means of existing values. However, the self-organizing map algorithm can treat the problem of missing data simply by considering at each learning step only those indicators that are available [7].

The second dataset consisted of financial data on international pulp and paper companies. The dataset covered the period 1995-2000, and consisted of seven financial ratios per year for each company. The ratios were chosen from an empirical study by Lehtinen [9], in which a number of financial ratios were evaluated concerning their validity and reliability in an international context. The ratios chosen were:

- Operating margin, a profitability ratio,
- Return on Equity, a profitability ratio,
- Return on Total Assets, a profitability ratio,
- Quick Ratio, a liquidity ratio,
- Equity to Capital, a solvency ratio,
- Interest Coverage, a solvency ratio, and
- Receivables Turnover, an efficiency ratio.

The ratios were calculated based on information from the companies' annual reports. The dataset consisted of 77 companies and 7 regional averages. The companies were chosen from Pulp and Paper International's annual ranking of pulp and paper companies according to net sales [12]. In total, the dataset consisted of 474 rows of data.

3.2. Choosing the Best Maps

The two datasets were standardized according to different methods. In [2] the authors used the standard deviations of each variable to standardize the data (Equations 1, 2), while in [3] the data have been scaled using histogram equalization [4]. It is not our intention to describe different methods for the standardization of datasets; however, in the literature there are examples of both standardization techniques used on similar datasets.

$$x_i = \frac{\sum_{j=1}^{n} x_{ij}}{n} \qquad \text{[Eq. 1]}$$

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2}{n}} \qquad \text{[Eq. 2]}$$

We have trained different maps with different parameters. As is stated in [2] a "good" map is obtained after several different training sessions. Best maps have been chosen based on two measures: one objective measure (the quantization error) and a subjective measure (ease of readability). However, the algorithm quantization error seems to be positively correlated with the dimension of the maps, while ease of readability is negatively correlated. In other words, we can obtain very "good" maps in terms of their quantization error if we use large dimension parameters, while they are poor in terms of readability. Cluster analysis is often a trade-off between accuracy and cluster clarity and manageability, by creating small maps we force the data into larger clusters. Consequently, when we compared the maps we restricted the maps' dimensions to be constant. The chosen maps and their clusters are presented in Figure 1.

3.3. Identifying the Clusters

We identify the clusters on the maps by studying the final U-matrix maps (Figure 1), the feature planes, and at the same time, by looking at the row data. Actually, the title of this paragraph, "identifying the clusters", should be "identifying the clusters of clusters". What we are saying is that we already have the clusters identified by SOM on the map (from now on we will refer to these clusters as row clusters). For example, in case we are using a 7x5 map, we have 35 row clusters. Next we have to identify the "real" clusters by grouping the row clusters. SOM helps us in this respect by drawing darker lines between two clusters that are "far" from each other (in terms of the Euclidean distance). The results for both datasets were

very similar in terms of the amount, and characteristics, of clusters (7 in each case).

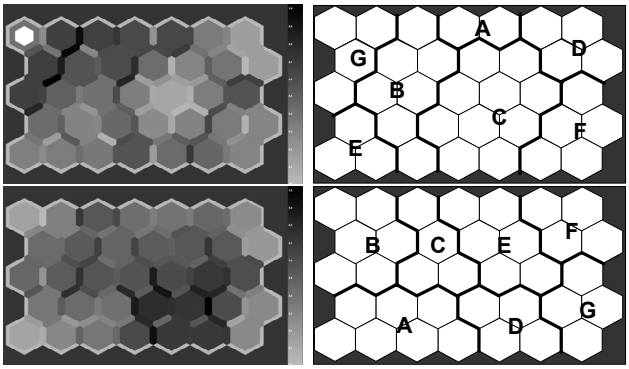

Figure 2. (a) The final U-matrix maps and (b) identified clusters on the maps for the EconomicPerf and FinancialPerf data sets

3.4. Defining the Outcome Values for each Row Data

Roughly speaking, we can state that the outcome values (the classes) in terms of economic/financial performance, were the same in both cases (Figure 1), so the classes are as follows:

A – best performance,
B – slightly below best performance,
C – slightly above average performance,
D – average,
E – slightly below average performance,
F – slightly above poorest performance, and
G – poorest performance.

Defining the outcome values for each data row is a straightforward process. Once we figure out which cluster each row cluster belongs to, the next step is to check which row data vectors are associated with each row cluster, and to associate the class code with those vectors. Consequently, in terms of methodology, we can divide the clustering process into two parts:

- creating the row-clusters – this part is entirely done by the SOM algorithm, the output being the U-matrix;
- creating the "real" clusters – this part is done by the map reader with the help of the SOM algorithm in terms of visualization characteristics.

This kind of multi-level clustering approach is not new. A two-level SOM clustering approach has been suggested before, in [16]. There, the row-clusters are "protoclusters" and our "real" clusters are the "actual" clusters. However, sometimes it is difficult to find good "real" clusters since the second part of the clustering process is highly subjective. Also, the standardization method has an important role, since for different standardization techniques we obtain different maps in terms of quantization error and ease of readability.

4. Applying multinomial logistic regression

In general, when multinomial logistic regression is applied as a predictive modeling technique for classification, there are some steps that have to be followed:

1. Check the requirements regarding the data sample: size, missing data, etc.,
2. Compute the multinomial logistic regression using an available software program (e.g. SPSS),
3. Assess the model fit (accuracy),
4. Interpret the results, and
5. Validate the model.

Below, we follow this methodology when applying logistic regression on our datasets.

4.1. Requirements

In the EconomicPerf dataset, the problem of missing data was overcome by using monthly means for each year. Averages were also used for missing data in the FinancialPerf dataset. The requirement of size, 15-20 cases for each independent variable, was exceeded for each dataset.

4.2. Computing the Multinomial Regression Model

We use SPSS to perform multinomial regression analysis selecting as dependent variables the class variables and as covariates the variables presented in Section 3.1.

4.3. Assessing the Model Fit

From the "Model Fitting information" output table of SPSS we observe that the chi-square value has a significance of < 0.0001, so we state that there is a strong relationship between dependent and independent variables (see Table 2). Next, we study the "Pseudo R-Square" table in SPSS, which also indicates the strength between dependent and independent variables. A good model fit is indicated by higher values. We will base our analysis on the Nagelkerke R^2 indicator (see Table 2). According to this, 74.5% for the EconomicPerf dataset and 97.8% for the FinancialPerf dataset, of the output variation can be explained by variations in input variables. Consequently, we would appreciate the relationships as very strong.

To evaluate the accuracy of the model, we compute the proportional by chance accuracy rate and the maximum by chance accuracy rate. The proportional chance criterion for assessing model fit is calculated by summing the squared proportion of each group in the sample, and the maximum chance criterion is the proportion of cases in the largest

group. We obtained the following indicators (Table 1):

Table 1. Evaluate the model's accuracy

	Model	Proportional by chance criterion	Maximum by chance criterion
EconomicPerf	61,3%	29,92%	49,8%
FinancialPerf	88%	15,62%	20,46%

We interpret these numbers as follows: for example, in the case of the EconomicPerf dataset, based on the requirement that the model accuracy should be 25% better than the chance criteria [5, p. 89-90], the standard to use for comparing the model's accuracy is 1.25 x 0.2992 = 0.374. Our model accuracy rate of 61.3% exceeds this standard. The maximum chance criterion accuracy rate is 49.8% for this dataset. Based on the requirement that model accuracy should be 25% better than the chance criteria, the standard to use for comparing the model's accuracy is 1.25 x 49.8% = 62.22%. Our model accuracy rate of 61.3% is slightly below this standard. The FinancialPerf dataset accuracy rate exceeds both standards.

4.4. Interpreting the Results

To interpret the results of our analysis, we study the "Likelihood Ratio Test" and "Parameter Estimates" outputs of SPSS. We find that the independent variables are all significant, in other words they contribute significantly to explaining differences in performance classification (for both datasets). However, not all variables play an important role in all regression equations (e.g. for the first regression equation, "CurrencyValue" is not statistically significant 0,125 > p = 0,05). Next, we can determine the direction of the relationship and the contribution to performance classification of each independent variable by looking at columns "B" and "exp(B)" from the "Parameter Estimates" output of SPSS. For example, a higher industrial output rate increases the likelihood that the country will be classified as a best country (B = +24,027) and decreases the likelihood that the country will be classified among the poorest countries (B = -11,137). It seems that the results for the EconomicPerf dataset are poorer, in the sense that for the FinancialPerf dataset we have more coefficients estimates that are statistically significant. For example, if we study the "Parameter Estimates" outputs of SPSS ("Sig." column), we find that EconomicPerf dataset has 33% significant coefficients, while FinancialPerf dataset has 62.5%.

4.5. Validating the Model

In order to validate the model, we split the datasets in two parts of, approximately, the same length. Our findings are illustrated in Table 2:

Table 2. Datasets' accuracy rates and accuracy rates estimators when applying multinomial logistic regression

		Main dataset	Part1 (split=0)	Part2 (spli=1)
EconomicPerf	Model Chi-Square (p < 0,0001)	291,420	200,779	136,852
	Nagelkerke R^2	0,745	0,855	0,721
	Learning Sample	61,3%	67%	58,4%
	Test Sample	no test sample	57,6%	67,1%
	Significant coefficients (p<0,05)	ALL	ALL except: CURRENCY[1]	ALL
FinancialPerf	Model Chi-Square (p < 0,0001)	1479,72	792,06	752,85
	Nagelkerke R^2	0,978	0,986	0,981
	Learning Sample	88%	89%	89,5%
	Test Sample	no test sample	76,1%	82,4%
	Significant coefficients (p<0,001)	ALL	ALL	ALL

With one exception, we obtained significant coefficients for the logistic regression equations. In both cases, the accuracy rates of the two split datasets were close to the accuracy rate of the entire dataset. For example, 89% and 89,5% are close to the entire FinacialPerf dataset accuracy rate of 88%. Again, the second dataset outperformed the first one, in the sense that for the FinancialPerf dataset, the accuracy rates for the test samples are closer to the learning sample accuracy rate. However, more investigations should be done to find problems that arise due to insignificant coefficients of each regression equation. Large standard errors for "B" coefficients can be caused by multicollinearity among independent variables, which is not directly handled by SPSS or other statistical packages. Moreover, the problem of outliers and variable selection should be carefully addressed. Also, the discrepancies between learning and test accuracy rates can arise due to the small sizes of the datasets. The larger the dataset is, the better the chance that we have correctly clustered data and, consequently, correct outcome values for each data row. We construct the outcome values based on SOM clustering. There is, of course, a chance that there are misclustered data, which can affect the accuracy of the model.

4.6. Predicting the Classes

The finished model was then used to test the classification of three new data rows for the FinancialPerf

[1] this coefficients is significant for p < 0,153.

dataset. These consisted of data for three Finnish pulp and paper companies: M-Real (no. 3), Stora Enso (no. 4), and UPM-Kymmene (no. 5), for the year 2001. These were used since they were among the first to publish their financial results. The results are illustrated in Table 3.

Table 3. Predictions using multinomial logistic regression

Operating Margin	ROE	ROTA	Equity to Capital	Quick Ratio	Interest Coverage	Receivables Turnover	Company no.	**Predicted Cluster**
5.621597	17.75955	8.979317	27.02372	0.857129	2.314056	6.8226657	3	**D**
11.0069	15.31568	7.67552	31.23215	0.830754	4.189956	6.2295596	4	**B**
16.27344	22.78149	11.16978	34.59247	0.629825	5.205047	6.0291793	5	**A**

5. Applying the Decision Tree Algorithm

For comparison reasons, a See5 decision tree builder system was applied on both datasets. The system was developed by a research team headed by Quinlan. The algorithm behind the program is based on one of the most popular decision tree algorithms, and was developed in the late 70's, also by Quinlan: ID3 [11]. The main idea is that, at each step, the algorithm tries to select a variable and a value associated with it that discriminate "best" the dataset, and does this recursively for each subset until all the cases from all subsets belong to a certain class. The method is called "Top-Down Induction Of Decision Trees (TDIDT)" and C4.5, C5.0/See5 represent different implementations of this method. The "best" discriminating pair (variable-value) is chosen based on so-called "*gain ratio*" criterion:

gain ratio(X) = gain(X) / split info(X) [Eq. 3]

where gain(X) means the information gained by splitting the data using the test X and:

$$\text{split info (X)} = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \log_2\left(\frac{|S_i|}{|S|}\right) \quad \text{[Eq. 4]}$$

represents the potential information generated by dividing S into n subsets. The See5 system implements these formulas along with some other features that are described in [11] and on the web page http://www.rulequest.com/see5-info.html.

5.1. Computing the Decision Tree

For both datasets, we performed three runs of the See5 software, exactly like we did when applying logistic regression: one for the whole dataset, another using first split dataset ("split=0"), and the other using the second half of data ("split=1"). When validating the entire dataset accuracy rate, we have used cross-validation, while when validating one split dataset accuracy rate we have used the other one as test sample. The results are summarized in Table 4.

The first line, for each dataset, represents the accuracy rates obtained using training datasets. The next two lines show us the validation accuracy rates calculated as follows: for the main dataset a 10-crossvalidation was conducted (64% being the average accuracy rate of 10 decision trees), for the "split=0" dataset we used "split=1" as test dataset (46,9% is the accuracy rate on the second dataset, based on the decision tree built with the first dataset), and the last accuracy rate was calculated by considering "split=1" as the training dataset and "split=0" as the test dataset (changing the roles).

Table 4. Dataset accuracy rates and accuracy rates estimators when applying decision tree algorithm

		Main dataset	Part1	Part2
EconomicPerf	Learning Sample	79,1%	77,7%	78,86%
	Test Sample	no test sample	46,9%	54,5%
	cross-validation	64%	no cross-validation	no cross-validation
FinancialPerf	Learning Sample	84,8%	86,5%	86,5%
	Test Sample	74,6%	71,7%	76,8%
	cross-validation	74,4%	no cross-validation	no cross-validation

When constructing the trees, we kept the two most important parameters constant: m = 5, which measures the minimum number of cases each leaf-node should have, and c = 25% (default value) that is a confidence factor used in pruning the tree.

5.2. Assessing the Model Fit

For the EconomicPerf dataset, it seems that our trees were not consistent due to poor accuracy rates and big discrepancies between learning and test accuracy rates, so further comparison with regression analysis cannot be performed in this case. There is at least a 10% difference between the accuracy rates for each split dataset used.

For the FinancialPerf dataset, the differences between accuracy rates are smaller. Therefore, we used this dataset for further investigation. The chosen decision tree is presented in the Appendix. Reading it we can state that the main attribute used to discriminate the data was ROE. The lower that we go down in the decision tree, the less important the attributes become. At each step the algorithm calculates the information gain for each attribute choosing the split attribute with the largest information gain – we call it *the most important* attribute.

5.3. Interpreting the Results

As we can see from the decision tree (Appendix), the second most important variable depends upon the values of ROE: if our ROE is greater than or equal to 10.71424, it is Equity to Capital, while if ROE is less than or equal to 9.179343, it is Receivables Turnover. We must note that we have used fuzzy thresholds, which allows for a much more flexible decision tree: the algorithm (C5.0) assigns a lower value (*lv*) and an upper value (*uv*) for each attribute chosen to split the data. Then a membership function (trapezoidal) is used to decide which branch of the tree will be followed when a new case has to be classified. If the value of the splitting attribute for the new case is lower than *lv*, the left branch will be followed, and if it is greater than *uv* then we will further use the right branch. If the value lies between *lv* and *uv*, **both** branches of the tree are investigated and the results combined probabilistically – the branch with the highest probability will be followed.

5.4. Validating the Model

Notice the asymmetric threshold values for almost every splitting attribute. In this case (FinancialPerf), the accuracy rate of the test sample is comparable with the accuracy rate of the learning sample. There is no specification on how close these two values should be; consequently, we conclude that the tree is validated. The only way to "really" validate the assumption that the two accuracy rates are "not far" from one another is to consider the two accuracy rates as random variables and then use a statistic test to see if their means differ significantly. This new step in validating the decision tree model would require splitting the dataset in different ways to obtain different training and test datasets, and then, under the assumption that the accuracy rates are random variables that follow normal distribution, which is not always the case, we would test if their means are or are not statistically different.

After training the decision tree, we tested it on the same data rows used in Section 4.

5.5. Predicting the Classes

The results are illustrated in Table 5. As can be seen in the table, the results are somewhat different from those obtained using logistic regression.

Table 5. .Prediction using the decision tree

Operating Margin	ROE	ROTA	Equity to Capital	Quick Ratio	Interest Coverage	Receivables Turnover	Company no.	**Predicted Cluster**
5.621597	17.75955	8.979317	27.02372	0.857129	2.314056	6.8226657	3	**B**
11.0069	15.31568	7.67552	31.23215	0.830754	4.189956	6.2295596	4	**B**
16.27344	22.78149	11.16978	34.59247	0.629825	5.205047	6.0291793	5	**A**

M-Real (no.3) was classified as a D company in Table 3, while it is a B company in table 5. The data rows of Stora Enso and M-real are generally similar, but the decision tree has placed more emphasis on ROE, while logistic regression seems to have emphasized Equity to Capital. Also, we can see from Table 6 that the decision tree has not quite correctly learned the pattern associated with Group D, only being able to correctly classify 58% of the cases in this group. The logistical regression model was much more successful, and we therefore consider its prediction the more reliable of the two. More study will be needed to judge why this happened.

6. Comparing the Classification Models' Accuracy

While this is not the only way to compare two classification techniques, comparing them using accuracy rates is the most used. In [10] the author compared five predictive models from areas of both machine learning and statistics. A comparison similar to ours was made in [13]. The authors compared logistic regression and decision tree induction in the diagnosis of Carpal Tunnel syndrome. Their findings claim that there is no significant difference between the two methods in terms of model accuracy rates. Also, they suggest that the classification accuracy of the bivariate models (two independent variables) is slightly higher than that of multivariate ones. It is not our goal to compare bivariate and multivariate models, while this can be a subject for further investigations using the datasets presented in this paper.

As we stated in section 5, we will consider only the second dataset when comparing the two methods, since for the first dataset the results were very poor in terms of the accuracy rate. In the last section, we will try to explain why we obtained such poor results using the EconomicPerf dataset.

Conversely, in the case of the second dataset (FinancialPerf) both logistic regression and decision tree models were validated against the split datasets. The differences between accuracy rates were smaller in this case, and the learning dataset accuracy rates were very good (88% and 84,8%). Also, both models performed similarly on the test datasets (89%, 89,5% and 86,5%, 86,5%). The bigger difference for the training datasets could be caused by the fact that when applying the decision tree algorithm, we split the data in two parts using 75% of the rows for the learning dataset. The remaining 25% was used as a test dataset. This was due to a number-of-rows restriction in the See5 demo-software (max 400 rows of data). Using logistic regression, changes in accuracy rates can occur when including/excluding some variables in/from the model. In the case of the decision tree, the accuracy rate of the model can be tuned using model parameters, e.g. the minimum number of

cases in each leaf (m) or the pruning confidence factor (c). The accuracy rates for the two methods are illustrated in Table 6.

Table 6. The observed accuracy rates of the two methods

Logistic Regression

Predicted \ Observed	a	b	c	d	e	f	g
a	88%	6%	2%	4%			
b	5%	89%	3%	2%			
c	6%	6%	77%	4%	4%	2%	
d		6%	2%	84%	8%		
e			7%	1%	88%	4%	
f					11%	89%	
g					3%		97%

Decision Tree

Predicted \ Observed	a	b	c	d	e	f	g
a	86%	10%		4%			
b	4%	87%	5%	1%			3%
c	3%	8%	76%	5%	8%		
d	0%	18%	6%	58%	12%		6%
e			2%		93%	2%	4%
f					3%	94%	3%
g				2%	4%	4%	90%

7. Discussion and conclusions

In this study, we have proposed a new two-level approach for making class predictions about countries'/companies' economic/financial performance. We have applied our methodology on two datasets: the EconomicPerf dataset that includes variables describing the economic performance of central-east European countries during 1993-2000, and the FinancialPerf dataset, which includes financial ratios describing the financial performance of international pulp and paper companies during 1995-2000. Firstly, SOM clustering was applied on both datasets in order to identify clusters in terms of economic/financial performance, and the optimal number of clusters to consider. By reading the SOM output (U-matrix maps), we have considered seven to be the most appropriate number of clusters for both datasets. Consequently, we construct the outcome values for each data row based on the SOM maps and the corresponding seven classes: best, slightly below best, slightly above average, average, slightly below average, slightly above poor, and poorest. Secondly, based on the new datasets (updated with the outcome values), we have predicted to which class a new input belongs. We chose and compared two predictive models for classification: logistic regression and decision tree induction.

Why is this approach important? Why combine clustering and classification techniques? Why not directly construct the outcome values and apply the predictive models without performing any clustering? We could perform surveys, asking experts how their company/country performed in different months or years, and then directly apply the classification technique to develop prediction models as new cases are to be classified. First of all, this kind of information (outcome values for each data row) is not easy to get (is costly), and secondly, even if we have it, in order for it to be useful, it has to be "true" and "comparable". What we mean by "true" is that when performing surveys, the respondents can be subjective, giving higher rankings for their country/company (not giving true answers). The outcome values can be un-"comparable" if, for example, one person has different criteria for the term "best performance" than another. In the best perspective, when answering our questions about their country/company performances the respondents would, most probably, classify their country/company using their knowledge and internal aggregate information. We think our methodology is an objective way of making class predictions about countries'/companies' performances since, using it, we can choose the correct number of clusters, define the outcome values for each data row, and construct the predictive model. Also, the problem of inserting new data into an existing model is solved using this method. The problem is that we normally have to train new maps every time, or standardize the new data according to the variance of the old dataset, in order to add new labels to the maps. Inserting new data into an existing SOM model becomes a problem when the data have been standardized, for example, within an interval like [0,1]. Also, the retraining of maps requires considerable time and expertise. We propose that our methodology solves these problems associated with adding new data to an existing SOM cluster model.

The results show that our methodology can be successful, if it is correctly implemented. Clustering is very important in our methodology, since we define the outcome values for each data row based on it. Our U-matrix maps clearly show seven identifiable clusters. More investigations should be performed on finding the utility of each clustering or, in other words, define "how well" we clustered the data. To evaluate the maps we used two criteria: the average quantization error and the ease-of-readability of each map. As a further research problem, we would try to develop a new measure, or use an existing one, to validate the clustering. When applying logistic regression, we obtained models with acceptable accuracy rates. All the coefficients of all regression equations were statistically significant except one (CURRENCY for the

EconomicPerf dataset). The accuracy rates were evaluated using two criteria: proportional by chance criterion and maximum by chance criterion. The first dataset's accuracy rate didn't satisfy the second criterion. When comparing the two classification techniques, we therefore only took into consideration the results of the second. However, like in [13] our findings claim that the results of the two classification techniques are similar in terms of accuracy rate. Also, when making predictions using the two models, we used data for the FinancialPerf dataset from year 2001. Two out of three new data rows were classified in the same class using both predictive models (Stora Enso and UPM-Kymmene to classes 2 and 1 respectively).

An improvement to our methodology would be to tackle the problem of variable selection for both the clustering and the classification phases, finding a new way to measure clustering utility, and generalizing the methodology. As further research, we will investigate different methods of improving our classification models.

Acknowledgements

The authors would like to thank Professor Barbro Back for her constructive comments on the article.

References

[1] B. Back, K. Sere, and H. Vanharanta, "Managing Complexity in Large Data Bases Using Self-Organizing Maps", *Accounting Management and Information Technologies 8 (4)*, Elsevier Science Ltd, Oxford, 1998, pp. 191-210.

[2] A. Costea, A. Kloptchenko, and B. Back, "Analyzing Economical Performance of Central-East-European Countries Using Neural Networks and Cluster Analysis", in *Proceedings of the Fifth International Symposium on Economic Informatics*, I. Ivan. and I. Rosca (eds), Bucharest, Romania, May, 2001, pp. 1006-1011.

[3] T. Eklund, B. Back, H. Vanharanta, and A. Visa, "Assessing the Feasibility of Self-Organizing Maps for Data Mining Financial Information", in *Proceedings of the Xth European Conference on Information Systems (ECIS 2002)*, Gdansk, Poland, June 6-8, 2002, pp. 528-537.

[4] J. F. Hair, Jr, R. Anderson, and R. L. Tatham, Multivariate Data Analysis with readings. Second Edition. Macmillan Publishing Company, New York, New York, 1987.

[5] D. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, Cambridge, 2001.

[6] S. Kaski and T. Kohonen, "Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World", in *Neural Networks in Financial Engineering*, N. Apostolos, N. Refenes, Y. Abu-Mostafa, J. Moody, and A. Weigend. (Eds), World Scientific, Singapore, 1996, pp. 498-507.

[7] J. P. Guiver and C. C. Klimasauskas, "Applying Neural Networks, Part IV: Improving Performance", *PC/AI Magazine 5 (4)*, Phoenix, Arizona, 1991, pp. 34-41.

[8] T. Kohonen, *Self-Organizing Maps*, 2nd edition, Springer-Verlag, Heidelberg, 1997.

[9] J. Lehtinen, *Financial Ratios in an International Comparison*, Acta Wasaensia 49, Vasa, 1996.

[10] J. R. Quinlan, "A Case Study in Machine Learning", in *Proceedings of ACSC-16 Sixteenth Australian Computer Science Conference*, Brisbane, Jan. 1993, pp. 731-737.

[11] J. R. Quinlan, *C4.5 Programs for Machine Learning*, Morgan Kaufmann Series in Machine Learning, Morgan Kaufmann Publishers, San Mateo, 1993.

[12] J. Rhiannon, C. Jewitt, L. Galasso, and G. Fortemps, "Consolidation Changes the Shape of the Top 150", *Pulp and Paper International 43 (9)*, Paperloop, San Francisco, California, 2001, pp. 31-41.

[13] S. Rudolfer, G. Paliouras, and I. Peers, "A Comparison of Logistic Regression to Decision Tree Induction in the Diagnosis of Carpal Tunnel Syndrome", *Computers and Biomedical Research 32*, Academic Press, 1999, 391-414

[14] A. Ultsch, "Self organized feature planes for monitoring and knowledge acquisition of a chemical process", in *Proceedings of the International Conference on Artificial Neural Networks*, Springer-Verlag, London, 1993, pp. 84-867.

[15] J. Vesanto "Neural Network Tool for Data Mining: SOM Toolbox", in *Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000)*, Oulun yliopistopaino, Oulu, Finland, 2000, pp. 184-196.

[16] J. Vesanto and E. Alhoniemi, "Clustering of the Self-Organizing Map", *IEEE Transactions on Neural Networks 11 (3)*, IEEE Neural Networks Society, Piscataway, New Jersey, 2000, pp. 586-600.

[17] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Academic Press, San Diego, 2000.

Appendix: the decision tree