ISSN 1392 – 124X INFORMACINĖS TECHNOLOGIJOS IR VALDYMAS, 2004, Nr. 4(33)

FORECASTING OF CREDIT CLASSES WITH THE SELF-ORGANIZING MAPS

Egidijus Merkevičius, Gintautas Garšva

Department of Informatics, Kaunas Faculty of Humanities, Vilnius University
Muitinės st. 8, LT- 44280 Kaunas, Lithuania

Rimvydas Simutis

Process Control Department, Kaunas University of Technology
Studentų st. 48-327, LT- 51367 Kaunas, Lithuania

Abstract. To determine the credit classes statistical and artificial intelligence methods have been used often recently. Particularly the artificial neural networks there have been often applied, one of them is a self-organizing map (SOM). SOM is a two-dimensional map of the credit units that is generated by similar characteristics (attributes) of the process. However this process is not specified by network outputs. If the credit units of one class dominate in the clusters it is valuable such SOMs to employ to forecast the new credit classes. In this paper we investigate the capabilities of SOM in forecasting of credit classes. We present the results of our investigations and show that SOM may distinctly reduce misclassification errors. On the other hand, we demonstrate the possibility of SOM to identify how dataset is liable for the valuable map generation.

Indroduction

The forecasting of the credit state has always been a relevant task in the finance market. Available algorithms of statistical and artificial intelligence methods, especially the combinations of them, adduce more and more accurate predictable results. Recently the algorithms of supervised and unsupervised artificial neural networks have been employed for determination of credit classes [9,16]. Self-organizing map (SOM) is unsupervised learning artificial neural network that is generated without defining of network output values. As a result of SOM-learning procedure the two-dimensional map of clusters is created, which represents the credit units by similar characteristics.The purpose of this paper is to investigate the capabilities of SOM in forecasting of credit classes.

Martin-del-Prio and Serrano-Cinca have been one of the first who applied SOM to the financial analysis. They generated SOMs of the Spanish banks and subdivided them into two large groups and this allowed establishing root causes of the banking crisis [14].

Having analyzed the state of Russian banks Shumsky and Yarovoy [18] subdivided them with the help of SOM into several groups and compared how banks migrated on the map within few years. Based on these results a trend for the further state of the banks was appointed. Similar investigations were completed with the data of Russian companies [19].

Kiviluoto [9] made a SOM-map which included 1137 companies, 304 companies from them were crashed. The created SOM was able to give useful qualitative information about similar input vectors. Visual exploration allowed to see the distribution of important indicator – bankrupt – on the map, thus, it was possibly to apply that map for the forecasting of companies bankrupt.

The mentioned authors have estimated only a current situation of credit state and afterwards they have interpreted it for forecasting bankrupt, causes of crisis period or market segmentation of banks. In this paper, we propose to generate the SOM which could be applied for forecasting of credit classes for new customers.

The first section of paper describes the core steps of standard SOM algorithm. In the second part we present the results of our investigations with real credit data and propose several recommendations for generating valuable SOM employing to credit class forecast.

1. Algorithm of SOM

In the self-organizing procedure the output data are configured for visualization of topologic original data

[3, 11, 12]. The learning of SOM is based on competitive learning algorithm („winner takes all"). The algorithm of standard stochastic SOM learning is based on 6 core steps:

Step 1. SOM-Weights are initialized.

Step 2. Learning data vector is represented as a grid. It's accomplished as:

- Random arrangement;
- By principal input components;
- By defining a large, enough hyper cube to cover all the training patterns [20].

Step 3. Each node is calculated to assess the best matching unit (BMU). One of the methods is to calculate Euclid's distance between each weight vector and input vector:

$$Dist = \sqrt{\sum_{i=0}^{i=n}(X_i - Y_i)^2}\ , \qquad (1)$$

where

X_i input vector to the node i;

Y_i - weights vector.

Step 4. It is calculated BMU radius of neighbourhood.

Step 5. Each neighbourhood weight is pitched to the input vector.

$$(y_j)' = y_j + h(r(y_j, y_{BMU}), t)(X_i - y_j), j = 1...p\ , \qquad (2)$$

where

y – node;

p – number of nodes;

$h(x,t)$ – neighbourhood function;

$r(y_1, y_2)$ – distance between nodes y_1 and y_2 in the grid;

t – iteration;

Neighbourhood function $h(x,t)$ is taken by assuming to maximum when $x = 0$. One of the popular functions is Gaussian:

$$h(x,t) = \alpha(t)\exp(-\frac{1}{2}\frac{x^2}{\sigma^2(t)})\ , \qquad (3)$$

where

$\alpha(t)$ – learning rate;

$\sigma(t)$ – neighbourhood width;

In addition, Bubble-function is used:

$$h(x,t) = \begin{cases}\alpha(t), x \le \sigma(t)\\ 0, x > \sigma(t)\end{cases}\ , \qquad (4)$$

where

$\alpha(t)$ - learning rate;

$\sigma(t)$ - Neighbourhood width;

Step 6. 1-5 steps are repeated while t iterations will been accomplished.

This described algorithm is realized in various software with refined capabilities [3]. In our investigations we chose Viscovery@SOMine software (VS), which allows choosing flexibly the parameters of the learning procedure [4].

2. Experiment

In this paper possibilities of SOM is studied by means of two real credit datasets (Australian credit approval (ACA) and German credit database (GCD)) taken from the public UCI Repository of Machine Learning Databases [15]. The used Viscovery@SOMine software has a suitable and simple user interface, large possibilities of data pre- and post processing, fast learning rate and comprehensive visualization and monitoring tools [4].

2.1. Australian credit approval dataset

This dataset is a database of credit cards. Data amount is 690 records, classes are two: 0, 1 (-, +). The distribution of classes:

1 class „bad"credits (1): 307 (44.5%);

2 class „good"credits (0): 383 (55.5%);

The attributes of records are continuous and categorical (see Table 1).

Table 1. The attributes of the ACA dataset

A1:	0,1	Categorical
A2:	Continuous	
A3:	Continuous	
A4:	1,2,3	Categorical
A5:	1,2,3,4,5, 6,7,8,9,10,11,12,13,14	Categorical
A6:	1,2,3, 4,5,6,7,8,9	Categorical
A7:	Continuous	
A8:	1, 0	Categorical
A9:	1, 0	Categorical
A10:	Continuous	
A11:	1, 0	Categorical
A12:	1, 2, 3	Categorical
A13:	Continuous	
A14:	Continuous	
A15	0,1	Categorical (value)

In this dataset the names of attributes are replaced with symbolic values.

In the experiment the data are grouped into training and test data with ratio 80 in 20. SOM is trained and tested by five-fold cross validation principle, i.e. ACA data are grouped five times randomly and five samples are trained and tested. For the network training all 14 attributes are used with similar relevance, however A15 attribute is employed for network reliability and for classification of credit unit (class „good" or „bad")

Figure 1 presents a sample of generated SOM.

The estimating of network training reliability shows how values of A15 attribute distribute in SOM

clusters. Table 2 presents a sample of training evaluation.

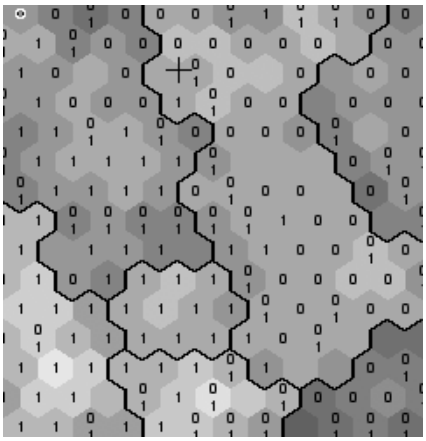

Figure 1. A sample of generated SOM

Table 2. Values distribution of A15 attribute on the SOM

Clusters	1 ("bad")	0 ("good")	Total:	"bad" (%)	"good" (%)
C 1	36	140	176		79.55%
C 2	81	55	136	59.56%	
C 3	69	2	71	97.18%	
C 4	5	63	68		92.65%
C 5	26	13	39	66.67%	
C 6	11	30	41		73.17%
C 7	21	0	21	100.00 %	
Total:	249	303		**80.85%**	**81.79%**

If in the cluster there are more "good" credits than "bad" credits, then this cluster is considered as "good cluster" and conversely.

The significant indicator in the validation of SOM is misclassification error which is estimated by performance matrix [7].

The sample of SOM performance matrix is presented in Table 3.

Table 3. The performance matrix of ACA dataset

Actual vs Predicted (Performance Matrix)				
	Predicted (by model)			
	0	1	Total	Total Error
Actual 0	233	70	303	0.231023
Actual 1	52	197	249	0.208835
	285	267	1104	0.110507

In the same way we estimate the reliability of the test data.

In Table 4 we present the investigation results by means of ACA dataset. We can see that global misclassification error composes approximately only 10%, thus we presume that SOM is efficient for the forecasting of credit classes because credit classes ("good" or "bad") dominate in the clusters (ascendancy presents over 90%).

On the other hand, we observe ~26-30% misclassification error of "bad" credits. It is explained by means of smaller amount of "bad" credits units in process of SOM training.

Table 4. The results of SOM reliability by ACA dataset

Misclassification error		1 sample	2 sample	3 sample	4 sample	5 sample	Standard deviation	Overall
Training set	552							
"Good" credit		23.10%	8.09%	19.54%	7.79%	18.87%	7.07%	15.48%
"Bad" credit		20.88%	31.69%	21.63%	30.33%	27.20%	4.93%	26.35%
Global error		11.05%	9.24%	10.24%	8.88%	11.32%	1.08%	10.14%
Test set	138							
"Good" credit		15.00%	6.76%	28.95%	12.00%	17.28%	8.24%	13.29%
"Bad" credit		37.93%	23.44%	22.58%	33.33%	33.33%	6.77%	31.90%
Global error		12.31%	7.25%	13.04%	10.87%	11.96%	2.28%	10.61%

2.2. German credit database

This dataset is a database of credit customers. Data amounts 1000 records, classes distribute into two: 0, 1 (-, +). The distribution of classes:

1 class„bad" credits (0): 300 (30%);

2 class „good" credits (1): 700 (70%);

Attributes of records compose 7 continuous and 13 categorical attributes (see Table 5).

In this experiment the data are grouped into training and test data with ratio 80 in 20 by five-fold cross validation principle, furthermore, all 20 attributes are used with similar relevance. In Figure 2 we observe primary generated SOMs.

Visual exploration allows presuming that distribution of A21 attribute is chaotic and Table 6 illustrates this assertion by means of performance matrix.

Table 5. The attributes of GCD dataset

A1:	Status of existing checking account	Categorical
A2:	Duration in month	Continuous
A3:	Credit history	Categorical
A4:	Purpose	Categorical
A5:	Credit amount	Continuous
A6:	Savings account/bonds	Categorical
A7:	Present employment since	Categorical
A8:	Instalment rate in percentage of disposable income	Continuous
A9:	Personal status and sex	Categorical
A10:	Other debtors / guarantors	Categorical
A11:	Present residence since	Continuous
A12:	Property	Categorical
A13:	Age in years	Continuous
A14:	Other instalment plans	Categorical
A15:	Housing	Categorical
A16:	Number of existing credits at this bank	Continuous
A17:	Job	Categorical
A18:	Number of people being liable to provide maintenance for	Continuous
A19:	Telephone	Categorical
A20:	Foreign worker	Categorical
A21:	"Good/bad" credit	Categorical

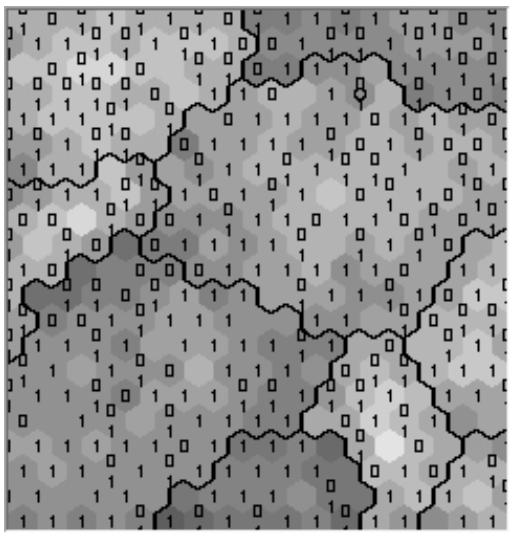

Figure 2. A primary SOM of GCD dataset

We observe that all clusters (100%) are assigned to "good" class, furthermore "bad" credit units are scattered on the all clusters. Under the circumstances this SOM generation is partially non-suitable.

When we enlarge the number of clusters, i.e. some clusters are decomposed into smaller, then we got the following results (see Table 7).

Table 6. A performance matrix of primary SOM generated from GCD

Actual vs Predicted (Performance Matrix)				
	Predicted (by model)			
	0	1	Total	Total Error
Actual 0	0	239	239	1
Actual 1	0	561	561	0
	0	800	1600	0.149375

Table 7. The performance matrix of primary SOM with the enlarged number of clusters

Actual vs Predicted (Performance Matrix)				
	Predicted (by model)			
	0	1	Total	Total Error
Actual 0	21	218	239	0.9121
Actual 1	17	544	561	0.0303
	38	762	1600	0.1469

In effect we get some better results: misclassification errors amount to 91% of „bad" credits and 3% of „good" credits.

For the achievement of better results the reconstruction of GCD dataset is performed:

1. First, normalization of data is completed because the attributes are categorical and continuous. The data normalization is executed by subtracting the means of records and dividing them by the standard deviation [13].
2. The sensitivity of various attributes are determined. For this purpose various back-propagation neural networks for prediction of credit classes were generated and by the means of normalized sum squared errors (NSSE) the sensitivity of each attribute was estimated. Table 8 shows first six sensitive attributes by NSSE and from the Figure 3 the variation of NSSE can be seen: NSSE decrease by composing of attributes.

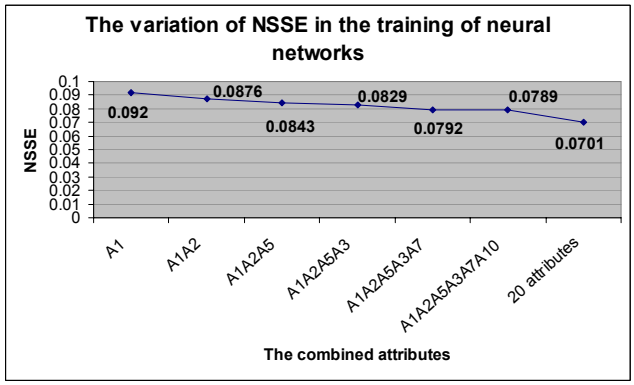

Figure 3. The variation of NSSE using various attributes

3. In pre-processing stage of SOM generating priority factors are assigned to the first six attributes by NSSE sensitivity results (see Table

8). To the rest attributes past value of priority factor (1) is assigned.

Table 8. The sensitivity of GCD dataset attributes

Attribute	Error	Difference of NSSE	Priority factor
A1	0.0920		2
A1**A2**	0.0876	0.0044	1.9
A1A2**A5**	0.0843	0.0033	1.8
A1A2A5**A3**	0.0829	0.0014	1.7
A1A2A5A3**A7**	0.0792	0.0037	1.6
A1A2A5A3A7**A10**	0.0789	0.0002	1.5
20 attributes	0.0701	0.0022	1

Table 9. The performance matrix of generated SOM by means of reconstructed GCD dataset

Actual vs Predicted (Performance Matrix)				
	Predicted (by model)			
	0	1	Total	Total Error
Actual 0	76	202	278	0.7266
Actual 1	45	477	522	0.0862
	121	679	1600	0.1543

The performance matrix of new generated SOM is presented in Table 9. In this sample the number of SOM clusters is 9 and amount of training data is 800. Table 10 presents performance matrix of SOM that is generated by means of test data (200 records).

Table 10. The performance matrix of generated SOM by means of test data taken from the reconstructed GCD dataset

Actual vs Predicted (Performance Matrix)				
	Predicted (by model)			
	0	1	Total	Total Error
Actual 0	5	17	22	0.7727
Actual 1	19	159	178	0.1067
	24	176	400	0.0900

The change over primary and reconstructed SOM generations shows that global error is similar, at the same time misclassification error of "bad" credits is distinctly reduced (approximately 15%).

For giving of more accurate and more satisfactory results we validate investigations by five-fold cross validation principle. The issues are represented in Table 11.

Table 11. The results of SOM reliability by GCD dataset

Misclassification error		1 sample	2 sample	3 sample	4 sample	5 sample	Standard deviation	Overall
Training set	800							
"Bad" credit		72.66%	76.95%	79.06%	82.37%	83.33%	4.31%	78.88%
"Good" credit		8.62%	9.07%	5.16%	5.36%	4.72%	2.08%	6.59%
Global error		15.44%	16.50%	15.38%	16.06%	13.56%	1.12%	15.39%
Test set	200							
"Bad" credit		77.27%	77.78%	91.30%	86.36%	81.94%	5.94%	82.93%
"Good" credit		10.67%	9.94%	6.21%	6.74%	4.69%	2.55%	7.65%
Global error		9.00%	8.04%	8.00%	7.75%	16.25%	3.63%	9.81%

As it can be seen from Table 11, the global misclassification errors represent 15.39% of training data and 9.81% of test data. However, misclassification errors of "bad" credit units compose 78.88% and 82.93%, respectively.

The comparison of investigations let us make several conclusions:

1. A small number of one class data impact on the sizeable misclassification error.
2. The misclassification error of "bad" credit units on the investigation of GCD dataset predicates some assumptions of the following causes:
   a) It could be that the given attributes can not describe the credit class rightly;
   b) In the collection (release, writing) of data may be some operational mistakes.

   According to the results of these investigations, we assume that SOM can identify how dataset is liable for the valuable map generation.
3. The pre- and post-processing possibilities of SOM software can distinctly reduce misclassification errors. It is accomplished by means of clusters scaling, by assess of consuming attributes etc.

3. Concluding remarks

In this paper we investigated the capabilities of SOM in the forecasting of credit classes. We proposed several techniques and recommendations to give better results in SOM generation.

In general we showed that SOM is valuable method for the forecasting of credit classes if data collection have been well-accomplished and attributes of data are correlated with the class of the credit units. In the investigation of ACA dataset global misclassification error represents 10.14% training data and 10.61% test data, also in the investigation of GCD dataset global misclassification error shows 15.39% training

data and 9.81% test data. Although, in the forecasting of „bad“ credit class by means of GCD dataset test data we found 82.93% misclassification error, thus we recommend to assess in the future works the relationship grade between attributes and to discriminate core attributes.

In the future we intend to compare SOM and statistical and other artificial intelligence methods in the forecasting of credit classes and make a new hybrid method in them applying the best features of thees methods.

References

[1] **B. Back, M. Irjala, K. Sere, H. Vanharanta.** Competitive Financial Benchmarking Using Self-Organizing Maps. *Reports on Computer Science & Mathematics, Åbo Akademi, Ser. A, No*. 1995, 169.

[2] **B. Back, K. Sere, H. Vanharanta.** Analyzing Financial Performance with Self-Organizing Maps. *Wsom*'97: *Workshop on Self-organizing maps*, 1997.

[3] **G. Deboeck.** Financial Applications of Self-Organizing Maps. *American Heuristics Electronic Newsletter, Jan*, 1998.

[4] **G. Deboeck, T. Kohonen.** Visual Explorations in Finance with Self-Organizing Maps. *London*, 1998, *Springer Finance*.

[5] **G. Deboeck.** Self-Organizing Maps Facilitate Knowledge Discovery In Finance. *Financial Engineering News, December* 1998.

[6] Viscovery SOMine. Eudaptics software Gmbh. *http://www.eudaptics.at.*

[7] **J. Galindo, P. Tamayo.** Credit Risk Assessment Using Statistical and Machine Learning: Basic Methodology and Risk Modeling Applications. *Computational Economics. April* 2000, *Vol*.15.

[8] **Z. Huanga, H. Chena, C.-J. Hsua, W.-H. Chenb, S. Wuc.** Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, (*accepted* 2003).

[9] **K. Kiviluoto.** Predicting bankruptcies with the self-organizing map. *Neurocomputing*, 21, 1998, 191–201.

[10] **K. Kiviluoto, P.Bergius.** Analyzing Financial Statements with the Self-Organizing Map. *Proceedings of the workshop on self-organizing maps* (*WSOM*'97), (*Espoo, Finland*), *Neural Networks Research Centre, Helsinki University of Tech- nology, June* 1997, 362 – 367.

[11] **T. Kohonen.** The Self-Organizing Map. *Proceedings of the IEEE*, 78, 1464-1480.

[12] **T. Kohonen.** Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.

[13] **Y. Le Cun, I. Kanter, S.A. Solla.** Eigenvalues of Covariance Matrices: Application to Neural-Network Learning. *Physical Review Letters, Vol*.66, *No*.18, 1991, 2396-2399.

[14] **B. Martin-del-Prio, K. Serrano-Cinca.** Self-Organizing Neural Network: The Financial State of Spanish Companies. *Neural Networks in Finance and Investing. Using Artificial Intelligence to Improve Real-World Performance. R.Trippi, E.Turban, Eds. Probus Publishing*, 1993, 341-357

[15] **P.M. Murphy, D.W. Aha.** UCI Repository of machine learning databases. *Department of Information and Computer Science, University of California, Irvine, CA,* *http://www.ics.uci.edu/~mlearn/MLRepository.html*.

[16] **M. Nørgaard.** Neural Network Based System Identification Toolbox Version 2. *Technical Report* 00-E-891, *Department of Automation Technical University of Denmark. 2000.* *http://kalman.iau.dtu.dk/research/control/nnsysid.html*.

[17] **S. Piramuthu.** Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research* 112, 1999, 310-321.

[18] **S.A. Shumsky, A.V. Yarovoy.** Kohonen Atlas of Russian Banks. *G.Deboeck and T.Kohonen(Eds). Visual Explorations in Finance with Self- Organizing Maps. Springer,* 1998.

[19] **S.A. Shumsky, A.N. Kochkin.** Self-organising maps of 200 top Russian companies. *Proceedings of Neuroinformatics*'99. *Moscow*, 1999.

[20] **M.C. Su, T.A. Liu, H.T. Chang.** An Efficient Initialization Scheme for the Self-Organizing Feature Map Algorithm, *Proceedings of the IEEE International Joint Conference in Neural Networks*, 1999.