Self-Organizing Feature Maps in Correlating Groups of Time Series: Experiments with Indicators Describing Entrepreneurship

Marta Czyżewska, Jarosław Szkoła, and Krzysztof Pancerz

University of Information Technology and Management
Sucharskiego Str. 2, 35-225 Rzeszów, Poland
{mczyzewska,jszkola,kpancerz}@wsiz.rzeszow.pl

Abstract. In the paper, we briefly describe a problem of identification of entrepreneurship determinants with respect to economic development of countries. In order to solve this problem, we need to identify correlations between entrepreneurship and macroeconomic indicators. The main attention in the paper is focused on selecting a proper computer tool for solving this problem. As a tool supporting identification, Self-Organizing Feature Maps (SOMs) have been chosen. Some modification of the clustering process using SOMs is proposed by us to improve classification results and efficiency of the learning process. At the end, we indicate some challenges of further research.

Keywords: self-organizing feature maps, correlation, entrepreneurship

1 Motivation

The phenomenon of entrepreneurship is the subject of various levels of observations such as the entrepreneur, industry, region or nation with respect to many aspects reflecting the entrepreneurship level. The worldwide interest in the entrepreneurship especially innovative entrepreneurship based on advanced knowledge and technology shows the importance of the phenomenon particularly for underdeveloped countries, for nations of aging societies, for those with youth unemployment growing rate, and for the global economy as well. Our research concerns designing effective methods for computer support of identification of entrepreneurship determinants as it is the key factor of countries economic development. Therefore, in our research, we are going to build a specialized computer aided system based on applying neural networks to determine the cross-countries differences referring to propensity for entrepreneurship and the country framework in order to assess policy gaps and opportunities for future actions. The multidimensional analysis enables us to form specific recommendations to a country government on how to lead a policy toward entrepreneurship development. The question how to increase the development level by the entrepreneurship stimulation policy is still open. Building an effective and boosting entrepreneurship system is challenge of the century, see e.g. [3], [5].

2 The Clustering Procedure using Self Organizing Feature Maps

The concept of a Self-Organizing Feature Map (SOM) was originally developed by T. Kohonen [6]. SOMs are neural networks composed of a two-dimensional grid (matrix) of artificial neurons that attempt to show high-dimensional data in a low-dimensional structure. Each neuron is equipped with modifiable connections.

In this section, we describe a clustering procedure used in experiments for finding correlations of groups of multidimensional objects using Self Organizing Feature Maps. We propose some modification to improve classification results and efficiency of the learning process, among others:

- a modified coefficient for adjusting weights,
- a modified way for adjusting weights of neighboring neurons (the modification coefficient is not constant, but it decreases along with the distance from the pattern neuron),
- a modified way of the learning process (only neighboring neurons of the pattern neuron for a given input vector are trained).

Input for the procedure is a matrix of real numbers. Each row of the matrix represents a feature vector of one object (corresponding to one country) subjected to clustering. All rows (feature vectors) have the same dimension. An input matrix must have at least two rows (feature vectors). A fragment of exemplary data (for the indicator "New business density") subjected to clustering is shown in Table 1.

Table 1. A fragment of exemplary data subjected to clustering.

Country / Year	2004	2005	2006	2007	2008	2009
Algeria	0.53	0.48	0.40	0.35	0.48	0.44
Argentina	0.56	0.55	0.61	0.62	0.57	0.46
Austria	0.60	0.65	0.68	0.66	0.65	0.58
...	...	...	...	...	...	...

Output for the procedure is a matrix of the size $n \times n$. The parameter n is determined as:

$$n = ceil(\sqrt{2m} + 0.5),$$

where *ceil* is a function rounding up elements and m is a number of feature vectors. An initial size of the output matrix is 2×2. This size is increased, during a learning process, up to $n \times n$. A learning process is performed iteratively. In our research, a number of iterations has been set as 100. More iterations did not improve a quality of classification. Each feature vector is associated with an individual map. A map represents a matrix of neurons. We have as many maps as many feature vectors is present. We will treat this set of maps as a multilayer

map labeled with M. An initial value of weights of the map is set on the basis of the following formula:

$$M[x][y][i] = \frac{random(min, max)}{10},$$

where x and y determine a position in the map and i is the index of the feature vector, i is integer included in the interval $[1, k]$, where k is a number of all feature vectors subjected to clustering, $random(min, max)$ is a pseudorandom-number generator returning a number from the interval $[min, max]$, where min and max determine minimal and maximal values of input feature vectors. A learning process includes the following steps:

1. Calculating a current coefficient for modification of weights of the map.
2. Calculating a new desired size of the map.
3. Random selection of the order of feature vectors for training the network.
4. Modification of weights of the map after calculation of the error on the basis of an input feature vector and current weights of the map.

Steps from 1 to 4 are performed iteratively up to the fixed number of iterations (in our case, 100). After finishing the learning process, the testing process is run. In this process, assessment of classification results is made for each input feature vector used in the learning process. Assessment consists in calculation differences between a given feature vector and weights of all neurons. A neuron with the smallest difference is selected and identified as the pattern neuron for this feature vector. On the basis of pattern neurons, a map including all feature vectors and their assignments to centroids is created.

The current coefficient η for modification of weights is calculated as:

$$\eta = e^{-\frac{e_c}{e_m}},$$

where e_c is a current epoch (its index changing from 1 to e_m), e_m is the maximal number of epochs.

The new desired size n_d of the map is calculated as:

$$n_d = \frac{2(e_c - 1)(n - n_s)}{e_m},$$

where e_c is a current epoch, n is the maximal size of the map, n_s is the initial size of the map, e_m is the maximal number of epochs. If the new desired size of the map is greater than the current one, the size n_c of the map is increased in the following way:

$$n'_c = n_c + 1$$

if n'_c is less than the maximal size of the map.

After changing the size of the map, weights need to be modified. For modification of weights in the map, we calculate auxiliary variables:

$$x_{temp} = x - (x - 1) \cdot 0.111$$
$$y_{temp} = y - (y - 1) \cdot 0.111$$

$$c_1 = (1 - x_{temp})y_{temp}$$
$$c_2 = x_{temp}(1 - y_{temp})$$
$$c_3 = (1 - x_{temp})(1 - y_{temp})$$
$$c_4 = x_{temp}y_{temp}$$

Weights of the map for each input feature vector are changed in the following way:

$$M[x][y][i] = c_1 M[x-1][y][i] + c_2 M[x][y-1][i] + \\ + c_3 M[x-1][y-1][i] + c_4 M[x][y][i]$$

for $i = 1, \ldots, k$.

A difference between a given feature vector *input* and weights of all neurons is calculated from the formula:

$$d = \sqrt{\sum_{i=1}^{k} (M[x][y][i] - input[i])^2}$$

for each neuron in the position x and y. Next, a neuron with the smallest d is selected and neighboring neurons (± 1 neurons in both directions, i.e., x and y) are modified according to:

$$M'[x][y][i] = M[x][y][i] + \\ + \eta(input[i]M[x][y][i])(1 - 0.4abs(x - x_{top}))(1 - 0.4abs(y - y_{top}))$$

for each $i = 1, \ldots, k$, where *abs* is the function of an absolute value, $x_{top} = x + 1$ and $y_{top} = y + 1$.

Results of the clustering process are presented in the form of minimal spanning trees with respect to distances between feature vectors and centroids.

3 Examined Data

Examined data consisted of entrepreneurship and macroeconomic indicators called World Development Indicators (WDI) published by the World Bank [1]. The exemplary indicators published in the report and describing the entrepreneurship we choose for our research are:

- New business density,
- Start-up procedures to register a business,
- Firms using banks to finance investment,
- Time to resolve insolvency,
- Strength of legal rights index,
- Time to prepare and pay taxes,
- Firms expected to give gifts in meetings with tax officials,
- Researchers in R&D,
- Patents and trademark application,
- High-technology exports.

Periodicity of data is annual. They cover developing and high-income economies. For each selected country, we have a time series consisting of annual values of a given indicator (cf. Table 1). Therefore, for the clustering process, we have as many feature vectors as many countries is selected. Each clustered object represents a time series. Examined indicators come from years of the first decade of 21st Century.

4 Challenges

The presented paper constitutes the first attempt to dealing with the problem of identification of correlations between groups of time series obtained from the clustering process. Therefore, it has rather a rudimentary (introductory) character. In this section, we give some challenges of further research.

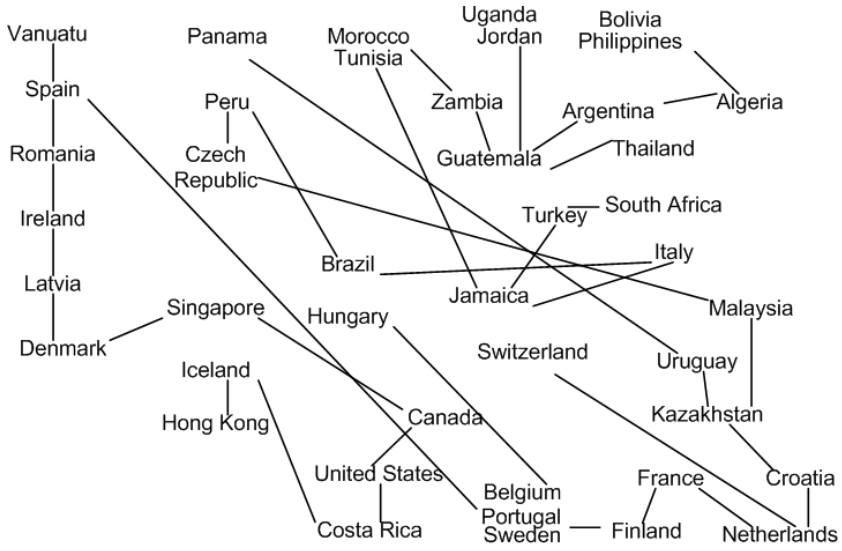

Fig. 1. An exemplary result of the clustering process: a spanning tree for the indicator "New business density"

As the result of clustering process of the set of time series corresponding to a given indicator, we obtain a minimal spanning tree with respect to distances between feature vectors and centroids. An exemplary spanning tree is shown in Figure 1. It presents clusters of countries regarding the indicator called "New business density" showing new businesses registrations per thousand population 15-64 years old. According to the Figure 1 we can notice several groups of countries with similar values of the indicator, i.e., one cluster form countries: Vanuatu, Spain, Romania, Ireland, Latvia, Denmark, Singapore; whereas the second one

covers: Bolivia, Philippines, Algeria, Argentina, Guatemala, Uganda, Jordan, Zambia, Morocco; and in the third we have: Malaysia, Kazakhstan, Uruguay, Croatia, Netherlands, France, Finland, Belgium, Portugal and Sweden.

In order to identify correlations between groups of time series formed in the clustering process, we need to apply some methods for comparison of topological structures of minimal spanning trees. In simple case, we can make one-to-one comparison, i.e., we compare a minimal spanning tree of one of the indicators with the one of another indicator. Results of comparison process should enable us to identify entrepreneurship determinants.

Moreover, we plan to test other clustering methods, among others, that proposed by us (see [7]) based on the ant principle. It is worth noting that we need to use clustering methods without a predetermined number of clusters. A fixed number of clusters can disturb the process of searching for correlations.

References

1. WDI, http://data.worldbank.org/data-catalog/world-development-indicators
2. Amit, R., Glosten, L., Muller, E.: Challenges to theory development in entrepreneurship research. Journal of Management Studies 30(5), 815–834 (1993)
3. Audretsch, D. (ed.): Entrepreneurship, innovation and economic growth. Edward Elgar Publishing Limited (2006)
4. Cios, K., Pedrycz, W., Swiniarski, R., Kurgan, L.: Data mining. A knowledge discovery approach. Springer, New York (2007)
5. Harper, D.A.: Foundations of Entrepreneurship and Economic Development. Routledge Taylor & Francis Group, London and New York (2003)
6. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (1982)
7. Pancerz, K., Lewicki, A., Tadeusiewicz, R.: Ant based clustering of time series discrete data - a rough set approach. In: Panigrahi, B.K., et al. (eds.) Swarm, Evolutionary, and Memetic Computing, Lecture Notes in Computer Science, vol. 7076, pp. 645–653. Springer-Verlag, Berlin Heidelberg (2011)