Kirt, T., Vainik, E., & Võhandu, L. (2007). A method for comparing self-organizing maps: case studies of banking and linguistic data. In Y. Ioannidis, B. Novikov, & B. Rachev (Eds.), *Proceedings of eleventh East-European conference on advances in databases and information systems* (pp. 107–115). Varna, Bulgaria: Technical University of Varna.

A Method for Comparing Self-Organizing Maps: Case Studies of Banking and Linguistic Data

Toomas Kirt[1], Ene Vainik[2], Leo Võhandu[3]

[1]Institute of Cybernetics at Tallinn University of Technology,
Akadeemia tee 21, 12618 Tallinn, Estonia
Toomas.Kirt@mail.ee
[2]Institute of the Estonian Language,
Roosikrantsi 6, 10119 Tallinn, Estonia
ene@eki.ee
[3]Tallinn University of Technology,
Raja 15, 12618 Tallinn, Estonia
leov@staff.ttu.ee

Abstract. The self-organizing map (SOM) is a method of exploratory data analysis used for clustering and projecting multi-dimensional data into a lower-dimensional space to reveal hidden structure of the data. The algorithm used retains local similarity and neighborhood relations between the data items. In some cases we have to compare the structure of data items visualized on two or more self-organizing maps (i.e. the information about the same set of data is gathered in different tasks, from different respondents or using time intervals). In this paper we introduce a method for systematic comparison of SOMs in the form of similarity measurement. Based on the idea that the SOM retains local similarity relations of data items those maps can be compared in terms of corresponding neighborhood relations. We give two examples of case studies and discuss the method and its applicability as an additional and more precise measure of similarity of SOMs.

Keywords: Neural networks, data mining, knowledge discovery, semantics.

1 Introduction

The self-organizing map (SOM) is a method to visualize multidimensional data. The SOM performs mapping of multidimensional data onto a two-dimensional map while preserving proximity relationships as well as possible. The results of the SOM analysis are usually assessed visually. Interpretation of the SOM and discovered knowledge depends mostly on an interpreter. Subjective factors such as one's attentiveness to both general patterns and local details of a large number of presented data items might diminish the objective value of data analysis.

When we use different sources of data that describe the same phenomenon but are collected somehow differently or the number of variables is varying then we have to assess whether the results of the two analyses are similar. As the SOM projects close units of the input space into nearby map units the local neighborhood should remain

quite similar. In this paper we propose a simple method to compare the results of different self-organizing maps. The methodology is based on the measurement of similarities of the local neighborhood.

In the first part of the paper the used methods and techniques including similarity measurement methodology are introduced. In the second part of the paper two data sets as case studies are used to illustrate the similarity measurement methodology. Finally there is a discussion to analyze the results and the accuracy of the methodology.

2 Self-Organizing Map

The self-organizing map [2] is a powerful tool to visualize high-dimensional data. It projects nonlinear relationships between high-dimensional input data into a two-dimensional output grid (map). The SOM is an artificial neural network that uses an unsupervised learning algorithm without prior knowledge how systems input and output are connected. For visualization of the self-organizing map a Unified distance matrix (U-matrix) is used. The analysis has been performed by the SOM toolbox [4].

3 Dimensionality Reduction

To reduce dimensionality of the data we use the principal component analysis (PCA). The main idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set [1]. The PCA transforms the data linearly and projects original data on a new set of variables that are called the principal components. Those are uncorrelated and ordered so that the first few components represent most of the variation of the original variables.

4 Matrix Reordering

The matrix reordering is a structuring method for graphs (and general data tables). The method reorganizes the neighborhood graph data vertices according to specific property – systems monotonicity [8], [9]. For example, we start with a simpleminded graph input variant. Then we calculate the Hamming similarity matrix S for the given graph. To reorder the graph for an easy visibility we will find the row sums of H. Then we take the weakest object in the system (one with the minimal row sum) and subtract that chosen object's similarities from the sum vector. We repeat that elimination step n times whereby z is the evolving list of graph nodes in the elimination order. And as the last step we print our graph g in the new order z. The examples of such reordering can be seen in our case studies (Fig.3, Fig.5).

5 Methodology of Similarity Measurement

While the SOM represents data on two-dimensional topological maps the local topological relations between data items can be used to assess whether the maps have similar structure. The local neighborhood is the basis of our approach to measure the similarity between maps and we expect the neighborhood relations to remain stable even when the overall orientation of the map changes.

The proposed methodology to measure similarity between the self-organizing maps consists of four main steps.

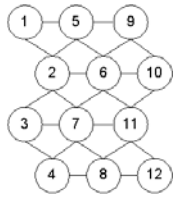


Fig. 1. Neighborhood relations on the SOM.

Firstly, to analyze general organization the resulting map is visually examined and clusters and their borders are identified, also the general orientation and locations of data items are identified. Thereafter the matrix of neighborhood relations is formed. Neighborhood assessment is based on the location of the best matching units (BMU - a point on the map that is the closest to the input data vector) on the self-organizing map. Two data items are neighbors if they are marked to locate on the same node or in the neighboring nodes depending on the neighborhood range. The neighborhood on the hexagonal map is demonstrated on Fig. 1. The neighborhood matrix is an n-by-n square symmetric matrix N where n is the number of data items and the matrix can also be regarded as a graph. If there is neighborhood relation between i*th* and j*th* element then the value of the matrix element is marked 1 and 0 otherwise.

$$n_{ij} = \begin{cases} 1, if neighbor \\ 0, otherwise \end{cases}$$

Next stage of similarity analysis is the calculation and assessment of similarity coefficients [6]. The coefficients have typically values between 0 and 1. A value 1 indicates that the two objects are completely similar and a value 0 indicates that the objects are not at all similar. We have used two coefficients, such as the Simple Matching Coefficient (SMC) and Jaccard coefficient (J).

$$\text{SMC} = \frac{\text{number of matches}}{\text{total number of variables}}. \quad (1)$$

The SMC rates positive and negative similarity equally and can be used if positive and negative values have equal weight.

Jaccard Coefficient (J) is used if the negative and positive matches have different weights (are asymmetric).

$$J = \frac{\text{number of positive matches}}{\text{number of variables - negative matches}} . \quad (2)$$

Jaccard Coefficient ignores negative matches and can be used if the variables have many 0 values.

If the value of the Jaccard coefficient and SMC is below 0.5 then the number of positive matches is less than half of the total matches.

Fourth part of the similarity measurement consists of finding how much the two neighboring matrixes are identical what is a maximum isomorphic subset. The task is not as complicated as the general isomorphic graph problem, because the order of the data items is known and to identify the maximum isomorphic subgraph we can use an AND operator. If a_{ij} & b_{ij} (elements of the neighborhood matrixes have both value 1), then the neighborhood relation is isomorphic. Here we can perform a new meta-level analysis and reorder and visualize the isomorphic sub-graph to see commonly shared information between two maps. For output the Graphviz[1] software has been used.

6 Case Studies

We use two sets of data to illustrate the method of similarity measurement. The first is a research into the concepts of emotion in Estonian language. The survey consisted of two parts and as a result two different data matrixes describe the same set of emotion concepts. In our meta-analysis we attempt to analyze whether and to what extent the results of two tasks are comparable. The second data set is banking data. In this case the purpose of our meta-analysis is to detect whether and to what degree the dimensionality reduction method (PCA) applied to the data has preserved its structure. Those two data sets reveal different aspects of the comparison methodology.

6.1 Study of Estonian Concepts of Emotion

The purpose of the study was to discover the hidden structure of the Estonian emotion concepts and test a hypothesis that the way the information about concepts is collected can influence its emergent structure. Two lexical tasks were carried out providing information about emotion concepts either through their relation to the episodes of emotional experience or through semantic interrelations of emotion terms (synonymy and antonymy).

Subjects and Procedures

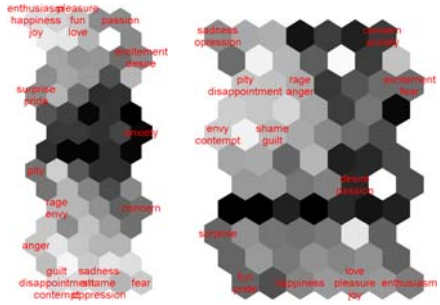
The inquiry was carried out in written form during the summer months of 2003 in Estonia. There were 24 emotion concepts selected for the study based on the results of tests of free listings [7] and also on word frequencies in the corpora. The participants

[1] Graph Visualization Software available from http://www.graphviz.org/

had to complete two tasks measuring the concepts by means of different levels of knowledge (see [10]). In the first task they had to evaluate the meaning of every single word against a set of seven bipolar scales, inspired by the Osgood's method of semantic differentials [5]. In the second task the same participants had to elicit emotion terms similar and opposite by meaning to the same 24 stimulus words.

Analysis by SOM and Meta-Analysis of Neighborhood Relations

The data of both tasks was analyzed by SOM (Fig. 2). In a visual comparison of the two maps we could see completely different structures, but there is a clear distinction of concepts of positive vs. negative emotions observable on both maps. The locations of these clusters are reversed, however. In addition, the upper part of Fig. 2b is divided into two subclusters as there is a group of concepts located in the uppermost right edge of the graph. It is hard to decide whether the obtained structures are different enough to claim the hypothesis that the way of approach (in form of our two tasks) can influence the emerging conceptual structure, proved.



a) b)

Fig. 2. a) Results of the Task 1 (24 concepts evaluated on the seven bipolar scales), b) Results of the Task 2 (24 concepts arranged according to relations of synonymy and antonymy)

Table 1. Neighbourhood similarity of the SOM of conceptual data

Neig. range	Task 1 neig.	Task 2 neig.	Similar neig.	Total neig.	SMC	Jaccard coef.
1	96	100	52	144	0.8403	0.3611
2	198	216	150	264	0.8021	0.5682
3	276	346	230	392	0.7188	0.5867

The summary of the neighborhood relations between two tasks is given in Table 2. The number of relations is measured with different range of neighborhood, starting from 1 to 3. Increase in neighborhood range causes also increase in the number of relevant neighborhood relations. The SMC coefficient is decreasing if the neighborhood is increasing because of possible connections between the words that actually do not belong to the same neighborhood. We could use the SMC as an indicator of stability. Jaccard coefficient is increasing if the neighborhood range is

widening and there is tendency to have more positive matches if the number of neighborhood relations increases.

As far as the neighborhood range remains open it is still difficult to decide, whether the two SOMs of our two tasks were different enough to claim our hypothesis of the case study proven.

The second step of meta-analysis is to find a maximum isomorphic sub-graph and to find a clue what the suitable range of the neighborhood could be. In the case the neighborhood range was provisionally set on 1, several separate fragments of conceptual networks were formed. The general structure of the data did not appear as a connected system. With the neighborhood range 2, the graph became connected. One can speculate that it represents the communal structure or a backbone of the conceptual data gathered from two tasks. The reordered data matrix and its graph are visible on Fig. 3. The lighter part of the reordered matrix is isomorphic part of the matrix.

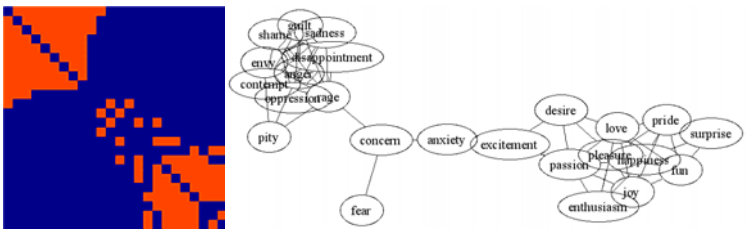


Fig. 3. Reordered and visualized isomorphic subgraph of lexical data (Task 1 vs. Task 2). Neighborhood range 2.

A conclusion can be drawn, that the match of the two structures based on our two tasks is partial, and it is measurable in principle. The degree of measured structural isomorphism depends on the rigidity of the selected criteria of neighborhood.

6.2 Study of Banking Data

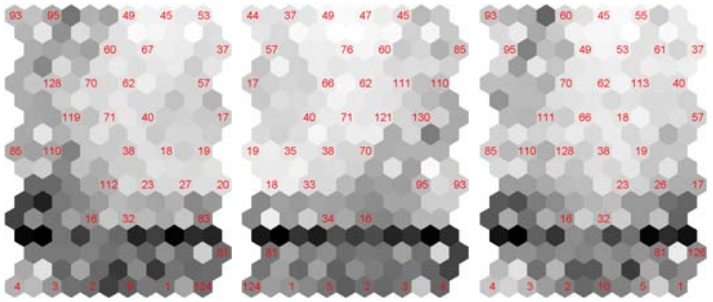
The second data set is used to illustrate the impact of dimensionality reduction by PCA on the SOM maps. The aim of the study is to measure the similarity between the results of SOM mapping of original data and reduced data.

The Banking Data

The second data set consists of banking data (1997—2000; http://www.bankofestonia.info). We have used 133 public quarterly reports by individual banks as a balance sheet and profit / loss statement (income statement). The 50 most important variables have been selected to form a short financial statement of a bank. All the variables are normalized by the variable of total assets to make the reports comparable.

Analysis by SOM and Meta-Analysis of Neighborhood Relations

We formed three sets of the banking data. The first set consisted of all 50 original variables (Original), for the second set 26 principal components describing 95% of variation were selected (PCA95) and for the third set 5 principal components describing 50% of variation were selected (PCA50). From those data sets three self-organizing maps were created (Fig. 4). Our aim has been to measure how similar those maps are and whether similar banks are projected into nearby map units in all cases.



a) b) c)

Fig. 4. a) SOM of 50 original variables, b) SOM of 26 principal components describing 95% of variation, c) of 5 principal components describing 50% of variation

Analyzing the maps visually we can see that in general the maps have a similar structure. As we are interested in overall structure we marked only the first BMUs on the map. The labels are referring to the number of a report. Comparing the SOMs we could identify one bigger group on top, another on bottom and a darker area between them. The original and PCA50 map seem to be rather similar but in case of the PCA95 left-right sides are interchanged. On the bigger light area on top of the SOM there are located the bigger and main retail banks. At the bottom some smaller and niche banks are gathered.

Table 2. Neighbourhood similarity of the SOM of banking data

Experiment	Neig. range	Orig neig.	PCA neig.	Similar neig.	Total neig.	SMC	Jaccard coef.
Orig vs. PCA 95%	1	1876	1772	1426	2222	0.9550	0.6418
Orig vs. PCA 50%	1	1876	2120	1480	2516	0.9414	0.5882
Orig vs. PCA 95%	2	4078	3972	3122	4928	0.8979	0.6335
Orig vs. PCA 50%	2	4078	4038	3046	5070	0.8856	0.6008

In Table 2 the similarity measurement coefficients of the banking data are given. The density of data items on the map is quite high and it is also visible in the number of neighborhood relations. There is a slight difference in the number of neighborhood relations between PCA95 and PCA50. It shows that the PCA retains the internal structure of the data items. The SMC value is very high in all cases, but is becoming lower if the neighborhood range is widening. The Jaccard coefficient shows about 0.6 similarities between the different representations of the data items.

As the banking data consisted of 133 data items the neighboring relations were much stronger than in case of linguistic data. In Fig. 5 there is given an isomorphic subgraph showing neighborhood relations between the SOM of original data and the SOM of PCA95. The neighborhood range is defined as 1. The graph illustrates quite well the structure within the data. The same grouping was visible on the graph representing only 50% of variations. When the neighborhood range was increased the isomorphic sub-graph became connected but at the same time the neighborhood relations became so dense that the structure was not clearly visible any more.

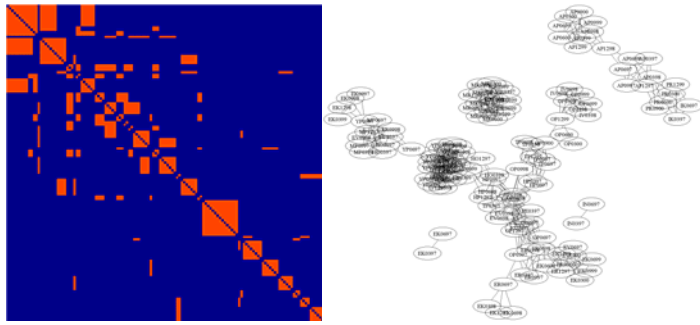


Fig. 5. Reordered and visualized isomorphic subgraph of banking data (Original vs. PCA 95% variation). Neighborhood range 1.

The analysis what is the impact of dimensionality reduction on the results gave us confirmation that even the dramatic dimensionality reduction by the PCA method retains the most important internal relations in the data.

7 Discussion and Conclusions

The case studies give an overview of possibilities to measure similarity between self-organizing maps that is based on the topology and neighborhood relations. We find local neighborhood relations between the data items and measure the similarity of relations by coefficients and by finding an isomorphic subgraph. There has been proposed another method to evaluate two- or three-dimensional visualizations and to measure distances between the two representations by Mandl and Eibl [3]. They calculate Euclidian distances between all the items and find correlation between two representations. We prefer to use local topological representation and not to convert it once more into the Euclidian space.

The similarity measurement coefficients together could give some additional information about the similarity. If the SMC value is high and at the same time the Jaccard coefficient has lower value then it indicates the presence of clustered structure. The bigger the difference is the smaller are the clusters. We can also use the SMC coefficient if the data items are exclusive like in the case of lexical study. There

were two exclusive groups of data – positive and negative emotion concepts – that had weak neighborhood relations.

We expected to use the maximum isomorphic subgraph as a measure to identify the similarity between the SOMs, but it became a new meta-level tool to find hidden structure and to reveal the grouping structure of the data. The main parameter in similarity analysis is the range of neighborhood. The number of neighborhood relations increases if the number of data items or the range of the neighborhood widens. The size of a map has also an impact on the density of data items on the map. Depending on the density of the data items range 1 or 2 gives good insight into the hidden structure or the so-called backbone within the data. In both case studies the visualized isomorphic neighborhood matrix gave us a new perspective on relations between the data items.

In this paper, we have proposed a methodology to measure similarity between the self-organizing maps if the maps are describing the same phenomenon but use different sources of data or the number of variables are different. We illustrated the methodology by two sets of data. The results of the two case studies have shown us that the suggested method to measure similarity between two self-organizing map is applicable and it gives new insights into the data.

Acknowledgements. This study was supported by the Estonian Science Foundation, grant No 7149 and G5918.

References

1. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (2002)
2. Kohonen, T.: Self-Organising Maps. 3rd edn. Springer, Berlin (2000)
3. Mandl, T., Eibl, M.: Evaluating Visualizations: A Method for Comparing 2D Maps. In: Smith, M., Salvendy, G., Harris, D., Koubek, R. (eds.): Proceedings of the HCI International 2001 (9th International Conference on Human-Computer Interaction). Lawrence Erlbaum Associates, London (2001) 1145–1149
4. Alhoniemi, E., Himberg, J., Parhankangas, J., Vesanto J.: SOM Toolbox (Version 2.0). [Computer software and manual]. (2005) Retrieved November 11, 2005, from http://www.cis.hut.fi/projects/somtoolbox/
5. Osgood, C.E., Suci, G.J., Tannenbaum, P.H.: The Measurement of Meaning. University of Illinois Press, Urbana and Chicago (1975)
6. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison Wesley, Boston (2005)
7. Vainik, E.: Emotions, Emotion Terms and Emotion Concepts in an Estonian Folk Model. Trames, 6(4) (2002) 322–341
8. Võhandu, L.: Express Methods of Data Analysis. Transactions of Tallinn TU, 464 (1979) 21–35
9. Võhandu, L.: Fast Methods in Exploratory Data Analysis. Transactions of Tallinn TU, 705 (1989) 3–14
10. Vainik, E.: Lexical Knowledge of Emotions: The Structure, Variability and Semantics of the Estonian Emotion Vocabulary. Tartu University Press, Tartu, Estonia (2004)