Contributing Authors

Timo Honkela is Professor of Applied Cognitive and Information Processing Science at Media Lab of University of Art and Design, Helsinki, Finland. He received his PhD from Helsinki University of Technology with his thesis on Self-Organizing Maps in Natural Language Processing.

Ilpo Koskinen is Professor at Department of Industrial Design of University of Art and Design, Helsinki, Finland. He received his PhD from University of Helsinki.

Timo Koskenniemi is Research Assistant at Media Lab of University of Art and Design, Helsinki, Finland.

Sakari Karvonen is Researcher at National Research and Development Centre for Welfare and Health and Docent in Medical Sociology at University of Helsinki, Department of Public Health. He received his PhD from University of Helsinki.

Chapter 1

KOHONEN'S SELF-ORGANIZING MAPS IN CONTEXTUAL ANALYSIS OF DATA

Timo Honkela
Media Lab
University of Art and Design Helsinki
timo.honkela@uiah.fi

Ilpo Koskinen
Department of Industrial Design
University of Art and Design Helsinki
ilpo.koskinen@uiah.fi

Timo Koskenniemi
Media Lab
University of Art and Design Helsinki
timo.koskenniemi@iki.fi

Sakari Karvonen
National Research and Development Centre for Welfare and Health
STAKES
sakari.karvonen@stakes.fi

Abstract Kohonen's Self-Organizing Map (SOM) is a means for automatically arranging high-dimensional statistical data. The map attempts to represent all the input with optimal accuracy using a restricted set of models or prototypes. The prototypes also become ordered on the map grid so that similar prototypes are close to each other and dissimilar prototypes far from each other. The SOM is useful in clustering, abstraction, and visualization through dimensionality reduction. It has been used in a multitude of application areas ranging from speech recognition to data mining of texts and from robotics to process monitoring. The unsupervised learning scheme of the SOM makes it well suited for applications in which the

input data cannot be labeled. A map is ordered and it follows the patterns of the input data in a non-linear but generalizing fashion. All this makes it well suited for data analysis and many areas in developing intelligent systems. In this article, the general principles of using the SOM in data analysis are considered reflecting on the concept of context. An illustrative experiment of data analysis is presented.

Keywords: Self-organizing map, SOM, data analysis, contextuality

Introduction

A single symbol or variable is associated with multiple occurrences of complex objects or processes in almost any representation of reality. In data analysis, it would thus be useful if one could directly link the symbol with the reality through some kind of artificial sensory mechanisms. However, as this linking is not most often possible, the complex relationships should be taken into account. One needs to consider, e.g., fuzziness, interdependencies between variables, and changes in the domain. Traditional artificial intelligence methods for knowledge representation, such as semantic networks, frame systems and rule-based systems seem to suffer from problems related to their approach of knowledge representation. For instance, the symbolic AI representations are not grounded in the perceptual domain and they are based on the assumption that the any domain consists of collection of objects, their distinct properties and the relationships between the objects.

The topics mentioned above are relevant also when statistical methods for data analysis are considered. Although the data may be in continuous numerical form, it is usually gathered and represented in such a way that a division into distinct well-defined variables is in use. Moreover, before any analysis, simplifying assumptions may be set, e.g., concerning mutual independence of the variables or their distribution. Moreover, the analysis is based a predefined set of hypotheses. One may try to find general principles that are valid in all or most of the cases under consideration. However, it may be that there are only very few rules or principles that are present in the overall data but there are local patterns that can serve as an important starting point for the data analysis. This kind of line of thinking is clearly present in the case-based reasoning approach that is considered below.

This article discusses how Kohonen's Self-Organizing Map (SOM) and its variants can be used in solving some of the problems outlined above, and provides directions for further development. Example of the use of the SOM for data organization is given. Both analysis of numerical data and written texts are considered. Specific emphasis is given to the point of view of contextuality: how can one represent and analyze data in such a way that the relevant context is taken into account.

1. SELF-ORGANIZING MAP

The Self-Organizing Map was developed by Teuvo Kohonen in the beginning of 1980s [12, 13]. The basic Self-Organizing Map (SOM, also denoted as Kohonen map) can be visualized as a sheet-like neural-network array. The map is an adaptive system. In the adaptation process the nodes in the array become specifically tuned to various input patterns or classes of patterns in an orderly fashion. The learning process is competitive and unsupervised: no teacher is needed to define the correct output for an input. The locations of the responses in the array tend to become ordered in the learning process as if some meaningful non-linear coordinate system for the different input features were being created over the network [13].

Perhaps the most typical notion of the SOM is to consider it as an artificial neural network model of the brain, especially of the experimentally found ordered "maps" in the cortex. There exists quite a lot of neurophysiological evidence to support the idea that the SOM captures some of the fundamental processing principles of the brain [13].

The SOM can also be viewed as a model of unsupervised learning, and as an adaptive knowledge representation scheme. The traditional knowledge representation formalisms (e.g., semantic networks, frame systems, predicate logic) are static and the reference relations of the elements are determined by a human. For a comparison, see [5].

The SOM is nowadays often used as a statistical tool for multivariate analysis. The SOM is both a projection method which maps high-dimensional data space into low-dimensional space, and a clustering method so that similar data samples tend to be mapped to nearby neurons. A related application area is data mining and visualization of complex data sets. Application areas include, for instance, image processing and speech recognition, process control, economic analysis, and diagnostics in industry and in medicine [13].

1.1 BASIC PRINCIPLES

Assume that some set of data have to be mapped onto a grid. Each input sample in the set is described by a real vector. Each node in the map (which is usually a two-dimensional grid) contains a prototype vector which has the same number of elements as the input vector. The basic idea in the Self-Organizing Map is that, for each sample input vector the best-matching winner and the nodes in its neighborhood are changed to represent the input data better. The net outcome in the learning process is that ordered values for the prototype vectors emerge over the array.

At the beginning of the learning process the radius of the neighborhood is fairly large, but it is made to shrink during learning. This ensures that the global order is obtained already at the beginning, whereas towards the end, as

the radius gets smaller, the local corrections of the prototype vectors in the map will be more specific. The learning rate factor also decreases during learning.

A number of details about the selection of the parameters, variants of the map, and many other aspects have been covered in the monograph [13].

1.2 CASE-BASED REASONING

In Case-Based Reasoning (CBR) problem solving is based in a collection of past cases rather than being encoded in generic rules or other knowledge descriptions (see, e.g., [15]). Each case typically contains a description of a problem and a solution. In order to solve a current problem it is matched against the cases in the case base, and similar cases are retrieved. The retrieved cases are used to suggest a solution which is reused and tested for success. The solution may be revised if necessary. The current problem and the final solution are stored as part of a new case.

While analyzing a collection of samples or cases of any kind one should consider whether the they come essentially from a similar source and are thus comparable with each other. Consider, e.g., a process monitoring system of a factory in which the machinery is changed to a large extent. Afterwards the comparison between the "old" and the "new" data is questionable. The same kind of question can be asked when medical measurements are considered. If a set of people is studied, one can question whether all the persons are similar enough with respect to the phenomenon being studied. It is important to discover patterns and hidden variables that are relevant only for a small portion of the whole set of individuals or cases. The SOM is very well suited for such an analysis: it can both form clusters as well as visualize the relationships between the clusters as neighborhood relations on the map. In addition, the variable planes illustrate effectively the interdependencies between the variables. Also the phrase component plane is often used but in this article we use the word variable systematically to refer to a component or element in the input and prototype vectors. In using the SOM, the number of variables can be high. For instance, in a study of the socio-economic status of the countries in the world based on World Bank data, 39 variables were used [10].

2. CASE STUDY

In this section we describe the data used in the case study and the results of our exploratory analysis.

2.1 DESCRIPTION OF DATA

Data for this paper comes from a sample of young people gathered in 1998 Helsinki, Finland. These data were collected from a sample of schools in this region. Out of schools in the area, 75% were covered by the sample. The

number of observations was 2373. Response rate was 75%. The study from which these data originated aimed at analyzing the relationship of lifestyle to health-related behavior among teenagers born between 1976 and 1984. Most respondents were born in 1982 (mode).

For this paper, a subset of variables was selected for closer analysis. This subset consisted of 38 variables, which described a number of health-related variables, lifestyles, as well as traditional sociological characteristics. Health-related and lifestyle variables measured weight and height, exercise habits, alcohol habits, possible contact with drugs, and a measure of depression (using Beck's scale [1] with one item, suicide, removed). Future orientations were measured by the respondents' plans for future education, varying from vocational training to university-level education. Respondents' attitude towards schooling was measured by their success in school, by their attitude towards school, and their unauthorized absence from school. More traditional sociological characteristics measured were family characteristics, and a possession of such electronic devices as CD players, TV, personal computer, and mobile phone. The main aim of the original study (see [8]) was to explore whether lifestyle variations explain health-related behavior in the Helsinki area. In addition, the study aimed to explore how lifestyle and social class were related health-related behavior.

A special feature of the data was that there were several groups of deviant responses to questions about weight and height. If we express the relationship with Body-Mass Index (BMI) in which weight is divided by the square of the weight, we get in index in which variation in population is normally between 15 and 40. In the Helsinki data, this index varied from 0 to 10 000.

In terms of the current paper, these data give us an opportunity to explore the ways in which health-related behavior is related to lifestyle issues, which are seen as contextual variables. Outliers in the body-mass index give us a further possibility to explore the use of the SOM in analyzing deviant cases in data.

2.2 PREPROCESSING OF DATA

Variables with nominal scale were split to separate binary variables, e.g., variable denoting educational plans which had five distinct values was split into four different binary variables (vocational, high school, university, or no opinion stated). Cases of persons with no plans for further education were coded with all the four variables having a zero value.

The SOM organizes data so that the variables with largest variance effect the result most. Because a neutral map was desired, data was normalized by dividing every variable with its variance.

Crude initial map was calculated and variable planes were visualized (Figure 1.1). Each variable's values in the prototype vectors of the organized map

are shown as a grayscale plot. Values have been scaled to fit full range from white (low values) to black (high values) in visualization. Thus, only the distribution on the map and not the absolute values of the variables is revealed in this figure.

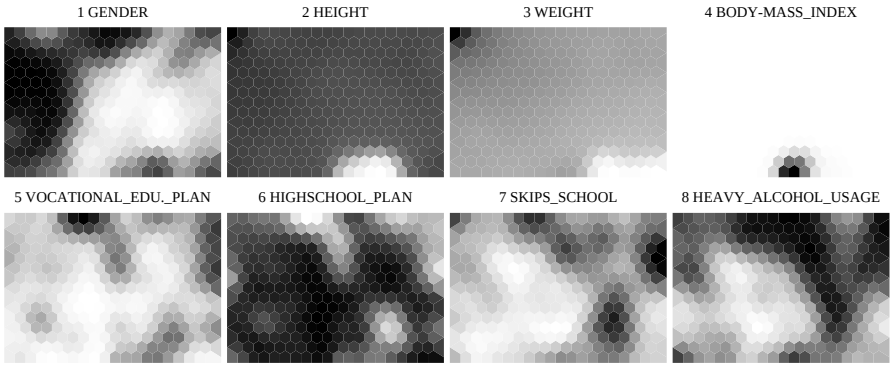

Figure 1.1 Collection of variable planes based on unpreprocessed, original data.

It is important to examine the data thoroughly before undertaking any formal analysis. With the large size of data sets a manual inspection is often not feasible or may even be impossible in practice. David J. Hand mentions [2] that anomalous data, or data with hidden peculiarities, can only be shown to be such if we can tell the computer what to search for. Peculiarities which we have not imagined will slip through the net and could have all sorts of implications for the value of the conclusions one draws. However, the SOM can provide useful insight on potentially problematic data. Indeed, visual inspection of the variable planes in our analysis revealed unrealistically high and low values of weight, height and body-mass index in some samples. These areas were clearly discernable in corresponding variable planes as sharp dark or light spots (see Figure 1.1). The outlier data was then removed from the input data set by finding nodes with strongly deviant values and then removing all the samples associated with these nodes (approximately 5% of all the samples). Even without removal of the data samples with peculiarities, many of the correlations found in the actual analysis (see next section) were already visible in the first analysis. This is possible because the SOM performs a non-linear analysis which means that the effects of the extremes in the data are restricted in small areas whereas linear models may suffer significantly from outliers.

2.3 RESULTS

The filtered data was used to create a 10 times 15 unit map of individuals. In this paper, however, we do not display this map but study the collection of variable planes (see Figure 1.2) based on the organized map. These variable planes illuminate the connections and correlations between the variables. In the following, some representative examples are considered. The depressed persons (variable 38) in the material are concentrated in one area near the lower right corner of the map. A very similar distribution is discernable in the variable denoting the number of friends (36) with a negative correlation. The same prototypic cluster can also be found on the variable map for owning stereo equipment (32). There one can also find two other clusters. Thus, those depressed tend to have stereos but only some (visually, roughly one third) of those who have stereos are depressed. Some variables are highly correlated which is already recognizable through visual inspection. Such pairs of variables are mother's education (26) and father's education (27), plans to aim at highschool studies (4) and university studies (5), alcohol usage (16) and heavy alcohol usage (15) and to some extent also smoking (14), and persons that like school (34) versus the ones who consider school as a waste of time (35) (negative correlation). Also height (9) and weight (10) are naturally positively correlated for which the exceptions are illuminated by the BMI (11). This variable has two clusters of high values. The one close in the lower part of the map seems to correlate strongly with those who have low general health level (12). Actually, these prototypes seem also be closely related to the area of depressed persons (38) and to the ones having low number of friends (36), as indicated earlier. A number of other similar and more fine-grained hypothesis could be made.

Figure 1.3 shows an analysis of the relations between the variables based on the basic map presented in Figure 1.2 (see also [18]). Variable planes shown in Figure 1.2 were scaled between one and zero. Each plane consisting of 10 times 15 hexagonal grid was converted into a 150-dimensional vector. These vectors were labeled with the name of corresponding variable. For each vector, an inverted copy was generated to find possible negative correlations. Inverted copies were marked with labels starting with an asterisk to distinguish them from the original ones. The map shown in Figure 1.3 was generated from this data derived from the variable planes. Every variable shows up twice on the map. Because of the inverted versions we can see also the negative correlations.

3. CONTEXT IN NATURAL LANGUAGE PROCESSING

Contextuality of interpretation is easily neglected, being, nevertheless, a very commonplace phenomenon in natural language (see, e.g., [6]). Contextuality of interpretation should be taken into account also when numerical variables are

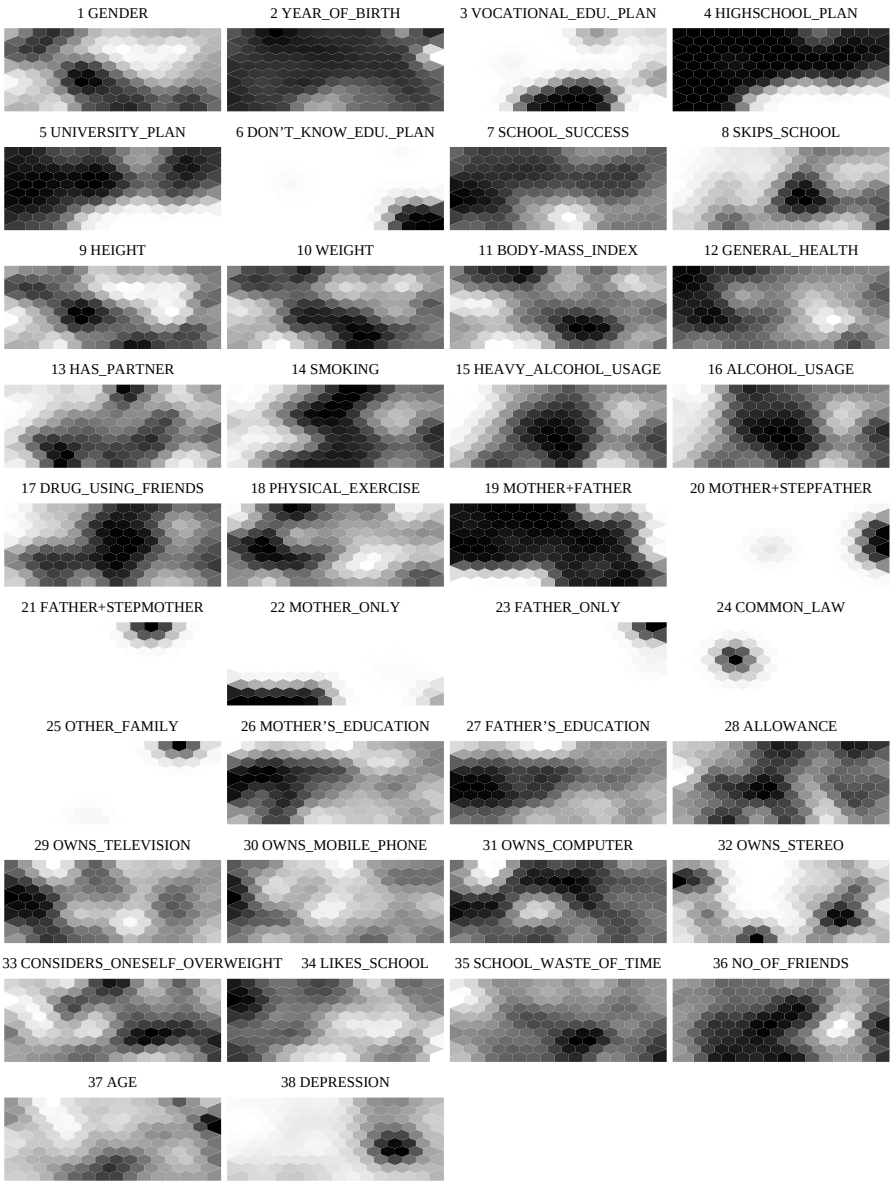

Figure 1.2 Collection of variable planes. Distribution of each variable over the map is indicated using shades of gray. Dark area corresponds to a high value of a variable.

considered: usually the semantics of a variable are defined and communicated through the use of natural language expressions. Thus, the interpretation of

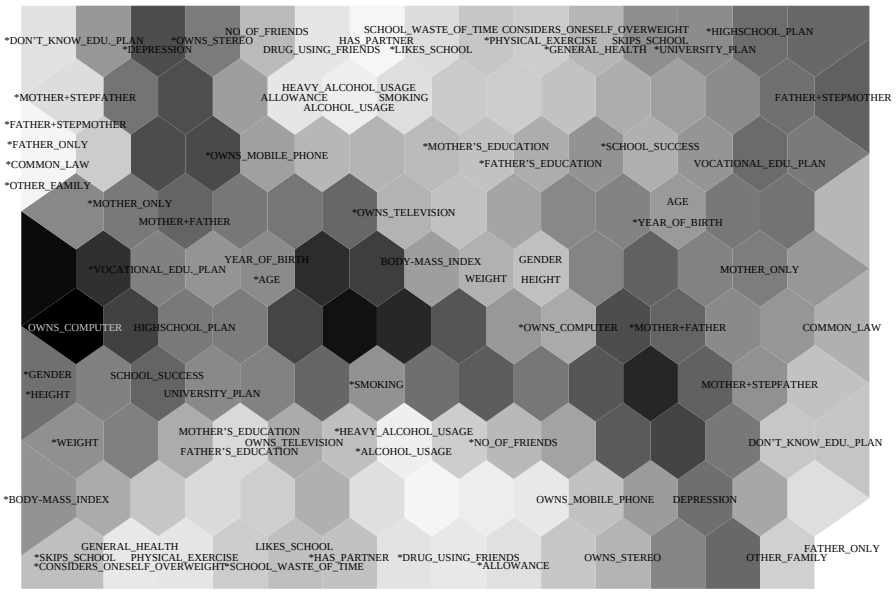

Figure 1.3 Variable map, i.e., a SOM with variables organized based on the variable planes. Each variable appears twice on the map. Inverted copies of the original variables are marked with labels with an additional asterisk. The level of gray indicates the distances in the input space. The darker the area between a pair of variables the larger is the distance between them.

the variables are subject to the potential fuzziness of the defining expressions (e.g., what is considered depression). Moreover, the interpretation of natural language is subjective. Traditionally, the variation in interpretation because the subjectivity has been either neglected (as in the case of traditional symbolic AI knowledge representation formalisms) or it has been considered as noise in the models (in statistical analysis). The nonlinear nature of the SOM analysis and the possibility of interpreting the levels on the maps as some kind of fuzzy membership functions already provides a means to approach fuzziness and subjectivity [5].

When the input consists of symbols, e.g., words in a text, context is required to provide basis for organization. Namely, similarity in the appearance of the words does not usually correlate with the content they refer to. As a simple example one may consider the words "window", "glass", and "widow". The words "window" and "widow" are phonetically close to each other, whereas the semantic relatedness of the words "window" and "glass" is not reflected by any simple metric. One useful numerical representation can be obtained by taking into account the sentential context in which the words occur (see, e.g.,

[17]). The SOM analysis of the words based on their contexts provides a map in which syntactic and semantic relations are visible [3, 17]. The areas on the map that reflect the traditional linguistic categories emerge automatically even though they have not been used in the input.

By virtue of the SOM, also text documents can be mapped onto a two-dimensional grid so that related documents appear close to each other. The visualized document map provides a general view of the document collection (see, e.g., [4, 9, 11, 14, 16]). The self-organized document map offers a general idea of the underlying document space. Moreover, the surroundings on the map for any document also serve as a context for the examination. In this article, using the the SOM for the analysis of numerical data is considered in more detail but the basic ideas are applicable regarding both text and numerical data.

4. ANALYSIS OF SOCIAL CONTEXT OF HEALTH-RELATED BEHAVIORS

In the social sciences, as well as in their derivatives, contextual analysis usually means analyses that use areal data in explaining some set of behaviors or beliefs [7]. Typically, contextual variables represent administrative or areal units in a linear equation. Thus, a typical contextual model might explain, say, the effect of wealth of neighborhoods in Helsinki to health-related behaviors within the city limits.

A major problem in contextual analysis of data is multicollinearity. An analysis based on the SOM makes the effect of multicollinearity clearer. Although an analyst might find a correlation of the wealth of neighborhood to a set of health-related behaviors, there are several additional thing that have to be taken into account before this correlation can be taken for real. Briefly, even after we have (statistically) controlled for a set of possible explanations, and still find that adding a contextual variable into our equation is significant, we have to take additional steps in analysis. Namely, even after performing regression diagnostics, there remains a possibility that the contextual variable correlates with several unmeasured variables that in turn correlate with measured variables in unknown ways.

This situation is typical. It is easy enough to imagine how in wealthier neighborhoods families have several cars, which improves their access to far-away exercise facilities such as ice-hockey arenas or football halls. If this is the case, our estimates may be inaccurate, and even though our results might still be valid, we may face difficulties in interpreting our data. It is this kind of a situation in which self-organizing maps may prove to be useful, at least in the explorative phase of data analysis, as previous figures show.

Now, if we read Figure 1.3 as well as Figure 1.2 anew from the contextual analysis viewpoint, we may first note that we can consider various lifestyle-

related variables, family structure variables, as well as variables that measure scholarly orientation as contextual variables for health-related behaviors. This exploratory analysis readily shows some results.

The first thing to note is that health-related behaviors form quite consistent clusters in the map. Alcohol drinking, smoking, and knowing drug users cluster together in the map. Not liking school, and a small amount of physical exercise are also fitted close to drug-related acquaintances and the use of intoxicants.

The location of other variables in Figure 1.3 helps us to make some interpretations concerning contextual variables. First of all, the amount of money young people can use seems to go with intoxicant-related behaviors. While possession of TV and other devices seem to be related to family structures, this may suggest that teenagers in the sample use their own earnings to intoxicants. Secondly, there appears to be a linkage between parents' educational level and health-related variables. Low parental education correlates with a the use of intoxicants, and with little exercise, although the latter relationship is not so strong as the former. Third, negative attitudes towards school correlate with the use of intoxicants and lack of exercise. In turn, high scholarly intentions correlate very strongly negatively with these behaviors. Finally, it is difficult to say anything consistent about the effects of family structure to health-related behaviors with these data, except that a family with both parents present goes with high scholarly intentions.

In terms of contextual inference, the SOM reported in Figure 1.3 allows us to situate health-related behaviors of teenagers in the sample into their behavioral, belief-based, and family structure contexts. Figure 1.3 makes some of the connections of these different types of variables to health-related behaviors easy to inspect, and provide a useful set of hypothesis for further testing. Perhaps most importantly, the map clearly establishes a point to be explained — health-related behaviors — and situates it into a spatial environment in which a mass of other variables are displayed in a readily analyzable form. In doing this, it also provides a way to identify possible contextual variables, and make hypotheses of their interrelationships for further analysis.

5. CONCLUSION AND DISCUSSION

In summary, the Self-Organizing Maps can support exploration of large data sets. i.e., it provides an efficient tool for data mining. This use has already been demonstrated in a large number of applications [13]. More specifically, when traditional linear models are cumbersome, the SOM provides a non-linear analysis and a visual access to a large number of variables and their relationships. This aspect has been studied in more detail in this article. The use of the SOM in a kind of case-based reasoning was considered. This method makes it possible for us to group variables into clusters, and this way to search

for similar collections of cases, even for small portions of cases in data. The importance of considering such local context based on the cluster analysis needs to be emphasized as a principled alternative of creating linear models over the whole data collection.

Moreover, a kind of microcausality can be considered: the analysis can suggest ways to approach small variations, and make hypotheses of them. Earlier in this article the symbolic nature of even numerical variables in traditional statistical analysis was discussed. For instance, in our analysis depression as phenomenon was represented as one variable. However, the phenomenon is highly complex, which is traditionally taken into account by building an index which consists of several items. In this case, the Beck scale consists of 13 such items. Instead of finally using only one composite variable it would be interesting to include such additional high-dimensional data into the SOM analysis process. These additional variables may often have correlations which the SOM analysis can reveal, but which would be difficult or even impossible to detect from the analysis of the composite variables.

References

[1] A.T. Beck and R.A. Steer. Screening for adolescent depression: A comparison of depression scales. *Journal of the American Academy of Child and Adolescent Psychiatry*, 30(1):58–66.

[2] M. Berthold and D.J. Hand (eds). *Intelligent Data Analysis; An Introduction*. Springer, 1999.

[3] T. Honkela, V. Pulkki, and T. Kohonen. Contextual relations of words in Grimm tales analyzed by self-organizing map. *Proceedings of ICANN-95, International Conference on Artificial Neural Networks*, F. Fogelman-Soulié and P. Gallinari (eds), vol. 2, EC2 et Cie, Paris, pp. 3-7, 1995.

[4] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, January, 1996.

[5] T. Honkela. *Self-Organizing Maps in Natural Language Processing*. PhD Thesis, Helsinki University of Technology, Espoo, Finland, 1997.
See $http://www.cis.hut.fi/\sim tho/thesis/$

[6] H. Hörmann. *Meaning and Context*. Plenum Press, New York, 1986.

[7] G.R. Iversen. *Contextual analysis*. Newbury Park, Calif. Sage, 1991.

[8] S. Karvonen and O. Rahkonen. Cultural variation in lifestyle and health-related behavior among young people (in Finnish). *Yhteiskuntapolitiikka* (in print), 2000.

[9] S. Kaski. Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. 82*. DTech Thesis, Helsinki University of Technology, Finland, 1997.

[10] S. Kaski and T. Kohonen. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. *Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, London, England, 11-13 October, 1995, World Scientific, Singapore, pp. 498-507, 1996.

[11] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM – Self-organizing maps of document collections, *Neurocomputing*, 21:101–117, 1998.

[12] T. Kohonen. Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[13] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1995. (Second, extended edition 1997)

[14] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive text document collection E. Oja and S. Kaski (eds.), *Kohonen Maps*, Elsevier, Amsterdam, pp. 171-182, 1999.

[15] J.L. Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, 1993.

[16] X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. In *Proceedings of 14th. Ann. International ACM/SIGIR Conference on Research & Development in Information Retrieval*, pp. 262–269, 1991.

[17] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61(4):241–254, 1989.

[18] J. Vesanto and J. Ahola. Hunting for correlations in data using the self-organizing map. *Proceedings of CIMA'99, International ICSC Congress on Computational Intelligence*. H. Bothe, E. Oja, E. Massad, and C. Haefke (eds.), ICSC Academic Press, pp. 279–285, 1999.