Incorporating Uncertainty Information into the Visual Analysis of Self-Organizing Maps

Tonio Fincke

Index Terms—Uncertainties, Self-organizing maps, Geovisualization

Data uncertainty is a frequent issue in data analysis. It describes the possible deviation of a true value from a value which is estimated to be true. In most cases, only this representing value is considered for analysis. Information about uncertainty can be useful to rate the quality of analysis results. It might also help to find alternative explanations for certain patterns or value combinations within a data set[5]. For the description of spatial uncertainties, several visualization methods have been developed, often with the objective of giving the user the possibility to get a good impression of the reliability of the data at a certain point or area in space[1].

Geovisualization tools provide users with the possibility to use multiple views simultaneously. This includes the display of a geographic map together with visualizations which emphasize different aspects of data. Often these views are linked, such that a selection of one entry would highlight it in all views. The user can thus interactively explore a data set[6].

One of the visualization methods which can be used for such an interactive analysis is the self-organzing map (SOM)[3]. A SOM is a neural network which is often used for data mining due to its capability to discover patterns in large and complex data sets[2]. SOMs map n-dimensional input data vectors onto a fixed number of neurons or codebook vectors of the same dimension. During a process of training, the attribute values of the codebook vectors are altered in such a way that each codebook vector becomes representative for several of the input vectors. The SOM then represents the structure of the input data set. Complementary to their positioning in the n-dimensional input space, the codebook vectors are arranged in a topological order. In most cases this would be a 2-dimensional grid. In such a grid, each codebook vector is represented by a grid cell. These cells can display information about the characteristics of the codebook vectors, like their values for a certain attribute or the distance to their topologically neighboring vectors[8]. Such values can explain the value distribution of a single attribute throughout the data set or group together codebook vectors which are geometrically close in the input space.

In the context of SOM, the topic of uncertainty propagation has not received much attention. This is because after the application of the SOM algorithm usually only the codebook vectors are considered. Although the codebook vectors have attribute values, these values are artificial. Therefore there is no uncertainty which would need to be described.

This leads to a special situation when using SOMs for visual analysis, as usually not the distinct input vectors are displayed on the map. Instead they are represented by the codebook vector they are geometrically closest to in input space. This vector is also called the Best Matching Unit(BMU). In a geovisualization environment, a user could not locate a distinct input vector on the grid, but only its BMU.

To overcome this problem, several ideas have been put forward about how to display distinct input vectors on a SOM output grid. One of those ideas is to place the input vectors randomly on the grid cell associated with its BMU[7]. Another approach is to determine a position on the output grid by calculating the distances between an input vector and all of the codebook vectors[4]. Since each area of the SOM has its own characteristics, the input vector would also have these characteristics. If, for example, a part of the map would represent a certain class, an input vector mapped onto this part would be considered a member of this class.

When an input vector is represented by a distinct point location on the grid, this would imply that the result of the SOM would be definite for the input vector. However, if the attribute values of an input vector are uncertain, it is improbable that the result of the SOM is certain for the input vector. Therefore, the placement of an input vector with uncertain values on a SOM grid should also consist of a mapping of its uncertainty. The idea presented in this abstract is to describe this uncertainty by a region surrounding an input vector's point location on the output grid. Figures 1 and 2 illustrate this idea.

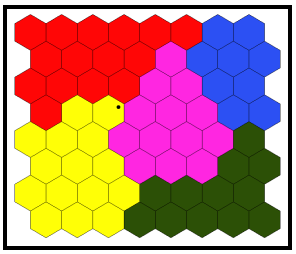

Fig. 1. A sample illustration of a SOM output grid. The grid cells represent output vectors. The colors indicate various classes. In the middle of the grid, a single input vector is placed in the cell of an output vector which is assigned to the yellow class. The input vector therefore seems to be also a member of the yellow class.

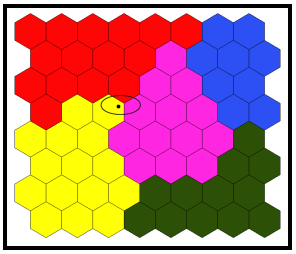

Fig. 2. The same grid as in figure 1. Here the uncertainty of the input vector mapping result is displayed as an area. The area extends over several differently colored grid cells, thus implying that the input vector cannot be unambiguously assigned to a single class.

If the cells in the figures would not depict the membership to a class but the values of one of the attributes, the uncertainty area would then describe the value range of that specific attribute. In this way the values of an attribute could be estimated. This can be helpful when these values are missing for a distinct input vector. Thus the SOM would serve for regression.

It should be noted that there is no single definite method to map an input vector onto the SOM output grid. Consequently, the method of

• *Tonio Fincke is with HafenCity University Hamburg, E-mail: tonio.fincke@hcu-hamburg.de*

how to map the uncertainty onto the grid is also still to be resolved. However, the approach presented in [4] in which exact spatial positions are assigned to each input vector seems to be a good starting point.

Another important issue is that the area will often be irregularly shaped. It might also become discontiguous when output vectors which are geometrically close in input space are not topologically close on the output grid. The size of the area and whether it will be depicted as one or multiple areas will depend on the size of uncertainty of the input vectors, on the structure of the SOM, and on the quality of its training.

It is hoped that when such an uncertainty aware representation of a SOM would be linked and simultaneously displayed with geographic maps and other data visualizations, it can improve an analyst's possibilities to gain insight into the structure of a data set.

REFERENCES

[1] L. Gerharz and E. Pebesma. Usability of interactive and non-interactive visualisation of uncertain geospatial information. In W. Reinhardt, A. Krüger, and M. Ehlers, editors, *Geoinformatik 2009 Konferenzband*, pages 223–230, Osnabrück, Germany, March 2009.

[2] T. Kohonen. *Self-Organizing Maps*. Springer, Espoo, 2001.

[3] E. L. Koua, A. M. MacEachren, and M.-J. Kraak. Evaluating the usability of visualization methods in an exploratory geovisualization environment. *International Journal of Geographical Information Science*, 20(4):425–448, 2006.

[4] G. Liao, T. Shi, S. Liu, and J. Xuan. A novel technique for data visualization based on som. In *International Conference on Artificial Neural Networks No15, Warsaw , POLOGNE*, pages 421–426, 2005.

[5] M. G. Morgan and M. Henrion. *Uncertainty: A Guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge University Press, Cambridge, United Kingdom and New York, 1990.

[6] J. C. Roberts. *Geographic Visualization: Concepts, Tools and Applications*, chapter Coordinated Multiple Views for Exploratory Geovisualization, pages 25–48. Wiley, 2008.

[7] A. Skupin. A cartographic approach to visualizing conference abstracts. *Computer Graphics and Applications, IEEE*, 22(1):50–58, 2002.

[8] A. Ultsch and H. Siemon. Kohonen's self organizing feature maps for exploratory data analysis. In *Proceedings Intern. Neural Networks*, pages 305–308, 1990.