In *International Conference on Artificial Intelligence (IC-AI'99), June 28-July 1 1999, Las Vegas.*

Data Mining Using Self-Organizing Kohonen maps: A Technique for Effective Data Clustering & Visualisation

Jason Ong
School of Computer Sciences
Universiti Sains Malaysia
Penang, MALAYSIA

Syed Sibte Raza Abidi
School of Computer Sciences
Universiti Sains Malaysia
Penang, MALAYSIA

Abstract - *Exploratory data mining using artificial neural networks offers an alternative dimension to data mining, in particular techniques geared towards data clustering and classification. In this paper, we argue the case for using neural networks as a viable data mining tool that can provide statistical insights and models from large data-sets. We demonstrate how Self-Organizing Kohonen Maps, an unsupervised learning neural network paradigm, can be efficaciously used for data mining purposes, in particular for data clustering applications. We show that high-dimensional data can be projected to a lower dimension and that data can be clustered together whilst preserving essential information. The Kohonen Map based data-clustering technique is applied to the 1991 World Bank social economics indicators, to show how multi-dimensional data sets can be reduced to two-dimensional (feature) maps, manifesting clusters of similar data items.*

Keywords: Neural Networks, Self-Organising Maps, Data Mining, Data Clustering, K-Means.

1. Introduction

Data mining offers a suite of algorithms, each addressing a different task and in the process elucidating a unique facet of the data [1]. Of the many facets of data mining, we are particularly interested in clustering problems, i.e. the process of finding similarities in the data and then grouping similar data into identifiable clusters.

Artificial Neural Networks (ANN) offer tremendous opportunities for performing data mining activities [2] [3], in particular problems pertaining to data classification and clustering. ANN have a natural propensity to learn–they learn how to solve problems from data as opposed to solving problems based on explicit problem specification. More attractively, the learning characteristics of ANN enable them to deal efficiently with noisy data—partial, incorrect and potentially conflicting data.

The use of ANN as data mining tools remains an area for further investigation [2] [3]. In this paper, we argue the case for using ANN as a viable data mining tool. We demonstrate the efficacy of Self-Organizing Kohonen maps (SOM) as a useful tool for the discovery of statistical insights and models from large data sets, i.e. exploratory data analysis [4] [5]. We show that by using SOM, high-dimensional data can be projected to a lower dimension representation scheme (a two-dimensional map) that can be easily visualised and understood. More attractively, the transformation leads to an automatic clustering of the data, i.e. similar data items are stored in proximity thereby forming clusters. Next, we present the U-Matrix method as a visualisation technique for demarcating the trained SOM into distinct clusters of similar data elements. Clusters emerging from a SOM are usually deemed ad hoc, we will show how traditional clustering techniques, such as K-Means, can be applied to a trained SOM to formally determine the clusters inherent in the organisation of the data items at the SOM's output layer.

2. Self-Organizing Maps (SOM) as a Data Mining Tool

The self-organising maps (SOM) introduced by Teuvo Kohonen [6] [7] are deemed as being highly effective as a sophisticated visualization tool for visualizing high dimensional, complex data with inherent relationships between the various features comprising the data. The SOM's output emphasises the salient features of the data and subsequently lead to the automatic formation of clusters of similar data items [8]. We argue that this particular characteristic of SOMs alone qualifies them as a potential candidate for data mining tasks that involve classification and clustering of data items [8].

2.1. Data Visualisation and Reduction with SOMs

A 'learnt' SOM can be used as an important visualization aid as it gives a complete picture of the data; similar data items are automatically grouped together. However, for practical purposes is still desired to demarcate the output layer of the SOM into visibly distinct clusters of similar data items. Traditionally, this is an ad hoc exercise and researchers tend to draw (by hand) boundaries dividing 'recognisable' clusters. We do not agree with this ad hoc practice and suggest the use of the U-Matrix method for drawing formal (mathematically sound) boundaries between different clusters.

The U-Matrix method [5] uses the distances between the units in a SOM as a boundary defining criteria. These distances can be then be displayed as heights giving a U-matrix landscape. Interpretation of the U-matrix is as follows: altitudes or the high places on the U-matrix will encode data that are dissimilar while the data falling in the same valleys will represent input vectors that are similar. Thus, data within the same valley can then be grouped together to represent a cluster. For illustration purposes, Figure 1 shows how the U-Matrix visualisation method can be used in conjunction with SOM to find cluster information in a data set (the rings data set). The darker hexagons represent units with a high U-Matrix value while low value units are represented by a lighter shade of grey. Through the U-Matrix visualisation two distinct regions, represented by the two black areas on the map can be seen. Each region contains all the data for one of the rings.

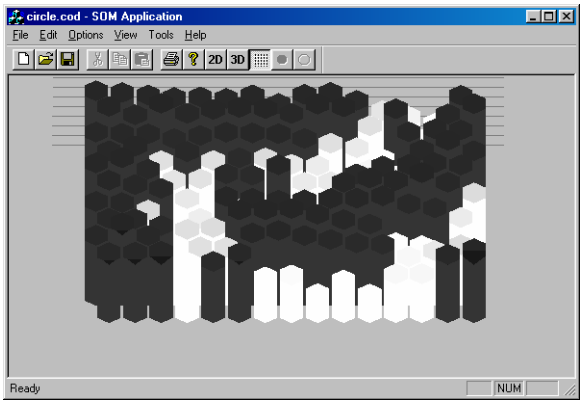

Figure 1. A U-matrix representation of the SOM trained for the *Rings* data set.

2.2. Automatic Detection of Clusters

By using the U-Matrix method, we managed to automatically cluster the trained SOM. Nevertheless, it was concluded that without any *a priori* knowledge of the classes, marking the regions based on the gray scale map is rather inconsistent and difficult. This is because (a) not all the clusters that were marked had clear walls to separate them from the other clusters and (b) small clusters are skipped or merged into the more predominant clusters.

Our conclusion is that, although the U-matrix can help greatly to choose the boundaries, but in sum it can be only regarded only as a data visualisation technique as much of the interpretation of the U-matrix values are subjective. To circumvent this situation we applied traditional clustering methods such as

hierarchical clustering, K-means clustering, K-nearest neighbor clustering, to a trained SOM.

3. Data Mining Using SOMs

To illustrate the efficacy of SOMs as a viable data mining tool we performed experiments involving two different data sets. Here, we present one experiment using a data-set comprising world social indicators as collected by the World Bank. The full data contains a total of 85 indicators on 202 different countries. Data Mining involves the learning of the data-set, followed by mining the learnt SOM vis-à-vis finding the emergent data clusters.

3.1 Experimental Results

After the training phase we obtained an ordered SOM in which each data item is represented by a unit and similar data items are stored in proximity leading to clusters of data items. Figure 2, illustrates how the huge data set is reduced to a two-dimensional representation for easier analysis and visualization. The learnt SOM represents many countries of the same geographic location in similar parts of the map. In region (d), many African nations fall into this region. Region (e) shows a region where some of the Middle Eastern countries can be found. Region (a) and (b) shows two regions of European countries. In region (c) there are some Asian countries. Figure 2, therefore, shows that similar items are grouped together by the SOM algorithm. Hence, based on these results one can safely assume that countries around a similar geographic region share similar qualities.

Analysing the learnt SOM (in Figure 2) from a social economic perspective, we can see certain patterns when we look at different regions of the map. We can see this more clearly by looking at the different plane maps. The plane map on the GNP per capita shows countries with a high GNP per capita are represented in the area with a lighter shade. By looking at this map, we can see that the bottom left corner of the map contains countries that have a higher GNP per capita. Similarly, for the attribute life expectancy, we see that the countries that are located at the top right corner are countries that has a lower life expectancy when compared with the other countries.

Further clustering of the data was performed by applying the a-dK means method to the trained SOM (shown in Figure 3). The results indicate that the trained SOM is divided into 8 different regions. More attractively, the regions correspond to the arrows marked in figure 2. Based on these results it is our contention that in an area where the class structure is not known, the SOM clusters can be used as a basis for further discovery.

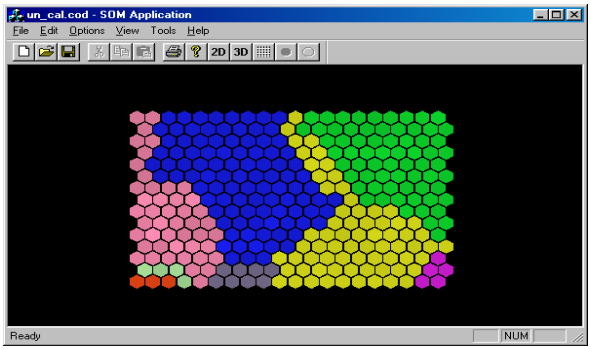

Figure 3: Data Clustering after applying the a-dK means method.

4. Conclusion

We have shown here that the SOM can be used as an effective tool for data mining, in particular for clustering applications. Our results have demonstrated the efficacy of our approach in that many countries with very similar geographic location were mapped to nearby units on the SOM, most attractively this was achieved in an unsupervised training session whereby the SOM was not directed in any way whatsoever to place similar countries together, rather the SOM

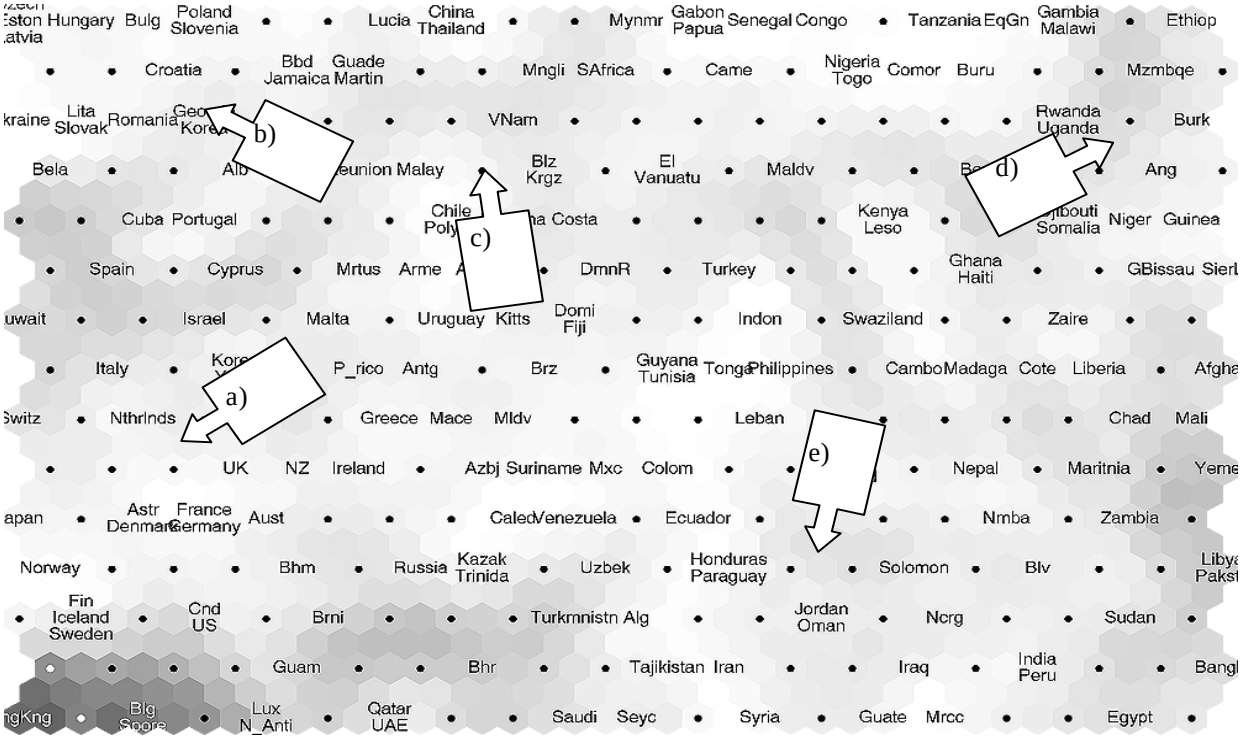

Figure 2: The SOM output after training it using the World Bank data set. The U-matrix representation (shades of grey) show the emergent clusters.

algorithm determined the similarity between various attributes and performed the clustering of similar countries. We have also shown that SOMs can also be used as a visualization tool to project a high dimensional data set to lower dimensions for visualization or for further mining efforts. The SOM along with data reduction qualities also offers the analyst different perspectives on which the analyst can view the data. In conclusion, these features validate the efficacy of SOM as a viable tool for exploratory data mining.

References

[1] U. Fayyad, G. Piatetsky-Shapiro, et al (eds.). Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park:CA, 1996.

[2] M. Craven, J. Shavlik. Using Neural Networks for Data Mining. Future Generation Computer Systems, 1997.

[3] H. Lu, R. Setiono, H. Liu. Effective Data Mining Using Neural Networks, VLDB'95 Proceedings, Springer, Singapore, 1995.

[4] S. Kaski. Data Exploration Using Self-Organizing Maps. Doctorate Thesis, Neural Networks Research Centre, Helsinki University of Technology, 1997.

[5] S. Kaski & T. Kohonen. Exploratory Data Analysis by the Self-Organizing Map: Structures of Welfare and Poverty in the World, Proceedings of the Third International Conference on Neural Networks in the Capital Markets, World Scientific, Singapore, 1995, 498-507.

[6] T. Kohonen. Self-organized Formation of Topologically Correct Feature Maps, Biological Cybernetics, 43:59-69,1982.

[7] T. Kohonen. The Self-Organizing Map. Proceedings of the IEEE, 78(9):1464-1480, 1990.

[8] J. Lampinen & E. Oja. Clustering Properties of Hierarchical Self-Organizing Maps. Journal of Mathematical Imaging and Vision, 2:261-272,1992.