TU Clausthal

ABSTRACT OF THE DISSERTATION

Matrix Learning for Topographic Neural Maps

by

Banchar Arnonkijpanich
Doctor of Natural Science in Compueter Science
Clausthal University of Technology, 2010

Since electronic data sets are increasing rapidly with respect to both, size of the data sets and data resolution, i.e. dimensionality, adequate data inspection and data visualization have become central issues of data mining. In consequence, a multitude of data organization schemes such as clustering or topographic mapping have emerged, two of the most popular methods being the self-organizing map (SOM) and neural gas (NG). SOM as proposed by Kohonen and NG by Martinetz et al. and generalizations thereof such as the generative topographic map constitute popular algorithms to represent data by means of prototypes arranged on a (hopefully) topology representing map. However, most standard methods rely on the Euclidean metric, hence the resulting clusters are isotropic and they cannot account for local distortions or correlations of data. A more general metric with K-means and fuzzy-K-means has been proposed in [**?**, **?**]. This general metric is given by a full adaptive matrix such that ellipsoidal clusters are accounted for. However, algorithms based on K-means are very sensitive to the initialization of the prototype distribution, which can lead to convergence to local minima.

In this thesis, we improve this algorithm by combining a general metric with SOM and NG such that the neighborhood based topology preservation in SOM and NG also reduces the influence of initialization. Thereby, our approach relies on a natural extension of the standard cost functions of NG and SOM (in the form of Heskes) and is conceptually intuitive. We give a strong mathematical foundation. We derive batch optimization learning rules for prototype and matrix adaptation based on the generalized cost functions and we prove convergence of the algorithm. Thereby, it can be seen that matrix learning implicitly performs local principal component analysis (PCA) and the local eigenvectors correspond to the main axes of the ellipsoidal clusters. Thus, the proposal also provides a cost function associated to alternative proposals in the literature which combine SOM or NG with local PCA models.

Since local PCA is a combination of vector quantization (VQ) and PCA, our approach can be directly applied to classification and clustering. The efficiency of the proposed method is evaluated by benchmark datasets for classification task. Image compression is used as an example of the (un)abilities of clustering for transform coding. Another focus of this work is to demonstrate the applicability of this implicit local PCA to low-dimensional data embedding for data inspection and data visualization. The proposed technique is based on matrix learning for neural gas and manifold charting. This provides an explicit forward mapping from a given high dimensional data space to low dimensionality and its inverse mapping is also estimated. Popular existing manifold learning methods such as Isomap and LLE lack such mapping. We demonstrate the usefulness of the forward mapping of the proposed model in several applications to manifold visualization and structure representation of images while an approximate inverse mapping is used in problems of dynamic texture synthesis as an interesting application of manifold learning.