Available online at www.sciencedirect.com
SCIENCE DIRECT®

European Journal of Operational Research 173 (2006) 729–745

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Optimal decisions in combining the SOM with nonlinear projection methods

Jolita Bernatavičienė, Gintautas Dzemyda, Olga Kurasova, Virginijus Marcinkevičius *

Institute of Mathematics and Informatics, Akademijos St. 4, 08663 Vilnius, Lithuania

Received 5 November 2004; accepted 24 May 2005
Available online 6 December 2005

Abstract

Visual data mining is an efficient way to involve human in search for a optimal decision. This paper focuses on the optimization of the visual presentation of multidimensional data.

A variety of methods for projection of multidimensional data on the plane have been developed. At present, a tendency of their joint use is observed. In this paper, two consequent combinations of the self-organizing map (SOM) with two other well-known nonlinear projection methods are examined theoretically and experimentally. These two methods are: Sammon's mapping and multidimensional scaling (MDS). The investigations showed that the combinations (SOM_Sammon and SOM_MDS) have a similar efficiency. This grounds the possibility of application of the MDS with the SOM, because up to now in most researches SOM is applied together with Sammon's mapping. The problems on the quality and accuracy of such combined visualization are discussed. Three criteria of different nature are selected for evaluation the efficiency of the combined mapping. The joint use of these criteria allows us to choose the best visualization result from some possible ones.

Several different initialization ways for nonlinear mapping are examined, and a new one is suggested. A new approach to the SOM visualization is suggested.

The obtained results allow us to make better decisions in optimizing the data visualization.

Keywords: Neural networks; Visual interactive modelling; Optimization; SOM; Visual data mining

* Corresponding author.
E-mail addresses: JolitaB@ktl.mii.lt (J. Bernatavičienė), Dzemyda@ktl.mii.lt (G. Dzemyda), Kurasova@ktl.mii.lt (O. Kurasova), VirgisM@ktl.mii.lt (V. Marcinkevičius).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.05.030

1. Introduction

Data points from real world are often described by an array of parameters, i.e., we deal with multidimensional data. These data points form a data set. The problem is to discover knowledge in the set of multidimensional points. There are a variety of possible data mining methods for the analysis (clustering, classification, visualization, etc.). A proper choice of method depends, e.g., on the goals of analysis, data structure, data amount, etc. Visualization is a powerful tool in data analysis. It makes easier the perception of data. The classic methods for visualization are, for example, Sammon's mapping [33], multidimensional scaling (MDS) [3], principal components [25], and others. Representatives of the modern visualization methods are neural networks – the self-organizing map (SOM) [20,21] and neural network-based realizations of Sammon's mapping (SAMANN) [28]. At present, combinations of the classic methods with modern ones (in particular, with neural networks) are under the rapid development.

In this paper, we investigate the consequent combinations of the SOM with Sammon's mapping or multidimensional scaling (MDS). The problems on the quality and accuracy of such visualization are discussed in this paper. Also, the best way of initialization for nonlinear mapping is suggested. When comparing different visualization methods, the problem arises to evaluate the quality of projection. Each method optimizes its own criterion of the quality or error. In this paper, a set of criteria that describe the projection quality is presented and applied.

This paper is organized as follows. In Section 2, a survey of basic data mapping algorithms is presented, some methods (SOM, MDS and Sammon's mapping) are analysed in detail. In Section 3, the ways of combining the visualization methods are investigated. In Section 4, the problems of initialization of nonlinear projection methods are discussed. In Section 5, a set of criteria of the mapping quality is described. In Section 6, the results of analysis are presented. The final section summarizes the conclusions.

2. The basic data mapping algorithms

There exist a lot of methods that can be used for reducing the dimensionality of data, and, particularly, for visualizing the n-dimensional vectors $X_1, \ldots, X_s \in R^n$, where s is the number of the vectors. A deep review of the methods is performed, e.g., by Kaski [19] and Kohonen [20,21]. The discussion below is based mostly on these reviews. The discussion shows the place of our approaches (consequent application of the self-organizing map and Sammon's mapping or MDS) in the general context of methods for reducing the dimensionality of data.

The first group of methods is so-called projection methods. The goal of projection is to represent the input data items in a lower-dimensional space so that certain properties of the structure of the data set were preserved as faithfully as possible. The projection can be used to visualize the data set if a sufficiently small output dimensionality is chosen.

One of these methods is a well-known principal components analysis (PCA) [25]. The PCA can be used to display the data as a linear projection on such a subspace of the original data space that preserves the variance in the data best. Effective algorithms exist for computing the projection. Even neural algorithms exist (see e.g., [30,32]). PCA cannot embrace nonlinear structures, consisting of arbitrarily shaped clusters or curved manifolds, since it describes the data in terms of a linear subspace. Projection pursuit [11] tries to express some nonlinearities, but if the data set is high-dimensional and highly nonlinear, it may be difficult to visualize it with linear projections onto a low-dimensional display even if the "projection angle" is chosen carefully.

PCA can be generalized to form nonlinear curves. While a good projection of a data set onto a linear manifold was constructed in PCA, the goal in constructing a principal curve is to project the set onto a nonlinear manifold. The principal curves [16] are smooth curves that are defined by the property that each