Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 18 (2013) 40 – 49

Procedia Computer Science

2013 International Conference on Computational Science

Interactive data mining by using multidimensional scaling

Piotr Pawliczek[a,b], Witold Dzwinel*[b]

[a]University of Texas, Department of Biochemistry and Molecular Biology, Houston, TX 77030, USA
[b]AGH University of Science and Technology, Department of Computer Science, Al.Mickiewicza 30, 30-059 Kraków, Poland

Abstract

Blind choice and parameterization of data mining tools often yield vague or completely misleading results. Interactive visualization enables not only extensive exploration of data but also better matching of clustering/classification schemes to the type of data being analyzed. The multidimensional scaling (MDS), which employs particle dynamics to the error function minimization, is a good candidate to be a computational engine for interactive data mining. However, the main disadvantage of MDS is both its memory and time complexity. We developed novel SUBSET algorithm of a lower complexity, which is competitive to the best, currently used, MDS algorithms in terms of efficiency and accuracy. SUBSET employs reduced dissimilarity matrix, which structure allows for efficient usage of both multi-core CPU and SIMD GPU processor architectures. Consequently, SUBSET enables visualization of datasets consisting of an order of 10^5 data items on a standard personal computer or laptop. We compare a few strategies of dissimilarity matrix reduction and we present typical timings obtained by respective MDS algorithms on selected multithread CPU and GPU architectures.

Keywords: multidimensional scaling; method of particles; incomplete distance matrix; multicore CPU; GPU.

1. Introduction

Data mining of large datasets consisting of data items $\boldsymbol{O}_i \in \boldsymbol{\Omega}$, ($i$=1,…,$M$), where $\boldsymbol{\Omega}$ is an abstract data space, involves application of many machine learning tools such as classifiers, regression and clustering schemes. Many of them are specialized for analysis of a special case of $\boldsymbol{\Omega}$ represented by $\mathbf{Y}$ space of multidimensional

* Corresponding author. tel.: +48 662130188, fax: +48 12 617 51 72
E-mail address: dzwinel@agh.edu.pl

doi:10.1016/j.procs.2013.05.167

feature vectors $\mathbf{y}_i$, $\mathbf{Y}=\{\mathbf{y}_i=(y_{i1},..y_{iN})\}_{i=1,...,M}$, where $N=\dim\mathbf{Y}$ (see Fig.1a). However, in general, the data items $\boldsymbol{O}_i$ can have more sophisticated structures, which cannot be directly represented by the feature vectors. This problem can be partially overcome provided that it is possible to define a dissimilarity measure $\delta(\boldsymbol{O}_i,\boldsymbol{O}_j)\rightarrow\Re^1$ between data items $\boldsymbol{O}_i$ and $\boldsymbol{O}_j$. We assume that $\delta(.,.)$ is symmetric and $\delta(\boldsymbol{O}_i,\boldsymbol{O}_i)=0$. Particularly, $\delta(,)$ can be a distance obeying also the condition of triangle inequality. In general, the $\boldsymbol{\Omega}$ space topology is represented by dissimilarity matrix $\boldsymbol{\Delta}=\{\delta_{ij}\}_{MxM}$. The vector representation of $\boldsymbol{\Omega}$ can be derived by employing multidimensional scaling (MDS) procedure [1-3].

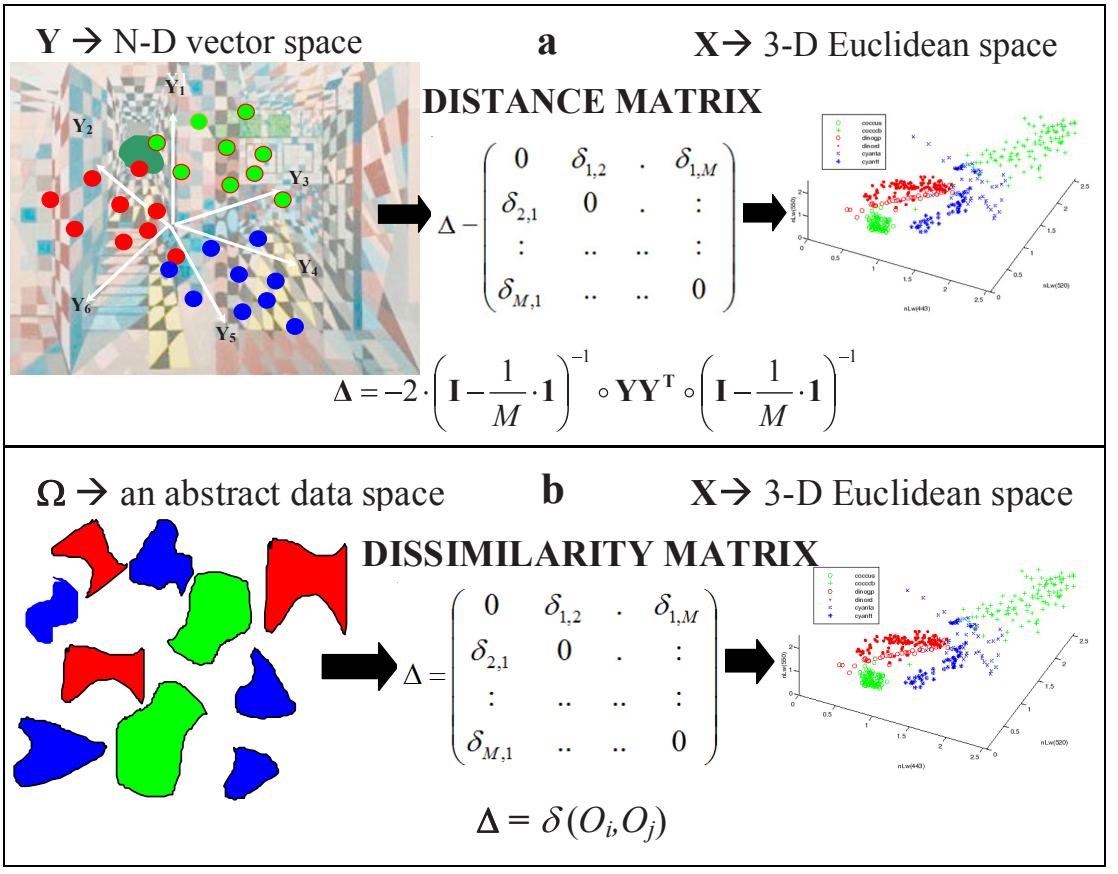

Fig.1. Multidimensional scaling applied for visualization in 3-D Euclidean space **X** data from a) a multidimensional feature space **Y,** b) an abstract $\boldsymbol{\Omega}$ space for which only dissimilarity matrix $\boldsymbol{\Delta}$ is known (e.g., dissimilarity measure between shapes).

Multidimensional scaling (MDS) (see e.g. [1,2,3]) is a bijection $B:\boldsymbol{\Omega}\rightarrow\mathbf{X}$ of a "source" space of abstract items $\boldsymbol{\Omega}=\{\boldsymbol{O}_i;\ i=1,...,M\}$ onto a "target" vector space $\Re^n\ni\mathbf{X}=\{\mathbf{x}_i=(x_{i1},..x_{in})\}_{i=1,...,M}$, where $n=\dim\mathbf{X}$, which reproduce topological structure of $\boldsymbol{\Omega}$ in **X** in respect to a given error criterion (see Fig.1b) .

Let us define the Euclidean matrix $\mathbf{d}=\{d_{ij}\}_{MxM}$ in the target vector space **X,** where d_{ij} is the Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ which correspond to $\boldsymbol{O}_i$ and $\boldsymbol{O}_j$, respectively. We assume that to preserve topological structure of $\boldsymbol{\Omega}$ in **X** an overall error $F(\|\boldsymbol{\Delta}-\mathbf{d}\|)$ should be minimized, where $F(.)$ - called the "stress" function [1-3] - is an increasing function $F:\Re^1\rightarrow\Re^1$. The value of $F(.)$ represents a discrepancy measure between dissimilarities $\boldsymbol{\Delta}$ from $\boldsymbol{\Omega}$ and corresponding distances **d** from **X**. The resulting matrix $\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots,\mathbf{x}_M)$, which minimizes the "stress" function $F(.)$, is the final outcome of multidimensional scaling. This way