# Decision Support via Big Multidimensional Data Visualization

Audronė Lupeikienė[1], Viktor Medvedev[1], Olga Kurasova[1],
Albertas Čaplinskas[1], Gintautas Dzemyda[1]

[1] Vilnius University, Institute of Mathematics and Informatics,
Akademijos str. 4, LT-08663 Vilnius, Lithuania
{Audrone.Lupeikiene, Viktor.Medvedev, Olga.Kurasova,
Albertas.Caplinskas, Gintautas.Dzemyda}@mii.vu.lt

**Abstract.** Business information systems nowadays should be thought of first of all as the decision-oriented systems supported by different types of subsystems. Multidimensional data visualization is an essential constituent of such systems, especially in the age of growing amounts of data to be interpreted and analyzed. As managers are faced with a federated environment and need to make time-critical decisions, data should be presented in a meaningful manner and easily understandable form. It is required more effective ways to cope with this situation. One of them is the visual presentation of complex data for human decisions. The paper focuses on the neural networks-based methods for visualization of big multidimensional datasets. The new strategy – to decrease the number of cycles of data reviews (passes of training data) up to the only one when training neural networks is proposed. The results of experiment on benchmark data to test this strategy are presented.

**Keywords:** data visualization, big multidimensional data, neural networks-based method, decision-oriented system.

## 1 Introduction

Characteristics of today's world, such as globalization, dynamics and often unpredictable changes, huge amounts of data, are being observed on any of its entities. Even a philosophy in general beyond business information systems (BIS) is frequently changing. Nowadays, they should be thought of first of all as the decision-oriented systems supported by different types of subsystems. Multidimensional data visualization is an essential constituent of such systems because this approach enables to discover knowledge hidden in big datasets. As managers are faced with a federated environment and a need to make time-critical decisions, data should be presented in a meaningful manner and easily understandable form. However, when datasets are becoming increasingly large we require more effective ways to process, analyze and interpret these data.

Copyright © 2015 by the authors. Copying permitted for private and academic purposes.
This volume is published and copyrighted by its editors.

We focus on neural networks-based methods for the visualization[1] of big multidimensional datasets. Two unsupervised learning methods are considered: SAMANN (a feed-forward neural network to learn Sammon's mapping) and SOM (Self-Organizing Map). To cope with data processing time problem the new strategy – to decrease the number of data passes (reviews) up to the only one when training neural networks is proposed. It is based on the assumption that huge amount of data includes many similar objects, so even in one pass, the neural network can see big amount of similar objects. After the training, network can be used for decision support – any number of new objects can be converted to meaningful form, i.e. presented as points on the plain.

Empirical research was carried out. To test the hypothesis, in other words, the proposed strategy, the experiment on a set of benchmark data was conducted. The additional evaluation of the outcome was done using the results of traditionally trained neural networks.

The structure of this paper is as follows. We start out, in Section 2, by positioning business information systems and their mission from today's perspective. Section 3 considers data visualization – describes the strategy of training neural networks through a single pass of training data and its experimental investigation. Section 4 concludes the paper and comments on open issues.

## 2 Decision Support as a Primary Aspect of BIS

The concept of information system[2] has significantly changed throughout it's more than 50 years history. These distinctions reflect its role and importance in a business enterprise and can be seen on the development approaches, methodologies, frameworks, architectural design decisions, technology.

At the very beginning, namely management information systems (MIS) served the business management. Their purpose was to cater to the information needs for planning, controlling and decision making. MIS is dependent on underlying transaction processing systems, but in fact, can itself be thought of as a transaction processing system, which possibly interacts with a decision-support subsystem. From technological point of view it was a set of applications centered around a database.

This philosophy has changed at the very beginning of the 21th century when it was realized that an enterprise system should be developed as a whole [1], [2], [3]. Information system (IS) here refers to a real world system which provides information services required to support business. It is a component of enterprise system and it should be aligned with business goals and mission, thus behaving as critical success factor. Consequently, the whole enterprise system is viewed as a three-layered system: business systems, information systems, and supporting software[3]. Application

[1] The term *visualization* means *dimensionality reduction* in this paper.

[2] We use the terms *information system* and *business information system* as synonymous in this context.

[3] There was one more result – separation of concerns, i. e., information processing and technology was clearly separated.

systems support or fully perform the information processing processes or only its parts.

Nowadays, information system should be thought of first of all as the decision-oriented system supported by different types of subsystems. The previous concept of IS cannot dominate, as manufacturing enterprises as a rule focuses on their core competences, so unable to produce alone. Decision-making is federated and synchronized between different divisions, within or between enterprises, to achieve total and autonomous optimization [4], [5]. Therefore, the main purpose of BIS is to suggest alternative decisions, which can be made by management, and to generate and evaluate scenarios for each alternative to describe possible impacts and consequences [6]. This can be seen when considering relation between enterprise resource planning (ERP) and advanced planning and scheduling systems (ASP) (ERP is one of the shapes of BIS). According to [4], [5], planning and scheduling process is primary aspect of decision making in manufacturing enterprises. APS system is not a part of ERP, but rather an entire planning and scheduling system within an enterprise supported by ERP system.

One of the challengers in this context is the ability to process big amounts of data in near-real time to make the decisions.

## 3 Neural Network Based Big Data Visualization Using a Single Pass of Training Data

### 3.1 Multidimensional Data Visualization

Big multidimensional data brings new challenges to data analysis because large volumes and different varieties must be taken into account. In many cases, data is just being generated faster than it can be analyzed. To analyze big data, many data mining and machine learning algorithms have been developed. We focus on dimensionality reduction algorithms which reduce data dimensionality from original high dimension space to target dimension (2D in visualization case). Data visualization is the presentation of multidimensional data in some graphical form. As more and more data should be collected and analyzed, it is very important to see analytical results presented visually, find dependences among a lot of objects.

Visualization is one of the basic operations in the toolbox of data analysts. Given a large set of some measured variables, the main idea is to represent them with a reduced set of more informative variables. Another reason for reducing the dimensionality is to reduce computational load in further processing. Today's large multidimensional datasets contain huge amount of data that becoming almost impossible to manually analyze them to extract valuable information. We require more effective ways to display, analyze and interpret the information contained within them.

Data from the real world are frequently described by an array of features $x_{1,}x_2, \dots, x_n$. Any feature may take some numerical values. A combination of values of all features characterizes a particular data object $X_j = (x_{j1,}x_{j2}, \dots, x_{jn}), j \in$

$\{1, \ldots, m\}$ from the whole set $X_1, X_2, \ldots, X_m$, where $n$ is the number of features, $m$ is the number of analyzed objects. If $X_1, X_2, \ldots, X_m$ are described by more than one feature, the data are called multidimensional data. Often the objects are interpreted as points in the $n$-dimensional space $R^n$, where $n$ defines the dimensionality of the space. In fact, we have a table of numerical data $\{x_{ji}, j = 1, \ldots, m, i = 1, \ldots, n\}$ for the analysis. An intuitive idea is to present multidimensional data, stored in such a table, in some visual form. It is a complicated problem considered by many researchers, as solution allows the human to gain a deeper insight into the data, draw conclusions, and directly interact with the data. A type of multidimensional data visualization is based on dimensionality reduction. The goal of dimensionality reduction is to represent the input data in a lower-dimensional space so that certain properties (e. g., clusters, outliers) of the structure of this dataset were preserved as faithfully as possible.

### 3.2 Related Works

A comprehensive review of the dimensionality reduction methods is presented in [7], [8]. Principal Component Analysis (PCA) [9] is one of the well-known dimensionality reduction methods. It can be used to display the data as a linear projection on such a subspace of original data space that best preserves the variance of the data. The PCA cannot preserve nonlinear structures, consisting of arbitrarily shaped clusters or curved manifolds, since it describes the data in terms of a linear subspace. An alternative approach to dimensionality reduction is Multidimensional Scaling (MDS) [10]. MDS is a classical approach that maps an original high dimensional space to a lower dimensional one, but does it in such a way that the distances of corresponding data points are preserved. The starting state of MDS is a matrix consisting of the pairwise dissimilarities of data points.

The effectiveness of PCA is limited by its global linearity. The MDS method is nonlinear method, however, unsuitable for large datasets: it requires too much computational resources. Therefore, the combinations of different data visualization methods are under active development today. The combination of different methods can be applied to make more efficient data analysis, while minimizing the shortcomings of individual methods.

Artificial neural networks (ANNs) may also be used for dimensionality reduction and data visualization. The MDS got some attention from the neural network researchers [11], [12]. As a result, several neural networks based methods for the visualization of big multidimensional datasets have been proposed, including SAMANN [7], [8], [12] and SOM [8], [13].

The most ANN based visualization methods are unsuitable for large datasets due to the demand of huge computational resources. One possible solution employs the hardware – increased memory, parallel processing and grid computing. The second solution is to go the other way and to develop more mature neural networks based visualization theory. Therefore the new strategies, approaches and methods for training artificial neural networks are required.

### 3.2.1 Visualization Methods Based on Neural Networks

A particular case of the metric MDS method is Sammon's mapping. It tries to optimize a projection error that describes how well the pairwise distances in a dataset are preserved. The application of original Sammon's mapping is not suitable for large datasets. Another disadvantage of this method is that the whole mapping procedure has to be repeated when a new data point has to be mapped. The back propagation-like learning rule (called SAMANN rule) [7], [8], [12] has been developed to allow a feed-forward artificial neural network to learn Sammon's mapping in an unsupervised way. This neural network is able to project new points after the training. In each learning step, two objects are given to this neural network. The weights of neural network are updated according to the update rule using the error measure. One training iteration of the neural network is completed if all possible pairs of objects from the dataset are shown to the neural network. After training, the network is able to project previously unseen data using the obtained generalized mapping rule.

The self-organizing map (SOM) is a class of neural networks that are trained in an unsupervised way using a competitive learning [8], [13]. A distinctive characteristic of this type of neural networks is that they can be used for both clustering and visualization of multidimensional data. SOM is a set of neurons, connected to one another via a rectangular or hexagonal topology. Each neuron is defined by the place in SOM and by the so-called codebook vectors. After SOM learning, the data $X_1, X_2, \dots, X_m$ are presented to SOM and winning neurons for each object are found. In such a way, the objects are distributed on SOM and some data clusters can be observed. Besides, according the position on the grid, the neurons are characterized by $n$-dimensional codebook vectors. An intuitive idea is to apply the dimensionality reduction methods to additional mapping of the codebook vectors of the winning neurons on the plane. MDS may be used for such the purposes. Moreover, the number of winning neurons is smaller than the number of data points, so smaller dataset should be visualized by MDS than, in the case, when the whole dataset is processed by MDS. This distinctive characteristic of SOM is very useful for big multidimensional data visualization.

## 3.3 ANN Training by Big Data: Strategy of a Single Pass of Training Data

To visualize big multidimensional data using SAMANN and SOM a new strategy for training these networks is proposed. The advantage of this strategy is that the network can be trained to visualize the multidimensional data through a single pass of training data. After the training, the network can be used for visual presentation of the desirable number of multidimensional objects on the plain. The strategy is based on the assumption that huge amount of data includes many similar objects, so even during one pass, the neural network can see big amount of similar objects.

A new strategy of big multidimensional data visualization using SAMANN is presented in Fig. 1: (1) training of SAMANN neural network using single pass (only 1 iteration); calculating its weights; (2) graphical presentation (visualization) of the dataset; (3) graphical presentation of new previously unseen points using calculated weights without additional neural network training.

A new strategy using SOM is presented in Fig. 2: (1) training of the SOM neural network using single pass; SOM winning neurons are calculated; (2) visualization of two-dimensional points that are two-dimensional representations of the codebook vectors of the winning neurons by MDS; (3) graphical presentation of the dataset; (4) graphical presentation of new previously unseen objects using the winning neurons by MDS without additional SOM training.

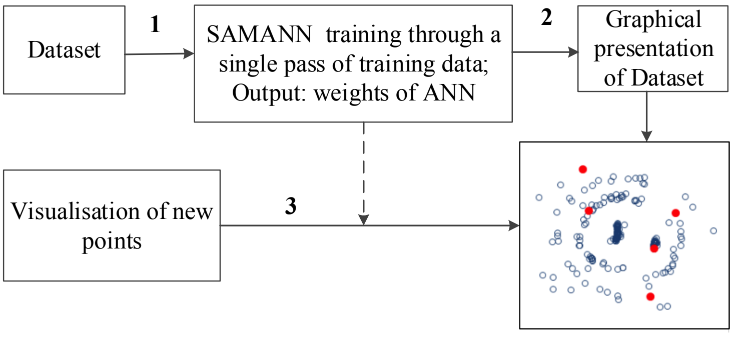
Dataset
1
SAMANN training through a single pass of training data; Output: weights of ANN
2
Graphical presentation of Dataset
Visualisation of new points
3

**Fig. 1.** A new strategy of big multidimensional data visualization using SAMANN.

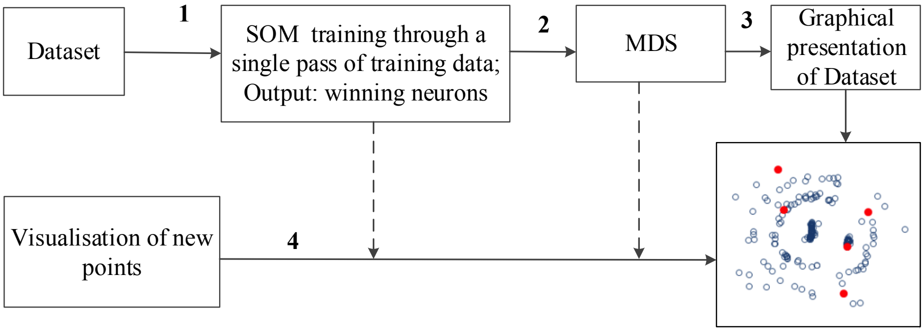
Dataset
1
SOM training through a single pass of training data; Output: winning neurons
2
MDS
3
Graphical presentation of Dataset
Visualisation of new points
4

**Fig. 2.** A new strategy of big multidimensional data visualization using SOM.

Ellipsoid dataset have been used to investigate the ability to visualize big multidimensional dataset using SAMANN and SOM. The ellipsoidal dataset consists of 7354 10-dimensional points from 10 overlapping ellipsoidal-type clusters. In the experiments, dataset, obtained using the ellipsoidal cluster generator [14], is used. The results of the experiments of multidimensional data visualization through a single pass of training data are presented in Fig. 3a and 4a. The points of the dataset are marked by black triangles. The circles correspond to the new points that were not used for training. For additional validation, the results of traditionally trained neural networks are presented in Fig. 3b and 4b.

The comparison of the results using SAMANN and SOM shows that it is possible to get the suitable projections of the primary dataset using single pass of training data and to visualize new points. The experiments show that even after one pass rather appropriate projection can be obtained.

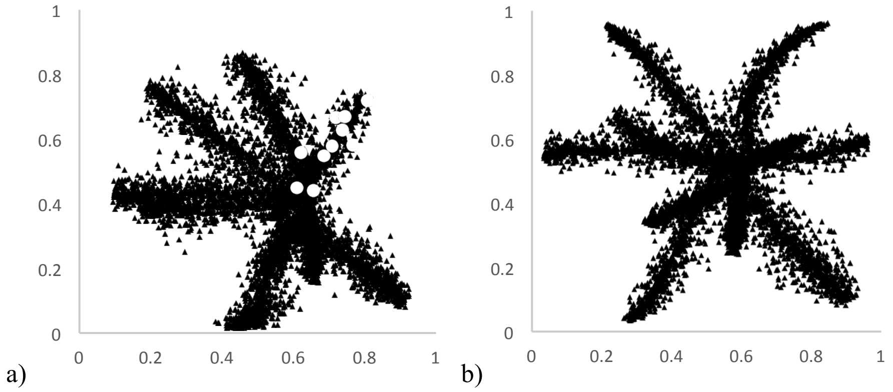

**Fig. 3.** a) Visualization results of the dataset and new points using SAMANN through a single pass of training data; b) Visualization results of the dataset using 10000 iterations.

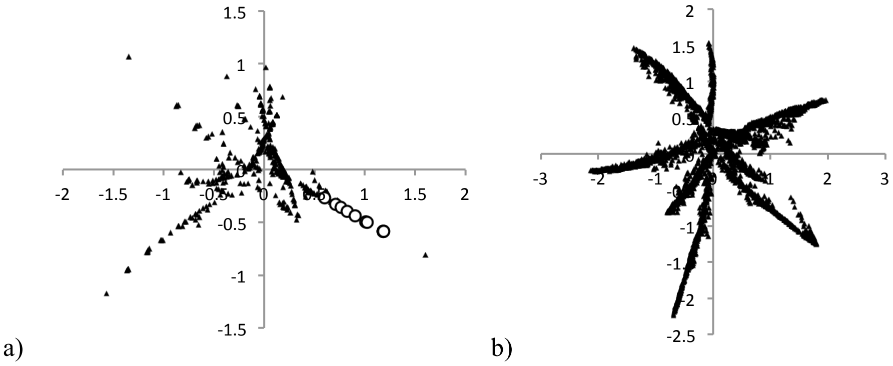

**Fig. 4.** a) Visualization results of the dataset and new points using SOM through a single pass of training data; b) Visualization results of the dataset using 100 epochs.

## 4 Conclusion and Future Research

Multidimensional data visualization is an essential constituent of business information systems, especially in the age of growing amounts of data to be interpreted and analyzed. Therefore, it is required the more mature neural networks-based visualization theory.

The new strategy to decrease the passes of training data up to the only one when training neural networks is proposed and examined. The results of experiment on the benchmark data to test this strategy allow us to conclude that the unsupervised learning of SAMANN and SOM neural networks are effective in producing the visual projections of the big multidimensional data, where we do not need any additional knowledge on the objects – the known numerical values of the features are sufficient. The obtained visualization results are good and computational expenses are acceptable if compare with traditional learning when a lot of iterations are required.

Further research should be focused on the theoretical background of such single pass strategy as well as on the discovering new domains (e. g., streaming data analysis [15]) where big multidimensional data are required to be visualized when reaching proper human decisions.

## Acknowledgment

This work has been supported by the project “Theoretical and Engineering Aspects of e-Service Technology Development and Application in High-Performance Computing Platforms” (No. VP1-3.1-ŠMM-08-K-01-010) funded by the European Social Fund.

## References

1. Maes, R., Rijsenbrij, D., Truijens, O., Goedvolk, H.: Redefining Business – IT Alignment through a Unified Framework. Report 2000-19, Amsterdam, Universiteit van Amsterdam, Department of Information Management (2000)
2. Caplinskas, A., Lupeikiene, A., Vasilecas, O.: Shared Conceptualisation of Business Systems, Information Systems and Supporting Software. In: Haav, H.-M., Kalja, A. (eds.) Databases and Information Systems II, pp. 109-120. Kluwer Academic Publishers (2002)
3. van Eck, P., Blanken, H., Wieringa, R.: Project GRAAL: Towards Operational Architecture Alignment. International Journal of Cooperative Information Systems 13(3), 235-255 (2004)
4. Nishioka, Y.: Advanced Planning and Scheduling (APS) Conceptual Definition and Implementation. White Paper, PSLX Consortium (2005)
5. Kristianto, Y., Helo, P., Mian, A.: Value Chain Re-Engineering by the Application of Advanced Planning and Scheduling. In: Gunasekaran, A., Sandhu, M. (eds.) Handbook on Business Information Systems, pp. 147-187. World Scientific Publishing Company (2010)
6. Lupeikiene, A., Dzemyda, G., Kiss, F., Caplinskas, A.: Advanced Planning and Scheduling Systems: Modelling and Implementation Challenges. Informatica 25(4), 581-616 (2014)
7. Dzemyda, G., Kurasova, O., Medvedev, V.: Dimension Reduction and Data Visualization Using Neural Networks. In: Maglogiannis, I., Karpouzis, K., Wallace, M., Soldatos, J. (eds.) Emerging Artificial Intelligence Applications in Computer Engineering. Frontiers in Artificial Intelligence and Applications, vol. 160, pp. 25-49. IOS Press (2007)
8. Dzemyda, G., Kurasova, O., Žilinskas, J.: Multidimensional Data Visualization: Methods and Applications. Springer, Heidelberg (2013)
9. Jolliffe I.T.: Principal Component Analysis. Springer, Heidelberg, (2002)
10. Borg, I., Groenen, P.: Modern Multidimensional Scaling: Theory and Applications. Springer, Heidelberg (2005)
11. Lowe, D., Tipping, M.E.: Feed-Forward Neural Networks and Topographic Mappings for Exploratory Data Analysis. Neural Computing and Applications 4, 83-95 (1996)
12. Mao, J., Jain, A.K.: Artificial Neural Networks for Feature Extraction and Multivariate Data Projection. IEEE Transactions Neural Networks 6, 296-317 (1995)
13. Kohonen, T.: Self-organizing Maps. Springer, Heidelberg (2001)
14. Handl, J., Knowles, J.: Cluster Generators for Large High-dimensional Data Sets with Large Numbers of Clusters, http://personalpages.manchester.ac.uk/mbs/julia.handl/generators.html
15. Bernatavičienė, J., Dzemyda, G., Bazilevičius, G., Medvedev, V., Marcinkevičius, V., Treigys, P.: Method for Visual Detection of Similarities in Medical Streaming Data. International Journal of Computers Communications & Control 10(1), 8-21 (2015)