Annales UMCS Informatica AI 4 (2006) 198-203

Annales UMCS
Informatica
Lublin-Polonia
Sectio AI

http://www.annales.umcs.lublin.pl/

# Visualization as a method for relationship discovery in data

Halina Kwaśnicka*, Urszula Markowska-Kaczmar, Jacek Tomasiak

*Department of Computer Science, Wrocław University of Technology,*
*Wybrzeże S.Wyspiańskiego 27, 50-370 Wrocław, Poland*

**Abstract**

Visualization techniques are especially relevant to multidimensional data, the analysis of which is limited by human perception abilities. The paper presents a hybrid method of multidimensional data analysis. The main goal was to test the efficiency of the method in the context of real-life medical data. A short survey of issues and techniques concerned with data visualization are also included.

## 1. Introduction

Analysis of real data sets is a very difficult task for the humans. It is extremely hard when examined data are represented by multidimensional vectors (records). The human mind is limited to the perception in two or three dimensional space. Therefore visualisation techniques lie in the area of interest of the data mining domain. A visualization process can be divided into several stages (Fig. 1). Rough data have to be preprocessed to the convenient form. The data presented in the form of multidimensional vectors are scaled or normalized. The next phase is a projection from multidimensional space into two dimensional (2D) space [1,2]. We can distinguish linear (Principal Component Analysis, Principal Coordinate Analysis, Discriminant Coordinates) and nonlinear methods (Self Organizing Map – SOM, Unified Distance Matrix, autoassociative neural networks, Non-Linear Mapping – for example well known Sammon mapping [3], triangulation [4], SAMANN [5]). Coordinates of points obtained after projection are subsequently plotted giving the final image. Different techniques used in this stage are presented in [6] and [7]. The source data can be used to bring in the final image a bit of information which was lost during the projection. In Fig. 1 this process is shown as a distinction.

Our hybrid method called *MedViz* combines known visualisation techniques. Using a neural network for the Sammon mapping is a novelty of the method.

*Corresponding author: *e-mail address*: halina.kwasnicka@pwr.wroc.pl

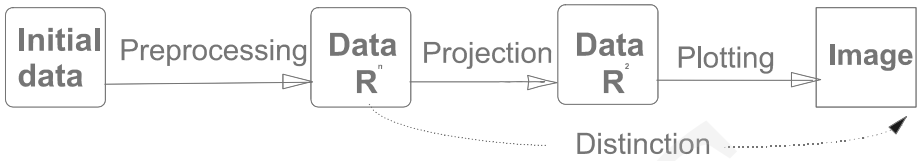



Fig 1. Stages of a visualization process

## 2. The *Medviz* method

According to the general scheme (Fig. 1), the data should be preprocessed before visualization. First, they have to be presented as a vector of numbers, with the relationship of partial order for each component. These vectors can be treated as points in the multidimensional Cartesian space ($\Re^n$).

The next step is normalization. Its aim is to obtain the distribution of values with mean $\approx$ 0.0 and variance $\approx$ 1.0 for each attribute (similar to $N(0.1)$). Owing to this process, attributes with large (e.g. thousand) and small (e.g. one) values are treated similarly. *Medviz* contains a number of popular normalization methods, the first one is *statistical normalization* in which each value is processed according to the following equation

$$x_{norm} = \frac{x - \overline{x}}{s} \tag{1}$$

where: $x$ – is the original value, $\overline{x}$ – the average value of the attribute, $s$ – the standard deviation of the given attribute. The next one is *simple scaling* to <-1.0;1.0>:

$$x_{norm} = \frac{x - \frac{1}{2}\left(x_{\max} + x_{\min}\right)}{\frac{1}{2}\left(x_{\max} - x_{\min}\right)} = \frac{2x - x_{\max} - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

where: $x$ is the original value, $x_{\min}$ – the minimal value of a given attribute and $x_{\max}$ – the maximal value of the attribute.

The next stage is the data projection by the Sammon mapping. It is nonlinear projection which retains the distance between pairs of training patterns. The general idea lies in a minimisation of the error function, defined as follows:

$$E = \frac{1}{\sum_{j=1}^{N}\sum_{i=1}^{j} D_{ij}} \sum_{j=1}^{N}\sum_{i=1}^{j} \frac{\left(D_{ij} - d_{ij}\right)^2}{D_{ij}} \tag{3}$$

where: $D_{ij}$ is the distance between $i$-th and $j$-th points in the source data, $d_{ij}$ is the distance between $i$-th and $j$-th points in the output data set, $N$ is the number of patterns in the input data set.

The above function describes the sum of the differences of distances calculated for each pair of points. Each distance is divided by $D_{ij}$ which ensures insensibility to the scaling method. For a given set of multidimensional data the output set of points in 2D space is created in this way that $i$-th point in the output data (2D space) is an image of $i$-th point in the input data (multidimensional space). Mostly the classic euclidian distance is used. In successive iteration, points in the space move in this way that the value of the error is minimised. A simple Newton gradient method is usually applied in this process [3].

After the projection, the image of the data is created by the simplified method *Glyph Plot*. Each element of the data set is represented as one *object* (point). The object has a form of a circle with a variable diameter and colour. These parameters represent the value of the same component (attribute). The component used for a distinction can be interactively chosen. The *MedViz* method has two possibilities of the point colouring: in a temperature scale of colour (blue-green-yellow-orange-red) or in a grey scale.

In *MedViz*, for the point in the image indicated by a mouse, it is possible to obtain a bit of information, like a label and an average value (of the actual selected attribute). The selection of point and click by the mouse gives information about the neighbourhood. Double click switches the user to the table with data. In this case it is possible to get a quick look in the values of attributes of the selected record. To facilitate an analysis of a huge data set it is also possible to zoom the selected fragment of image. This function helps in the case of some small data sets. Nonuniform distribution in the space causes that the analysis can be difficult. An enlargement of the part containing interesting subset of points makes the observation much easier.

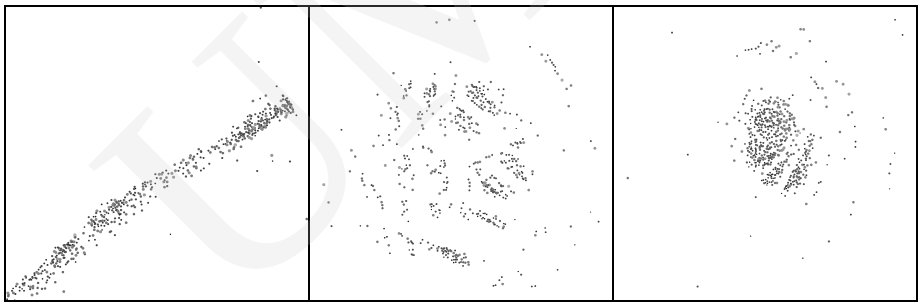
Because of the complexity of class $O(N^2)$ of Sammon mapping, the time of calculation for huge data sets becomes unacceptable. The similar situation arrives when a new record (vector) has to be visualised. Some solutions of this problem are described in ([8]).

In our method we use a simple back propagation (BP) neural network (NN) with momentum. In the output layer of the NN a linear activation function is used because new values can lie outside the range <-1.0;1.0>. At the beginning, a part of original multidimensional vectors is projected to two dimensional space according to the Sammon mapping method. They constitute a training set for the NN realizing Sammon mapping. The error of the mapping is calculated as the sum of distances of the output vectors from the training patterns. After training the rest of original multidimensional vectors, as well as new vectors when necessary, are projected by the NN. The new (added) points are visualised as empty circles. The unknown components are automatically supplied according to the method described in the previous section. They are distinguished in the image by the colour besides the assumed scale.

## 3. Experimental studies

The aim of the experimental studies was the evaluation of an efficiency of the implemented method in data analysis in order to find relationships between attributes. Additionally, the influence of different methods of normalization was investigated. If there is no other explanation, the following values of parameter are set: number of iteration of Sammon mapping – 500, simple scaling as a normalization method, parameters of NN: number of layers – 2, min error – 0.4%, momentum – 0.9, learning coefficient – 0.001, max number of iteration – 3000.

We have experimented with the three data sets: *y14c* – synthetic data set [9], *benthic* [7] – it contains information about the water purity in Chesapeake Bay and *cervix* – it has information about patients with carcinoma of the cervix uteri. Fig. 2 shows the results obtained for *cervix* appropriately (a) without normalization, (b) with simple scaling, (c) with statistical normalization. The *cervix* data set in Fig. 2 is a set of real data, which was not specially prepared before.



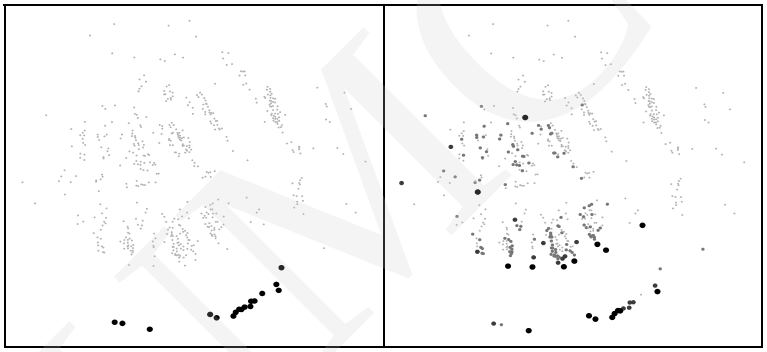
(a) (b) (c)

Fig. 2. The comparison of the normalization methods (data set *cervix*)

It is easy to notice that there is big a difference in the plots obtained with the different normalization methods. The values of attributes differ a lot which can be observed in Fig. 2.a for the data not normalized. The oblong shape of points means that the values of one attribute significantly differing from other values have dominated the projection process. The plot obtained by simple scaling of the set of points (Fig. 2b) has more groups than after statistical normalization (Fig. 2c). At the same time statistical normalization demonstrates better capability to show outliers. In general, we can say that the only way to choose the method of normalization is to experiment with them.

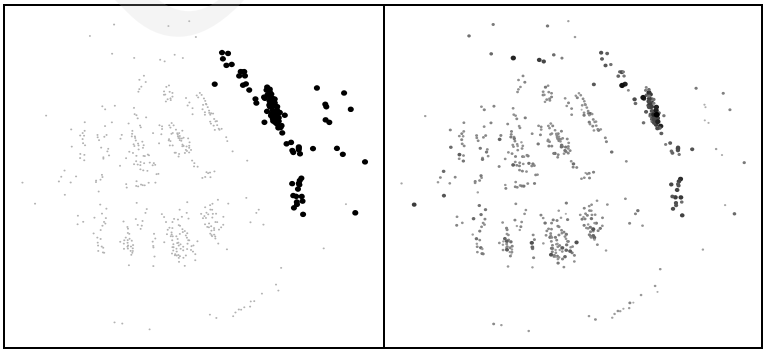
The successive experiments tested the relationship between geometric distribution of points in the two dimensional space and the values of attributes. The criterion of colouring was joined with the values of respective attributes. We

used two data sets: *benthic* and *cervix*. The obtained plots were analysed with respective attributes after the appropriate colouring of points (*the Glyph Plot method*). It was the basis for searching similarity between the distribution of points and distribution of colours in the different plots. The first two images (Figs. 3 and 4) present the data set *cervix* shown by the perspective of the attribute *P*8.2 – the application of radiotherapy (External Pelvic RT), *P*10 – the general evaluation of the therapy efficiency, *P*7.3 – the application of surgery operation (Radical Abdominal Hysterectomy without Pelvic/Paraortic Lymphadenectomy) and *P*11 – the time between the diagnosis and the discharging from a hospital.



(a) P8.2 (b) P10

Fig 3. The geometric separation (data set *cervix*, attributes *P*8.2 I *P*10)



(a) P7.3 (b) P11

Fig. 4. The geometric separation (data set *cervix*, attributes *P*7.3 and *P*11)

Fig. 3a shows a distinct group of points. It is easy to observe similar tendency with the other attributes. For example in Fig. 3b a similar group of points can be found. It represents the patients with a good efficiency of the radiotherapy (bottom part of image). In this way we can conclude that radiotherapy (External Pelvic RT) has a positive influence on the efficiency of treatment. In Fig. 4

similar relationship between the attributes *P*7.3 and *P*11 can be found. This means that the patients after the surgery operation of Radical Abdominal Hysterectomy without Pelvic/Paraortic Lymphadenectomy, need longer stay at hospital.

## 4. Conclusions

We have conducted numerous experiments using a developed method and different data sets. The results show large possibilities of visualisation in the area of data mining. Visual analysis uses much better human capabilities and very frequently leads to the conclusions which are very difficult to draw from rough data. The method and application were used in medical data analysis, but this is much universal – possible to use successfully with other data sets.

## References

[1] König A., *A survey of methods for multivariate data projection, visualisation and interactive analysis*. In Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems, (1998) 55.

[2] De Backer S., Naud A., Scheunders P., *Non-linear dimensionality reduction techniques for unsupervised feature extraction*. Pattern Recognition Letters, (1998) 711.

[3] Sammon J.W., *A nonlinear mapping for data stucture analysis*. IEEE Computer, C-18 (1969) 401.

[4] Lee R.C.T., Slagle J.R., Blum H., *A triangulation method for the sequential mapping of points from N-space to two-space*. IEEE Transactions on Computers, C-26 (1977) 288.

[5] Jain Anil K., Mao Jianchang, *Artificial neural network of nonlinear projection of multivariate data*. In IEEE International Joint Conference on Neural Networks (6th IJCNN'92), Baltimore, MD. IEEE/INNS. MI State U., III (1992) 335.

[6] Friendly M., *Statistical graphics for multivariate data*. SAS SUGI 16 Conference, (1991).

[7] McLeod A.I., Provost S.B., *Multivariate data visualization*. In El-Shaarawi A. H. and Piegorsch W. W., editors, Encyclopedia of Environmetrics, John Wiley and Sons, (2001) 1333.

[8] Pekalska E., de Ridder D., Duin R.P.W., Kraaijveld M.A., *A new method of generalizing sammon mapping with application to algorithm speed-up*. In 5th Annual Conference of the Advanced School for Computing and Imaging, ASCI'99, (1999) 221.

[9] Dembélé D., Kastner P., *Fuzzy c-means for clustering microarray data*. Bioinformatics, (2003) 973.