Pixel-based Parallel-Coordinates technique for outlier detection in Cardiac Patient Dataset

Hasimah Hj Mohamed*, Abdul Razak Hamdan , Azuraliza Abu Bakar

Dept.of Science and Management System, Faculty of Technology and Information Science, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.
Phone: (62)-22-2500995; Fax: (62)-22-2534185

Visual Exploration technique applies human visual perception to explore large data sets and have proven to be of high value in exploratory data analysis. Pixel-based visual exploration relies on basic features that human perceptual system inherently assimilates very quickly: color, size and shape. Patterns in data that indicate trends can be immediately obtained and gaps in the data can be recognized. We can also discover outliers or errors in data, determine the minimum and maximum values and identify the clusters. As a result we can have better understanding on complex systems, make better decision and discover information that might otherwise remain unknown. In this paper, we proposed a framework on pixel-based and parallel coordinate techniques in visualizing the cardiac dataset. The main ideas are to represent as many data items as possible by the pixels of the display device and identifying outliers by arranging and coloring the pixel according to the relevance for the query. The data is represented in a graphical format and it will support the medical expert to have new insights and encourage better problem solving and gains deeper domain knowledge.

1. Introduction

Visualization technology seems to provide important potentials to improve the process of querying, analyzing and understanding the data. With visualization techniques, larger amounts of data can be presented at the same time on the screen, colors allow the user to instantly recognize similarities or differences of thousands of data items, the data items may be arranged to express some specific relationships and so on.

Visual data exploration aims at integrating humans in the data exploration process, applying their perceptual abilities to the large data sets available in today's computer systems. The basic idea of visual data exploration is to present the data in some visual form, allowing the human to get insight into the data, draw conclusions, and directly interact with the data.

Different research groups in Information Visualization have proposed taxonomies for the different kinds of Visual Data Exploration. One of the approaches classifies Visual Data Exploration along three dimensions: *data type to be visualized, the technique itself* and *the interaction and distortion methods(1).*

Some visual techniques are used for the effective exploration of multidimensional data, such as geometric (e.g. landscapes (2), parallel coordinates (3)(4) and treemap (5)) and pixel-oriented (e.g. scatterplot, spiral technique (6), circle view (7)) techniques. For example, the approach presented in Keim (7) makes use of a circle-segment technique which visualizes k-dimensional data (resulting from the execution of a specific query) as a circle divided into k segments.

*Responsible author : *Email: hasimah@cs.usm.my*

In this paper we propose a pixel-based parallel coordinate visualization technique for visualizing and detecting the outliers in the large amounts of multidimensional data. We use this technique to visualize our cardiac patient data (raw data) in order to assist the data miner to decide how to preprocess the data.

This paper is organized as follows. In the next section we will discussed briefly the pixel based and parallel coordinate techniques. Section 3 will describe the use of both techniques towards cardiac patient dataset and finally, the result and conclusion are presented in section 4 and 5 respectively.

2. Visual Exploration Techniques

In the following, we briefly classify and describe the pixel-oriented and parallel coordinate technique, the visualization techniques that we have implemented on our cardiac patient dataset.

2.1 Pixel-Based oriented technique.

The basic idea of pixel-oriented techniques is to map each data value to a colored pixel and present the data values belonging to one dimension (attribute) in a separate subwindow (cf. Fig. 1) (8).

Since, in general, it use only one pixel per data value, the techniques allow us to visualize the largest amount of data which is possible on current displays (up to about 1,000,000 data values). All pixel-oriented techniques partition the screen into multiple subwindows. For data sets with m dimensions (attributes), the screen is partitioned into m subwindows–one for each of the dimensions. In the case of a special class of pixel oriented techniques – the query-dependent techniques – an additional (m + 1)th window is provided for the overall distance. Inside the windows, the data values are arranged according to the given overall sorting, which may be data driven for the query-independent

techniques or query driven for the query-dependent techniques. Correlations, functional dependencies, and other interesting relationships between dimensions may be detected by relating corresponding regions in the multiple windows.

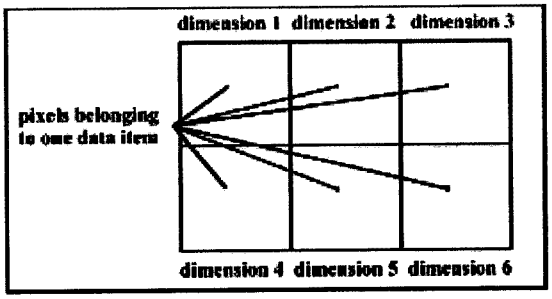

Figure 1: Basic arrangements of sub windows for data with six dimensions

Several pixel-oriented visualization techniques have been proposed (9), i.e the recursive pattern technique, the circle segment technique and spiral techniques such as Morton and Z-order technique.

2.2 Paralell coordinate Technique

Parallel coordinates were proposed by Alfred Inselberg as a new way to represent multidimensional information. Parallel coordinates (3) is a two-dimensional technique to visualise multidimensional data sets. An *n*-dimensional data tuple

$$(x1,x2,x3,\ldots,xn)$$

is visualised in parallel coordinates as a *polyline*, connecting the points x1,x2,x3,..,xn in *n* parallel *y*-axes, as can be seen in Figure 2 (10). For a large set of tuples, this technique will produce a compact two-dimensional visualisation of the whole multidimensional data set.

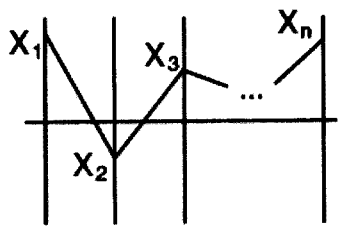

Figure 2: A tuple visualised with parallel coordinates.

By drawing the axes parallel to one another, one can represent data in much greater than three dimensions. Each variable is plotted on its own axis, and the values of the variables on adjacent axes are connected by straight lines. Thus, a point in an n-dimensional space becomes a polygonal line laid out across the n parallel axes with n-1 line segments connecting the n data values. Many such data points (in Euclidean space) will map to many of these polygonal lines in a parallel coordinate representation.

One important aspect of this visualization scheme is that it provides opportunities for human pattern recognition: by using color to distinguish lines, and by supporting various forms of interaction with the parallel coordinates system, patterns can be picked up in the given database of multidimensional data. The number of dimensions that can be visualized using this scheme is fairly large, limited only by the horizontal resolution of the screen. However, as the number of dimensions increases, the axes come closer to each other, making it more difficult to perceive patterns. It is also important to note the flexibility of the parallel coordinates approach in that each coordinate can be individually scaled: some may be linear with different bounds, while others may be logarithmic.

Parallel coordinates are generally considered to be one of the standard techniques to visualise multidimensional data sets. What is not well established yet is the interaction with parallel coordinates and especially the dynamic aspects of interacting with them.

3. The Cardiac Patient Dataset

The cardiac patient medical dataset is the information of patients admitted to National Heart Institute from the time they entered until they are discharged. This data is collected from the year 1998 to 2003. It contains of 6015 records with 195 attributes. Basically, there are four category of information in the dataset namely *Initial Consult Information*, *Anesthesia Information*, *Operative Information* and *Discharge Information*. *Initial Consult* contains information such as patient identification (*Name, NRIC, MRN*), socio demographic (*Age, Ethnicity*), risk factors (*smoking, CAD history, chronic diseases*), pre-operative medication (*type of medication*) and diagnosis (*type of heart problem*). *Anesthesia* contains information such as name of staff involved in the anesthesia procedure, bypassing and cardiopulmonary support information and anesthetic information. *Operative* contains information such as name of staff involved in the operative procedure, operative information, coronary artery bypass information and valve operative information. Finally, *Discharge* contains information such as discharge date, post-operative complication, type of blood product used, mortality information and readmission information.

For pre and post-operative complication on cardiac patients based on their risk factors information, some variables in *Initial Consult* and variables related with post-operative complication in *Discharge* are selected.

4. Visualising the Cardiac Patient Dataset

The pixel based recursive pattern algorithm is implemented onto raw cardiac patient dataset. The attributes are visualized in a separate subwindow. Within a subwindow, each attribute value is represented by one colored pixel with the color reflecting the attribute value. In order to enable the user to relate attribute values of different attributes but at the same positions, the order of the objects is reflected by the same arrangement of pixels in each subwindow.

The algorithm has been implemented using JAVA language with the ODBC/JDBC capabilities to read the data stored in the MS Access databases. We use 2D array to store the data from the database. We have enhanced the recursive algorithm to handle the 2D array data structure.

We have design the frame size of 800 x 800 for the output display. The size of each subwindow has been set as 64 x 94, which means, the maximum data item that can be visualize per attribute is 6016 records. We used eight subwindows to represent those eight selected attributes. The total data items can be displayed on this screen output is 6016 x 8 (48128 data items).

We used different color to represent different values of the attributes. The categorical attributes has been map to different colors by taking into consideration that the number of Just Noticeable Differences (JND) (11) between colors is high. Since age is a continuous attribute, we use a color scale ranging from yellow to green, green to blue, blue to maroon and maroon to brown This will enable the user to see the differences between values clearly. We use black color to represent any null (missing value) and white color to represent outlier / noisy value.

Figure 3 shows the pattern of the cardiac patient dataset with 8 subwindows, representing 8 attributes (Race, age, sex, redo, risk, operation, complication and mortality) and 6016 data items. We have execute the queries from the database by sorting the values according to certain attributes, i.e operation, risk factor and age. The pattern shows that most of the elderly patients (represented by maroon to brown color) undergo the CABG operation (green color) with a poor (yellow color) and high (red color) risk factor. Most of the patients did the CABG operation (green color), and very few did the CABG+VALVE operation (yellow color).

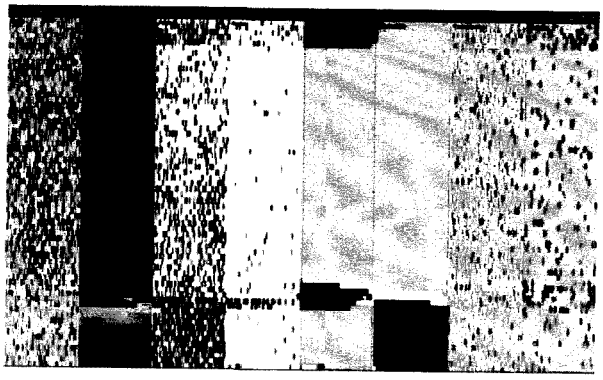

Figure 3: Pattern of Cardiac Patient Data sorted by operation, risk and age

There are also some null values (representing by black color) and outlier values (white color). This pattern has shown clearly that some of the data has missing values in age and risk attributes (represented by black color – null values), so that particular data need to be cleaned. In order to visualize the outliers and null values clearly, we have extract only the records containing of outliers and null values, as depicted in figure 4.

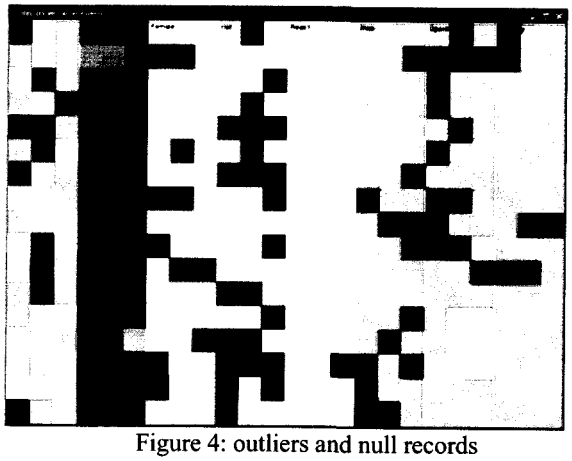

Figure 4: outliers and null records

Our pixel based visualization in figure 4 above does not show directly the relationships between attributes. Therefore, it is quite difficult for the users to see the values of individual records. We have revisualized the data using parallel coordinate technique to overcome this problem.

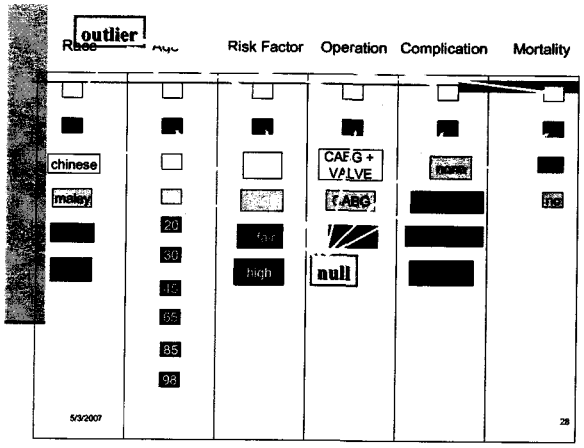

Figure 5: value arrangement for parallel coordinate

Our layout is similar to parallel coordinate layout, but the continuous axes are replaced with sets of boxes that represent the values of each attributes. Figure 5 shows the arrangement values (using different colors) of each attributes. The first two boxes represent the outliers and null values, indicated by white and black color respectively. Figure 6a and 6b shows the linking relationship between attributes for each records in the dataset. This view can help the data miner in the preprocessing task of the data.

Our future research is integrating the data cleaning algorithm with this visualization technique to help the user more understand the data in preparation of data mining process.

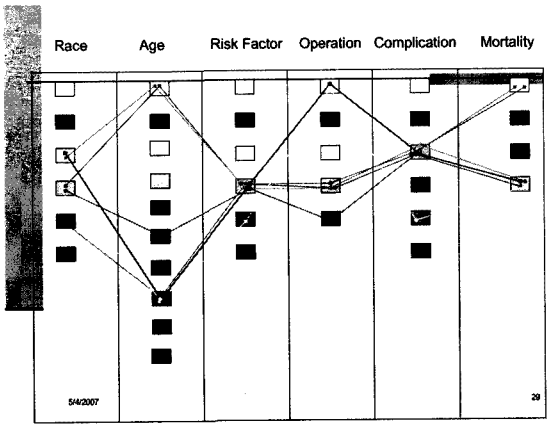

Figure 6a: layout visualization

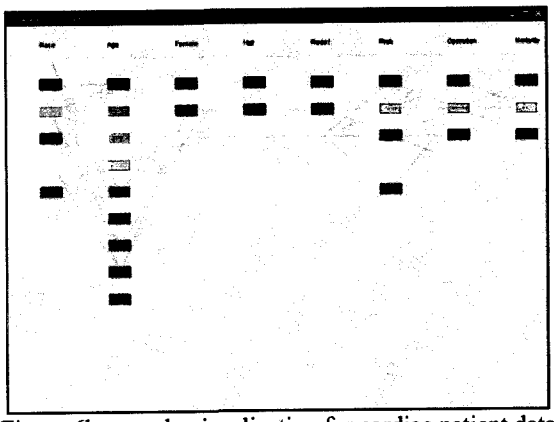

Figure 6b: sample visualization for cardiac patient data

5. Concluding Remark

This paper presented the recursive pattern and parallel coordinate technique to visualize the outliers and null records in cardiac patient dataset. The large dimension of data set (6015 x 195) make it impossible to explore the data in their numerical forms. Therefore visualization is a good and practical alternative to have better insight of the data. This research is a preliminary work of visual data mining of cardiac patient data. The patterns of the cardiac patient dataset obtained are yet to be analyzed for further investigation. However the first impression of the data can be seen in a better way as presented.

Acknowledgement.

We would like to express gratitude to the Ministry of Science, Technology and Innovation, Government of Malaysia under E-Science Fund 01-01-05-SF0087 and Universiti Sains Malaysia under short term grant for funding this research, and National Heart Institute for providing data and domain expert.

References

(1) Keim, D. A (2002)., Information Visualization and Visual Data Mining, IEEE Transactions on visualization and Computer Graphics, vol. 8, no.1, Jan-March 2002, pp 1-8.

(2) Wright, W. (1995), Information Animation Applications in the Capital Markets, *Proc. Int. Symp. on Information Visualization*, Atlanta GA, USA, pp. 19–25.

(3) Inselberg, A. and Dimsdale, B. (1990), Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry, *Visualization'90*, San Francisco CA, USA, pp. 361–370.

(4) Bendix, F., Kosara, R., Hauser, H. (2005), Parallel Sets: Visual Analysis of Categorical Data, *IEEE Symposium on Information Visualization 2005, MN*, USA, pp 133-140.

(5) Zhao, S., McGuffin M.J., Chignell, M.H. (2005), Elastic Hierarchies: Combining Treemaps and Node-Link Diagrrams, *IEEE Symposium on Information Visualization 2005, MN, USA*, pp 57-164.

(6) Keim, D. and Kriegel, H. (1994), VisDB: Database Exploration using Multidimensional Visualization, *IEEE Computer Graphics and Applications* pp. 40–49.

(7) Keim, D., Schneidewind, J., Sips, M. (2004), CircleView – A New Approach for Visualizing time-related Multidimensional Data Sets, *AVI'04, Italy* pp 179-182.

(8) Keim, D. A. (2000)., Designing Pixel-Oriented visualization techniques: Theory and applications, IEEE Transactions on visualization and Computer Graphics, vol. 6, no.1, Jan-March 2000.

(9) Keim, D. A, Kriegel H., and Ankerst, M (1995), Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data., Proc. of the 6th IEEE Visualization conference, 1995, pp 279-286.

(10) H. Siirtola, "Direct Manipulation of Parallel Coordinates", Proceedings of the International conference on Information Visualization, London, Jul. 2000, IEEE Computer Society, pp 373-378.

(11) Herman G. and Levkorwitz H.(1992), Color scales for image data, Computer Graphics and Applications, 1992, pp 72-80.