International Journal of Electrical and Computer Engineering (IJECE)
Vol. 5, No. 6, December 2015, pp. 1564~1568
ISSN: 2088-8708

Clutter Reduction in Parallel Coordinates using Binning Approach for Improved Visualization

Swathy Sunil Kumar, Teenu Krishnan, Sreeja Ashok, M.V.Judy
Department of Computer Science & I.T, Amrita School of Arts & Sciences, Kochi,
Amrita Vishwa Vidyapeetham

Article Info

Article history:

Received Apr 23, 2015
Revised Aug 11, 2015
Accepted Aug 30, 2015

Keyword:

Binning
Data visualization techniques
parallel coordinates
Multivariate data
Parametric methods

ABSTRACT

As the data and number of information sources keeps on mounting, the mining of necessary information and their presentation in a human delicate form becomes a great challenge. Visualization helps us to pictorially represent, evaluate and uncover the knowledge from the data under consideration. Data visualization offers its immense opportunity in the fields of trade, banking, finance, insurance, energy etc. With the data explosion in various fields, there is a large importance for visualization techniques. But when the quantity of data becomes elevated, the visualization methods may take away the competency. Parallel coordinates is an eminent and often used method for data visualization. However the efficiency of this method will be abridged if there are large amount of instances in the dataset, thereby making the visualization clumsier and the data retrieval very inefficient. Here we introduced a data summarization approach as a preprocessing step to the existing parallel coordinate method to make the visualization more proficient.

Corresponding Author:

Sreeja Ashok,
Department of Computer Science & I.T,
Amrita School of Arts & Sciences,
Amrita Vishwa Vidyapeetham.
Email: sreeja.ashok@gmail.com

1. INTRODUCTION

Data visualization techniques often play a vital role while representing large quantities of data, help analysing these data and landing at convincing conclusions [3]. Data visualization, when done faultlessly, proves to be an efficient means to comprehend the data that are buried in large datasets to discover the relationships, correlations, outliers and hidden patterns. It is accepted as the most efficient method for conveying the information in the dataset to the intended users with the help of graphical aids such as tables and charts. Processing, analyzing and communicating the data residing in large data sets present a variety of ethical and analytical challenges for data visualization.

Data visualization finds its applicability in different spheres such as chemical imaging, crime mapping, biological data visualization, medical imaging and so on. Chemical imaging is an analytical capability of creating a visual images of components distribution from concurrent measurement of spectral, spatial and time information. Medical imaging is the practice of creating visual illustration of the interior of a body for clinical study and medical interferences.Understanding the raw data which is residing in the large dataset is an important, yet challenging dilemma in the current situations where a large amount data gets accumulated everywhere. As the quantity of records increases, the effectiveness with which the data can be interpreted reduces. There are many visualization techniques that can be used to visualize high dimensional data such as scatter plot , glyphs, parallel coordinates, hierarchical Techniques [3].

Scatter plots are the elderly and generally used method to project high dimensional data into a two dimensional space. In this, the parallel projections are positioned in grid structure to aid the user to memorize the dimensions related with each projection. Glyphs are graphical objects that are designed to convey multiple data values. This technique can be used only when there is a limited number of a data element to be displayed simultaneously, as it may require a large amount of screen space to be viewed. Parallel coordinates are principally popular today due to its theoretical simplicity and solid appearance.

Parallel coordinates are high-dimensional data visualization technique which was invented in 1980's that represents N-dimensional data in a 2-dimensional space with mathematical rigorousness [2], [20]. It finds its applicability in diverse sets of multidimensional problems in many domains, such as WinViz, XmdvTool, and SPSS Diamond [12], [17]. It uses parallel axis for dimensions and represents N dimensional data in two dimensional spaces.Identifying the clusters in the plots is an important part of understanding and interpreting the data [4], [8]. Figure 1 represents the parallel coordinate plot of a data set which consits of 5 attributes and 4 data objects.

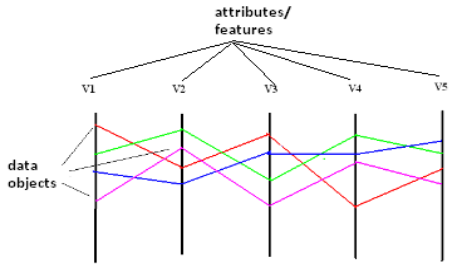

Figure 1. Parallel coordinate plot, 5 attributes and 4 data objects

In the recent years many research efforts have been directed at the display of large data sets as well as in the area of parallel coordinates. Different enhancement techniques can be in-cooperated with the parallel coordinate for better and improved data visualization. Use of colors, animation,3D viewing, and brushes, all can help us in the better understanding of the data [2], [11]. Parallel coordinate plot, accompanied by scatter plot and the radar chart has been extensively worn for visualizing multivariate datasets [1]. The use of colors and opacity can enhance the visualization. Highlighting the specific data can improve the visual understanding and thereby increase the visual clarity. Allen R Martin studied the high dimensional brushing of multivariate data, it is an operation established in the visualization systems to interactively select the subsets from the original dataset. He described N-dimensional brushes which are defined on data space. It provides advancement to the Xmdv Tool which can be used for data visualization by providing a highlighting operation to the user by using a single brush [18]. It also provided Xmdv tool with a range of methods of brush specification as well as manipulation. Jimmy Johansson studied different methods in which a data tuple with n dimension can be represented as polylines connecting n points [20]. It is a space efficient as well as an interactive method to represent a large data set [9]. Here clustering algorithm is used in combination with the parallel coordinate's methodology to represent the data. High precision textures are used for better visualization and the clusters are highlighted in different colors.

Parallel coordinates have been proved to be an efficient tool due to its efficiency in pointing out the similarity between each attributes, but efficiency reduces due to polylines and over plotting [7]. Due to the clutter impression and interference with crossing lines, the operation such as selection as well as data clustering becomes a challenging problem with respect to large dataset.The presence of polylines reduces the visibility of hidden patterns in large data set. Data reduction techniques improve the visualization and decision making by retaining the pair wise correlation between each attribute values [13]. Here we are proposing a data centric approach for data summarization to reduce the effect of clutters where the grouping is done ahead of pattern generation. This improves the comparison of individual characteristics of each data object with respect to the complete data set.

2. RESEARCH METHOD

In the proposed methodology we attempt to reduce the shortcomings of parallel coordinate primarily over plotting, by combining binning methods with parallel coordinates to boost the efficiency of data interpretation. This is achieved by adding binning as a preprocessing step to the normal parallel coordinate approach. Figure 2 depicts the work flow of the proposed system.

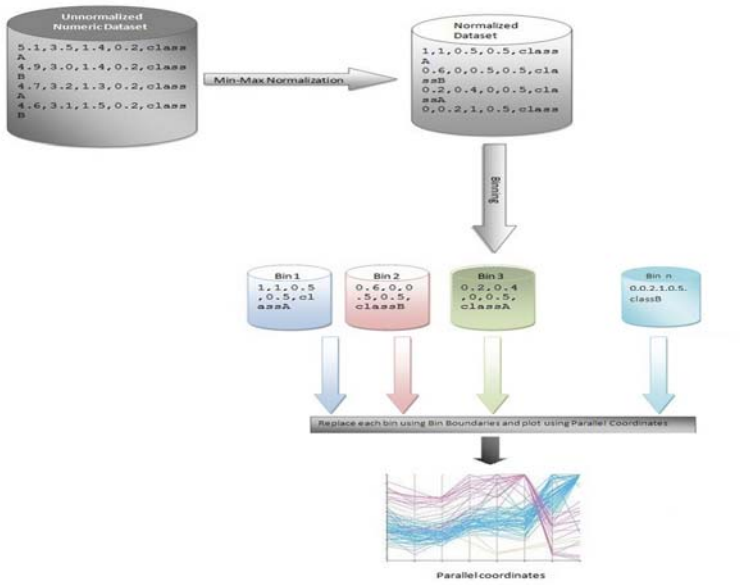

Figure 2. Work flow of the proposed system

The following steps are involved in the proposed system

1) Data Normalization
2) Binning using parametric models
3) Visualization using parallel coordinates

2.1. Data Normalization

To avoid dependence on the selection of measurement unit, the data should be normalized. It involves the linear transformation of data to fall within a common range. Through normalization, all attributes will be given an equal weight. Many normalization methods are available such as min max normalization, Z-score normalization, decimal scaling etc.Here min–max normalization method is being used to get a positive range of values with in a limit to observe the variances of each attribute clearly. The minimum and maximum values are set as 0 and 1. So this makes it into a non negative range.

Z-score normalization is formulated as:

$$X' = \frac{x-\mu}{\sigma}$$

Min max normalization is formulated as given below.

$$X' = \frac{x - x\,min}{xmax - xmin}$$

2.2. Binning using Parametric Approaches

By reducing the size of the dataset, the cluttering effect can be reduced. One approach is to use data discretization by grouping data objects into intervals. Binning is a top-down tearing technique based on precise number of bins. There are different types of binning approaches: equal width binning, equal frequency binning. The proposed system uses equal width binning technique. The main challenge here is to identify the optimum bin width and bin number. Sturges' formula for normal distribution is a statistical procedure for deciding the optimum bin size which is more efficient when compared to other methods like risk minimization technique and Bayesian optimal binning [22]. According to Sturges' theory, the optimal bin size k is derived using the formula.

$$k = [\log_2 n + 1]$$

Where ‘n’ is the size of data

The number of bins, k is decided based on the size of the database. The existing dataset is replaced by smoothing using bin boundaries and the newly generated dataset is given as input for visualization using parallel coordinates. The binning approach partitions the dataset into suitable bin size to avoid cluttering and simplifies the dataset. The bins thus created are plotted which brings more clarity for further process.

2.3. Visualization using Parallel Coordinates

The input dataset can be simplified and executed very fast after the binning process. Binned parallel coordinates provides context views of the dataset rather than the focus views. The clutters can be effectively reduced through which we can easily distinguish the patterns. Each attribute is represented using a vertical line and the bin samples are highlighted as horizontal lines. The user receives immediate feedback about the characteristics of the data objects and the correlation between each attributes. The pair wise comparison of each data and comparison of variables associated with each data item is clearly differentiated using this process. Data clusters appear as dense regions which show the similarity of features associated with each data item.

3. RESULT AND DISCUSSION

We examined the effectiveness of the proposed approach through experiments on different databases such as Data_User_Modeling_Dataset_Hamdi_Tolga KAHRAMAN having 258 instances and 6 attributes, Sitka89 having 632 instances and 4 attribute, IRIS having 150 instances and 5attributes. The number of bins depends on the number of instances in the database. Number of bins are calculated based on Sturges' formula .The simulation was done using R programming language.

The proposed method provides more clarity and understanding of the dataset. The convergences are more accurate and clear to study the influence of each attribute on the outcome. The parallel coordinate representation of each dataset before and after applying the binning approach is shown in Figure 3, 4 and 5. The left image represents the traditional line based parallel coordinates and right image shows binning based parallel coordinates with 9 ,10 and 8 bins per data dimension based on the data size.

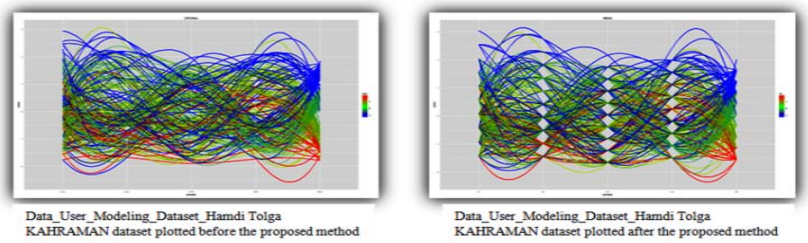

Figure 3. Comparison of two Parallel coordinate renderings in the same dataset Data_User_Modeling_Dataset_Hamdi_Tolga KAHRAMAN

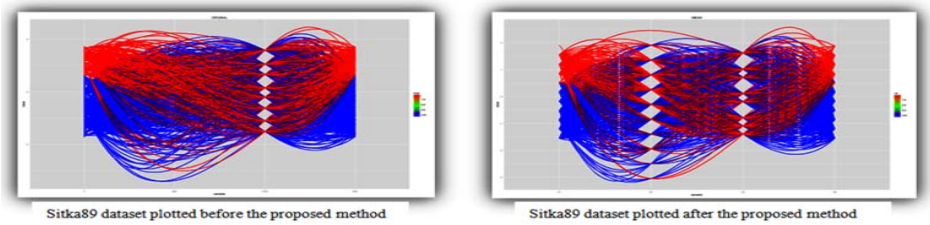

Figure 4. Demonstrate the appearance of data set, sitka89 using traditional (left) and binned parallel coordinate system

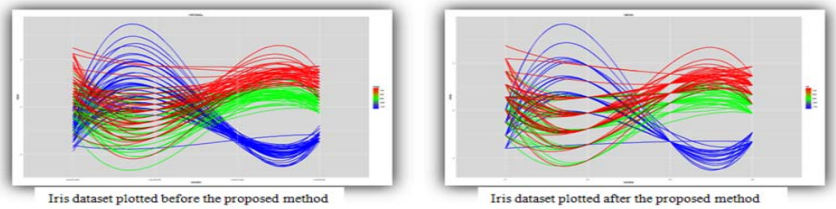

Figure 5. Illustrating the patterns available in IRIS dataset

Display of the data values using binning approach improves pattern optimization. By combining binning and parallel coordinates, the clutters are reduced drastically. This improves the performance and the effectiveness of decision making from the large datasets

4. CONCLUSION

This paper focused on reducing the over plotting in parallel coordinates by reducing the clutters using binning approach. The incorporation of binning methodology into the parallel coordinate plot assists bundling of binned data, which improves the perceptibility of data when compared to the original plot. The plots obtained after incorporating the proposed methodology support its users to arrive at valid conclusions from the large datasets. The visualization becomes more valuable and patterns become more detectable, thereby improving the effectiveness of visualization.

ACKNOWLEDGEMENTS

This work is supported by the DST Funded Project, (SR/CSI/81/2011) under Cognitive Science Research Initiative in the Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham University, Kochi.

REFERENCES

[1] Mao Lin Huang, Liang Fu Lu, Xuyun Zhang. "Using arced axes in parallel coordinates geometry for high dimensional BigData visual analytics in cloud computing", Springer-Verlag Wien, 2014.
[2] Hong Zhou, Xiaoru Yuan2, Huamin Qu, Weiwei Cui, Baoquan Chen, "Visual Clustering in Parallel Coordinates", *Eurographics/ IEEE-VGTC Symposium on Visualization*, Vol. 27, No. 3, 2008.
[3] Michael Schroeder, David Gilbert, Jacques van Helden, Penny Noy, "Approaches to visualization in bioinformatics: from dendrograms to Space Explorer", 2000.
[4] Almir Olivette Artero, Maria Cristina Ferreira de Oliveira, and Haim Levkowitz, "Uncovering Clusters in Crowded Parallel Coordinates Visualizations", *IEEE Symposium on Information Visualization*, 2004.
[5] Ronald R. Yager, Dimitar P. Filev, "Summarizing data using a similarity based mountain method", 2007.
[6] Hadley Wickham Bin, "A framework for visualising large data", *transtats*, 2013.
[7] Aritra Dasgupta, Min Chen, and Robert Kosara, "Conceptualizing Visual Uncertainty in Parallel Coordinates", The Eurographics Association and Blackwell Publishing Ltd., 2012.
[8] Hani Siirtola. "Direct Manipulation of Parallel Coordinates", *IEEE*, 2000.
[9] Jimmy Johansson, Patric Ljung, Mikael Jern, Matthew Cooper, "Revealing Structure within Clustered Parallel Coordinates Displays IEEE Symposium on Information Visualization", *IEEE*, 2005.
[10] Rayner Alfred, "Discretization Numerical Data for Relational Data with One-to-Many Relations", *Journal of Computer Science*, Vol. 5, No. 7, pp. 519-528, 2009.
[11] Helwig Hauser, Florian Ledermann, and Helmut Doleisch, "Angular Brushing of Extended Parallel Coordinates".
[12] Inselberg and B. Dimsdale, "Parallel Coordinates: A Tool For Visualizing Multidimensional Geometry", *IEEE Conf. on Vis. '90*, pp.361-378, 1990.
[13] Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner, "Hierarchical Parallel Coordinates for Exploration of Large Datasets".
[14] Aritra Dasgupta and Robert Kosara, "Screen-Space Metrics for Parallel Coordinates".
[15] Myung-Hoe Huh, Dong Yong Park, "Enhancing parallel coordinate plots", Published by Elsevier Ltd.
[16] Pak Chung Wong, and R. Daniel Bergeron, "Multiresolution Multidimensional Wavelet Brushing", *IEEE*, 1996.
[17] Hing Yan lee and Hwee-Leng Ong, "Visualization Support for Data Mlnlng", *IEEE*, 1996.
[18] Allen R. Martin, Matthew O. Ward, "High Dimensional Brushing for Interactive Exploration of Multivariate Data", IEEE, 1995.
[19] Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases", *SIGMOD '96*, Montreal, Canada IQ, 1996.
[20] Matthew O. Ward, "XmdvTool: Integrating Multiple Met hods for Visualizing Multivariate Data", *IEEE*, pp. 1070-2385, 1994.
[21] J. Han, and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, 2001.
[22] Sturges H. A., "The choice of a class interval". *Journal of the American Statistical Association*, pp. 65–66, 1926.