Stroke Data Analysis through a HVN Visual Mining Platform

Mao Lin Huang[1], Zhixiong Yue[1,2], Jie Liang[1], Quang Vinh Nguyen[3] and Zongwei Luo[2]
[1]University of Technology, Sydney, Australia,
[2]Southern University of Science and Technology, China
[3]University of Western Sydney, Australia
{Mao.Huang@uts.edu.au; Zhixiong.Yue@student.uts.edu.au}

Abstract

Today there are abounding collected data in cases of various diseases in medical sciences. Physicians can access new findings about diseases and procedures in dealing with them by probing these data. Clinical data is a collection of large and complex datasets that commonly appear in multidimensional data formats. It has been recognized as a big challenge in modern data analysis tasks. Therefore, there is an urgent need to find new and effective techniques to deal with such huge datasets.

This paper presents an application of a new visual data mining platform for visual analysis of the stroke data for predicting the levels of risk to those people who have the similar characteristics of the stroke patients.

The visualization platform uses a hierarchical clustering algorithm to aggregate the data and map coherent groups of data-points to the same visual elements - curved 'super-polylines' that significantly reduces the visual complexity of the visualization. On the other hand, to enable users to interactively manipulate data items (super-polylines) in the parallel coordinates geometry through the mouse rollover and clicking, we created many 'virtual nodes' along the multi-axis of the visualization based on the hierarchical structure of the value range of selected data attributes.

The experimental result shows that we can easily verify research hypothesis and reach to the conclusion of research questions through human-data & human-algorithm interactions by using this visual platform with a fully transparency manner of data processing.

Key words: *multi-dimensional data visualization, visual analytics, data mining, risk prediction, stroke*

1. Introduction

Acute stroke care has highly time-dependent treatments that require teams of personnel to achieve good outcomes. It is estimated that for every minute the middle cerebral artery remains blocked in an ischemic stroke, approx. 2 million neurons are lost.

Information visualization is a discipline in its own right that combines graphical display in static or dynamic form to reveal a new understanding of data. In clinical medicine, novel methods of information visualization can lead to improved clinical outcomes at both the population and individual levels.

Clinical data, including stroke data, is a collection of large and complex datasets that commonly appear in multi-dimensional data formats. It has been recognized as a big challenge in modern data analysis tasks. Thus, there is an urgent need to find new and effective technique to deal with such huge datasets and present the outcomes in real-time.

Parallel coordinates are a well-established geometrical system for visualizing multidimensional data [1, 2] that has been extensively studied for decades. There is also a variety of associated interaction techniques currently used with this geometrical system. However, so far only a few existing techniques can achieve the function of selecting a single data item in parallel coordinates geometry because of that theoretically a particular visual poly-line (a visual object) does not have a geometric region.

Therefore, in this project we use a new interactive parallel coordinate visualization proposed in 2016 [3] to create a visual data mining platform for analyzing the available stroke data for predicting the risk of stroke.

This visual mining platform uses a 'virtual node' approach [3] to enable the interactive visualization, and analysis of stroke data. We use optimized data mining, including clustering algorithms to process the data under this visual platform and then visually present the analytical results in real-time. A visual highlighting method is also implemented to support the decision making. The paper demonstrates its effectiveness on a case study of stroke patients.

2. Stroke Data Collection

The stroke data we used for this research project is collected from an open source Healthcare Dataset Stroke Data at [www.kaggle.com/asaumya/healthcare-dataset-stroke-data]. The test dataset consists of 43,400 data items and each item contains 12 attributes.

We use our new visual data mining platform to process the data. Particularly, we use a hierarchical clustering algorithm to generate many 'virtual nodes" along the parallel axis. That allows users to interactively select the values of attributes to generate user's expected view of the output, which can directly verify expert's (Doctor's) hypothesis in real-time.

978-1-7281-2850-4/19/$31.00 ©2019 IEEE
DOI 10.1109/IV-2.2019.00010

2.1. Data Format and Definitions

We now introduce the data format and the medical description of data attributes one by one. Figure 1 shows the table that illustrates the detail of each data attributes.

DATA DICTIONARY

Variable	Definition
id	Patient ID
gender	Gender of Patient
age	Age of Patient
hypertension	0 - no hypertension, 1 - suffering from hypertension
heart_disease	0 - no heart disease, 1 - suffering from heart disease
ever_married	Yes/No
work_type	Type of occupation
Residence_type	Area type of residence (Urban/ Rural)
avg_glucose_level	Average Glucose level (measured after meal)
bmi	Body mass index
smoking_status	patient's smoking status
stroke	0 - no stroke, 1 - suffered stroke

Figure 1: 12 data attributes and their descriptions.

The data attributes are described as follows:

- Hypertension (high blood pressure): Boolean type of data which consists of 0/1 (False/True). The 0 value stands for the sample person did not suffer from hypertension while the 1 value stands for suffering from hypertension.

- Heart Disease (Cardiovascular disease): Boolean type of data which consists of 0/1 (False/True). The 0 value stands for the sample person did not suffer from heart disease while the 1 value stands for suffering from heart disease.

- Work Type: categorical type of data including Private, Self-employed, Government job, Children and Never worked. These five categories are represented by label encoding as 0, 1, 2, 3 and 4 accordingly.

- Average Glucose Level (Blood Sugar Level): is a numerical type of data measured after meal in unit of mg/dL. The normal average glucose level is below 125 mg/dL for non-diabetes people who are not fasting [4] Generally, when average glucose level is higher than 200 mg/dL, the person is in a condition of Hyperglycemia (High blood sugar) [5]. When average glucose level is lower than 70 mg/dL, the person is in a condition of Hypoglycemia (Low blood sugar) [6].

- BMI (body mass index): is a numerical type of data which is derived from the mass and height of an individual. The formula of calculating BMI is expressed as follows:

$$BMI = \frac{mass}{height^2}$$

Where the mass is in unit of kilograms, the height is in un it of meters. The BMI is universally expressed in kg/m^2. The normal (healthy weight) range of BMI is from 18.5 to 25. The World Health Organization (WHO) regards a BMI of less than 18.5 as underweight, greater than 25 as overweight and above 30 as obese.

- Smoking Status: is a categorical type of data including never smoked, formerly smoked and smokes. These three categories are represented by label encoding as 0, 1 and 2 accordingly.

- Stroke: is a Boolean type of data which consists of 0/1 (False/True). This variable is the target variable of our dataset. The 0 value stands for the sample person has never been suffering from stroke while the 1 value stands for the sample person has suffered from stroke.

3. Interactive Parallel Coordinates Visualization

The original Parallel Coordinates visualization [1] does not support direct selection of single data item via a mouse click interaction. This is because that a visual poly-line (a visual object) have theoretically no geometric region. Therefore, we have to use a new Hierarchical Virtual Node parallel coordinates visualization technique proposed in 2016 [3] to enable the "select" operation in Parallel Coordinates Visualization.

3.1. Hierarchical Virtual Node (HVN)

The basic idea behind HVN is to interpolate nodes directly in parallel coordinates for hierarchical data selection. A node structure is an intuitive interface of interaction because it has a duality of representing both the data and coordinates. In our design, a node represents a collection of homogeneous data points based on location proximity and clustered by hierarchical clustering.

Hierarchical Clustering [7] is a widely used data mining method for clustering data into a hierarchy of disjointed clusters. It needs a 'stopping rule' to terminate the process when the optimal number of clusters has been determined. Fortunately, we build the entire hierarchy for each dimension so finding a suitable 'stopping rule' is not a concern in this work. Hierarchical Clustering initializes each data item into a cluster and iteratively merges clusters based on the shortest inter cluster distance to form a new cluster.

There are several advantages of this design over other techniques in parallel coordinates family.

First, it achieves the direct data selection in parallel coordinates via mouse clicking which is intuitive and efficient instead of through the means of a widget.

Second, it does not requires user input of data range in absolute value. Quantization is always difficult and it requires the user to be familiar with the dataset in order to produce precise value patterns. Instead, it allows users to adjust the value ranges by disabling some virtual nodes through mouse clicking.

Third, the granularity of multi-level interaction with multidimensional data is a much more distinctive feature than other methods. The HVN provides hierarchical aggregation of data with greater flexibility to quickly explore data patterns between nearest neighbors.

Fourth, the interpolation of virtual nodes allows the user to perceive the distribution of the data density through the distribution of the virtual nodes in an elegant way. Data density is an important characteristic of a continuous variable in a multidimensional dataset because this is usually hidden in parallel coordinates due to over-plotting. Hauser et al. [8] had provided a similar implementation by embedding histograms in parallel coordinates.

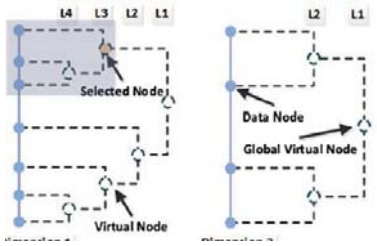

Figure 2: Hierarchical Virtual Nodes (HVN): shaded rectangle indicates the selection of data points if the orange node has been clicked.

4. Case Study: Stroke Data Analysis

We collected the data from an open source Healthcare Dataset-Stroke-Data downloaded from the following site at [www.kaggle.com/asaumya/healthcare-dataset-stroke-data]. The test dataset consists of 43,400 data items and each item contains 12 attributes.

We use our new visual data mining platform to process the data. Particularly, we use a hierarchical clustering algorithm to aggregate the data items into 'clustered and curved polylines' that will pass through some 'virtual nodes" along the parallel axis. Note that these virtual nodes are created based on the clustering algorithms that allows users to interactively select the values of attributes to generate user's expected view of the output, which can directly verify expert's (Doctor's) hypothesis in real-time.

4.1. Displayed by Original Parallel Coordinates Visualization

Parallel coordinates [1, 2] is a common way of visualizing and analyzing multidimensional datasets. It has been extensively studied for decades. To show a set of (value) points in an n-dimensional space, a backdrop is drawn consisting of N parallel lines, typically vertical and equally spaced. A (value) point in n-dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the i^{th} axis corresponds to the attribute value in i^{th} coordinate. An example of the parallel coordinate visualization of the stroke dataset is illustrated in Figure 3.

We can see from the visualization shown in Figure 3 that it does not give us any specific knowledge. What we can only see is the abstract view of the density distribution of the polylines (or data items). We can see that there are three over-crowded and overlapped display regions. Because that they are too overcrowded, so that we cannot see the detail of these visual objects. We usually call them 'visual clut" or 'over-plotting'.

Therefore, we, at this stage, got the conclusion that the visualization of stroke data with the original parallel coordinate geometry does not give you any useful information and knowledge for further stroke prediction.

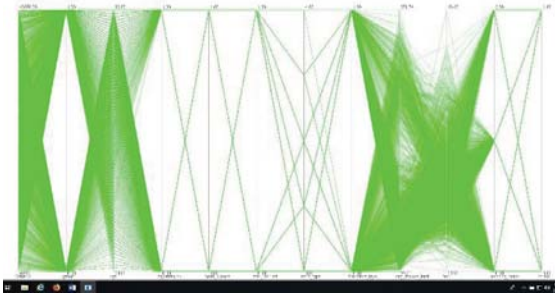

Figure 3: A screenshot collected from our HVN visual platform that displays of entire stroke dataset in the ordinary parallel coordinates visualization.

4.2. HVN Clustering in Parallel Coordinates Visualization

We use a hierarchical clustering algorithm to aggregate polylines (or data items) into curved "super polylines". Thus, the number of visual objects (polylines) are greatly reduced and accordingly the visual complexity of the visualization is also reduced significantly.

To enable the direct interaction with one or small groups of data items, we generated many 'clickable' virtual nodes based on the clustering structure. We then use another algorithm to partition the data points into hierarchical clusters.

The basic idea behind HVN is to interpolate nodes directly in parallel coordinates for hierarchical data selection. A node structure is an intuitive interface of interaction because it has a duality of representing both the data and coordinates. In our design, a node represents a collection of homogeneous data points (items) based on location proximity and clustered by hierarchical clustering.

Figure 4 shows a HVN visualization of the stroke data, from which we can see that the number of polylines is greatly reduced. The display becomes much clear and readable than the previous display appeared in the original parallel coordinate visualization. In the current HVN visualization, the leftover curved 'polylines' are representing clusters, rather than data points (items). The visual effects of these curved polylines are much better than the data point (items) polylines displayed in the original parallel coordinate visualization.

This is that because that we uses a hierarchical clustering algorithm to aggregate data items into curved

'super-polylines' that significantly reduces the visual complexity. On the other hand, to enable users to interactively manipulate data items (super-polylines) in the parallel coordinate geometry through the mouse rollover and clicking, we created many 'virtual nodes' along the multi-axis of the visualization based on the hierarchical structure of the value range of the data attributes. We can see from the Figure 4 that

We can see from Figure 4 that the number of polylines is greatly reduced and now it remains only a small number of the 'super-polylines", those are used to represent clusters.

We may also find that three over-crowded display regions have been disappeared. Now the quality of the visualization is much better than the quality of previous visualization as shown in Figure 3.

5. Visual Analysis by Interactions on HVN Visualization

Now we are going to demonstrate the usefulness of HVN visual platform in analyzing the stroke data, verifying expert's medical hypothesis and predicting potential stroke illness.

As Doctors, they usually want to know the impact of people's current health symptoms and personal hobbies to the stroke.

This is because that if we could find out some direct relationships between stroke and one or more such symptoms or hobbies, we can then effective prevent the stroke illness by providing people with targeted medical services and changing people's living habits.

These personal data including *Hypertension (high blood pressure), Heart Disease (Cardiovascular disease), Work Type, Average Glucose Level (Blood sugar level), BMI (body mass index)* and *Smoking Status*.

In our HVN visual analysis platform, we use direct interaction with 'virtual nodes' to find out such relationships and the impact of these personal health symptoms and other attributes to the stroke illness.

We now discuss our interactive visual analysis result and the conclusions gained from our HVN visual mining platform. Figures 5 and 6 are screenshots that showing us the analysis outcomes as described below:

5.1. Interactive Visual Analysis Finding 1

Suppose that we try to find out the impact of people's *'Blood Sugar Level*' to the Stroke illness, we then can implement the following interactions on the HVH visualization as shown in Figure 4:

1) De-select clusters of data points, those with the AGL value lower than 250 (AGL <250 mg/dL), by simply clicking on a series of corresponding 'virtual nodes' in the visualization shown in Figure 4.

2) After performing the above 'mouse clicking' interactions, the HVN visualization will be transformed from the one shown in Figure 4 to another one shown in Figure 5.

Analysis Result and Conclusion 1: In Figure 5, we can find the following conclusion that is "if a person has high level blood sugar (AGL > 250 mg/dL), and s/he is Self-employed, and aged between 50 to 82, and s/he is currently smoking, then s/he must have suffered from stroke, regardless s/he is fat or thin".

We have found that there is no direct relationship between stroke and BMI index. This means that whatever you fat or thin, as long as your blood sugar level is higher than 250 mg/dL, smoking, self-employed and aged 50 +, you must be suffered from stroke.

5.2. Interactive Visual Analysis Finding 2

In the second visual interaction experiment, we implement the following interactions on the HVH visualization in Figure 4:

1) De-select clusters of data points, those with the AGL value higher than 70 mg/dL (AGL >70 mg/dL), by simply clicking on a series of corresponding 'virtual nodes' in the visualization shown in Figure 4.

2) After performing the above 'mouse clicking' interactions, the HVN visualization will be transformed from the one shown in Figure 4 to another one shown in Figure 6.

Analysis Result and Conclusion 2: In Figure 6, we can find the following conclusion that is "if a person has low level of blood sugar (e.g. AGL < 70 mg/dL), and s/he is Self-employed, and never had any record of Heart Disease, and even if s/he is very fat (obese, BMI index >50), then s/he must not suffered from stroke".

As shown in Figure 6, all clusters crossing through the leftover “virtual nodes” belongs to “no stroke” category. This means that in our dataset if a sample has relatively low average glucose level, who is in a condition of Hypoglycemia (lower than 70 mg/dL), then s/he must not suffered from stroke. Furthermore, all these samples have not been suffering from heart disease. All these findings are well defined by our visual platform.

Conclusions and Future Work

We have presented our research outcomes that is to use a new HVN visual data mining platform for interactively and transparently analyzing the stroke data. Doctors and domain experts may use this analysis platform to adjust the data mining algorithms through direct 'mouse clicking' on a series of virtual nodes.

They can also use 'mouse clicking' interaction on these virtual nodes to select or de-select specific data item(s) or group(s) of data items of their interest. Thus, they could easily verify their research hypothesis and find out the explicit and implicit relationships between a certain disease or illness and patient’s health attributes. Consequently the domain experts could use these findings to produce some mechanisms and medical advises for future prediction and prevention of these diseases.

The preliminary and informal evaluation has proved the effectiveness of using HVN visual mining platform for

finding implicit and explicit impact and relationships. We will next carry out a formal usability study.

In the future, we want to also conduct formal aesthetic evaluation to the new HVN visualization from user's perspectives. Aesthetics are often used to measure the layout quality of graph drawings and it is commonly accepted that drawings with good layout are effective in conveying the embedded data information to end users [9, 10].

However, the existing aesthetic criteria are currently only applying to graph visualizations and only few are applied to parallel coordinates geometry. However, in parallel coordinate visualization, the edge-crossings [11] and over-plotting among polylines are also the serious problems that significantly affect to the quality of visualization.

Currently the parallel coordinate visualization are mainly evaluated based on personal judgments and user studies for their overall quality. However, personal judgments are not reliable while user studies can be costly to run. Therefore, there is a need for a direct measure of overall quality of parallel coordinate geometry. Therefore, in the future we will investigate a common measurement standard that measures overall quality based on individual aesthetics and gives a single numerical score on the aesthetic and readability quality of a given multi-dimensional data visualization.

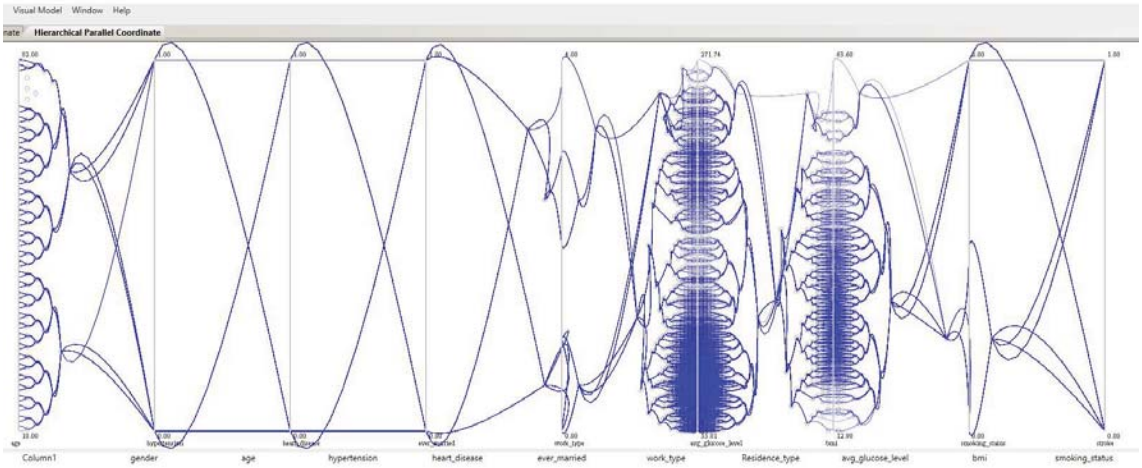

Figure 4: A screenshot collected from HVN visual platform that displays of entire stroke dataset downloaded from Healthcare Dataset Stroke Data. The Curved Polylines represent clusters that are created base on the hierarchical clustering structure.

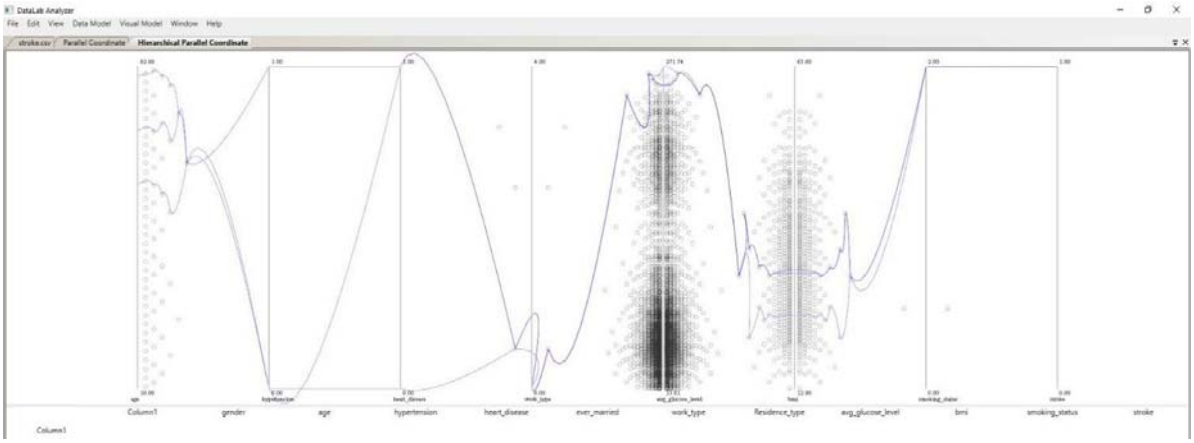

Figure 5: A screenshot collected from HVN visual platform. We can find the following fact from this picture, that is "*if a person has high level blood sugar (AGL > 250 mg/dL), and s/he is Self-employed, and aged between 50 to 82, and s/he is currently smoking, then s/he must have suffered from stroke, regardless s/he is fat or thin*".

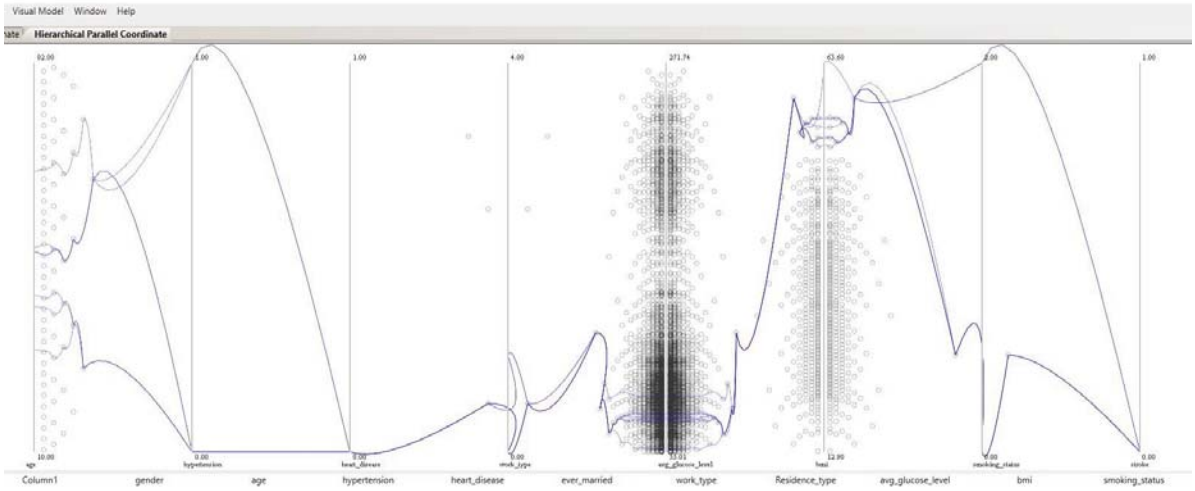

Figure 6: A screenshot collected from HVN visual platform. We can find the following fact from this picture that is "*if a person has low level of blood sugar (e.g. AGL < 70 mg/dL), and s/he is Self-employed, and never had any record of Heart Disease, and even if s/he is very fat (obese, BMI index >50), then s/he must not suffered from stroke*".

References

[1] Alfred Inselberg, 1985. 'The Plane with Parallel Coordinates', *the Visual Computer, vol. 1(2), pp. 69-91, Springer-Verlag.*

[2] Lu, L., Huang, M. and Huang, T., 2012. 'A new axes re-ordering method in parallel coordinates visualization', *Proceedings of 11th International Conference on Machine Learning and Applications, vol. 2, pp 252-257.*

[3] Huang, M., Huang, T. and Zhang X., 2016. 'A novel virtual node approach for interactive visual analytics of big datasets in parallel coordinates', *Future Generation Computer Systems vol. 55, pp. 510-523.*

[4] [https://medlineplus.gov/ency/article/003482.htm].

[5] [http://care.diabetesjournals.org/content/37/Supplement_1/S81]

[6] [https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/low-blood-glucose-hypoglycemia]

[7] Eisen, M.B., Spellman, P.T., Brown, P.O. Botstein, O., 1998. 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl. Acad. Sci. USA (1998), pp. 14863–14868.*

[8] Hauser, H., Ledermann, F., Doleisch, H., 2002. 'Angular brushing of extended parallel coordinates', *IEEE Symposium on Information Visualization, 2002, pp. 127–130.*

[9] Huang, W., Huang, M., Lin, C., 2016. 'Evaluating overall quality of graph visualizations based on aesthetics aggregation', *Information Sciences 330, 444-454.*

[10] Nguyen, Q., Huang, M., Hawryszkiewycz, I., 2004. 'A new visualization approach for supporting knowledge management and collaboration in e-learning', *Proceedings of Eighth International Conference on Information Visualization, pp. 693-700.*

[11] Huang, W., M Huang, M., 2011. 'Exploring the relative importance of number of edge crossings and size of crossing angles: A quantitative perspective', *International Journal of Advanced Intelligence 3 (1), 25-42.*