Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 42 (2014) 247 – 254

Procedia Computer Science

International Conference on Robot PRIDE 2013-2014 - Medical and Rehabilitation Robotics and Instrumentation, ConfPRIDE 2013-2014

Multidimensional Data Medical Dataset Using Interactive Visualization Star Coordinate Technique

Noor Elaiza Abd Khalid[a], Marina Yusoff[b], Ezzatul Akma Kamaru-Zaman[c], Izyan Izzati Kamsani[d]*

[a,b,c,d] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Shah Alam, 40000, Malaysia

Abstract

The twenty first century sees the tremendous advancement of computer and machine technologies that are able to produce ginormous amout of data. Current software architecture, management and analysis approaches are unable to cope with the flood of data. The challenge of understanding large and complex data includes issues such as clutter, performance, information loss and limited cognition. Medical field involves analyzing the body system which includes many different scientists and medical professionals. The datasets are a hybrid of many different medical areas databases to understand and answer the many questions of the human body. This paper explores the capability of interactive star coordinate visualization technique to identify clusters correlation between selected attributes using interactive star coordinate for multi-dimensional datasets An interactive Star coordinates is designed consists of four stages that includes Information Objects Transformation; Dimension Mapping; Interactive Features design and Coloring. Finally the performance of the interactive star coordinates is compared to histograms of the data of interest. Interactive star coordinate is found as a promising method of visualizing information clusters pattern which provides one of the means for fast decision making.

Peer-review under responsibility of the Center for Humanoid Robots and Bio-Sensing (HuRoBs)

Keywords: Big data; Clutter; Decision making; Histogram graph; Multidimensional data; Star coordinate technique; Visualization

* Corresponding author. Tel.: +60192692717; fax: +60355435501.
E-mail address: elaiza@tmsk.uitm.edu.my

Peer-review under responsibility of the Center for Humanoid Robots and Bio-Sensing (HuRoBs)
doi:10.1016/j.procs.2014.11.059

1. Introduction

Recent advent of technology has provided tremendous amount of data and information in various fields such businesses, economy, healthcare, biomedical and bioinformatics [1][2]. Datasets in these fields usually include multidimensional datasets consisting of a significantly larger range of attributes [3]. Gaining insight about of data is not just a matter of presenting it, but translating data into information as observed by Spence [4]. Foraging and unearthing vital hidden links between the diverse variables and parameters within voluminous datasets is a slow painstaking, tedious and complex tasks. Analysing information within a pool of data may cause complexity in term of clutter, performance, information loss and limited cognition [5].

Data visualization method provides the means to summarize and interpret large data promptly [6]. Even though many methods have been developed, the scientific community has yet to create and agree upon standard tools for data visualization, manipulation and analysis. The existence of multi-dimentional datasets further complicates the analysis. Thus, more robust visualization technique is needed to create intelligent visualization designs such an interactive view [7]. It allows users to view data in different angles, axis and attributes manipulation. Applying colors allows instant recognition of similarities or differences of the large data items and expressed attributes relationship [8]. Interactive visualization is able represent huge amount of information coherently, compactly from different viewpoints and provides several level of details [6]. Star coordinate allows interactive online manipulation of attributes dimension[9]. Kandogan [10] found it to be useful in gaining insight (not numerical analysis) into hierarchical clustered datasets.This research aims apply the flexibility of interactive star coordinate data visualization technique to uncover clusters of hidden associations and relationships between the data within a span of attributes.

2. Proposed Method

This research work consists of four phases as illustrated in Figure 1 including the data collection; interactive star coordinates design; manipulation of the interactive features and usability and accessibility.

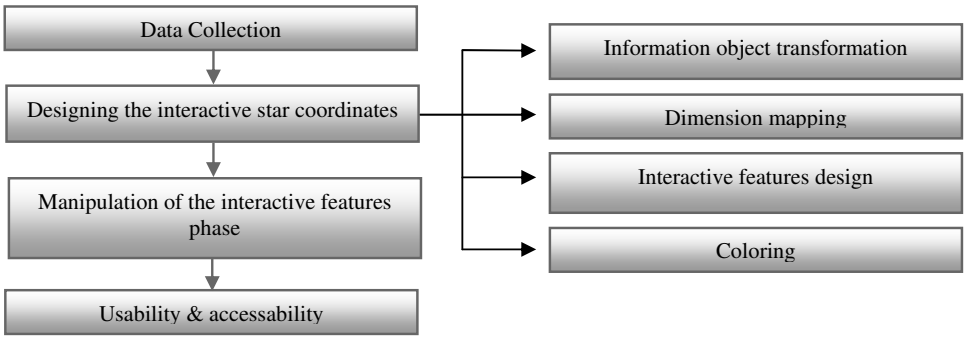


Fig. 1. Design process for ViStar.

2.1. Data Collection

In this study, six hundred and ninty nine retrospective datasets of early stage (Stage I and II) Estrogen Positive (EC+) Breast Cancer of women treated and not treated with tamoxifenmonotherapy diagnosed between 1980 to1995 from GEO Database [11].

2.2. Interactive Star Coordinate engine design phase

This phase consists of four stages that includes Information Objects Transformation, Dimension Mapping, Interactive Features design and Coloring.

Stage1 involves the transformation of information objects from the data file. This involves assigning numerical

values to non numerical values. Subsequently, the data are arranged into a matrix where columns represent the dimensions and row values for each field in a record. Figure 2 shows a matrix of information objects $P_1,P_2,...,P_n$ which are associated with each of the six fields or attribute of the dataset denoted as.

	F_1	F_2	F_3	F_4	F_5	F_6
P_1	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}
P_2	d_{21}	d_{22}	..			
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
P_n	d_{n1}	d_{n2}				d_{n6}

where
P_n: n information objects ; n = 699 .
F_i: i attributes or dimensions; i = 6

Fig. 2. Information objects to matrix transformation.

Stage 2 involves mapping each information object onto the star coordinate axes. The axes $C_1,...,C_6$ denoted the fields or dimensions share a common origin, which in the Cartesian Coordinate system may be conveniently denoted by (0,0) shown in Figure 3. Each field vector $f_1,...,f_6$ is calculated by multiplying the distance d_{ij} with its corresponding unit vector, c_j oriented in a direction along the axis,C_j. Subsequently, vector $P_j(x,y)$ denoting the final point are calculated based on equation 1.

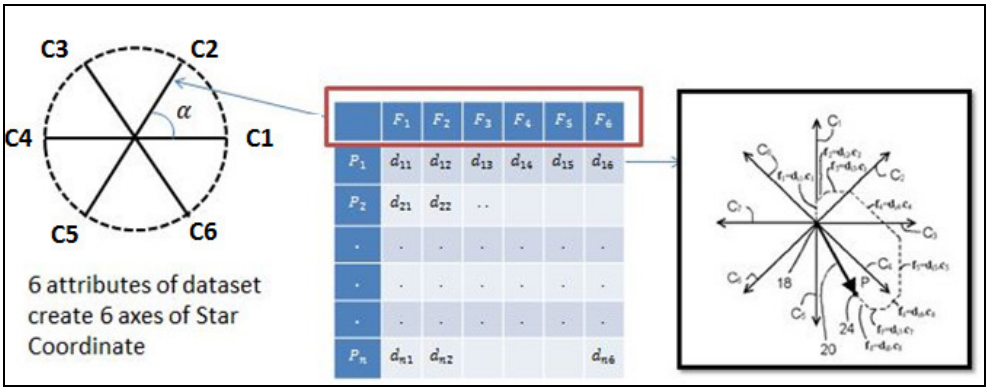


Fig. 3. Mapping Architecture.

Each point are calculated according to Equation 1 from Kadogan [12].

$$P_j(x,y)=\left(\sum_{i=1}^{n} u_{xi}.(d_{ji}-\min_i),\sum_{i=1}^{n} u_{yi}.(d_{ji}-\min_i)\right) \quad (1)$$

where

$$D_j=(d_{jo},d_{j1},...,d_{ji},...,d_{jn}),u_i=\frac{c}{\max_i-\min_i}, \quad (2)$$

where,

$$u_i=(u_{xi},u_{yi})=(\cos\alpha,\sin\alpha)$$

$$\max_i=\min\{d_{ji},0\le j\langle number\_of\_elements\},$$

$$\min_i=\max\{d_{ji},0\le j\langle number\_of\_elements\}.$$

Stage 3 involves incorporating interactive features including scaling and rotation features for online data manipulation. Scaling allows users to change one or more axis length concurrently. Scaling feature involves recalculating the contribution of attributes by multiplying the ratios accordingly to 'mapping' equation and re-

mapped in accordance to the new scaling factor as in Equation 3 [10].

$$\frac{c}{\max_i - \min_i} * scale \tag{3}$$

Rotation on the other hand, provides users the facility to rotate the axis which re-correlate the relationships between atributes. Rotation changes the axis angle and re-distributes the scatter plots as in equation 4.

$$v_i = \left(\cos\frac{2\pi i}{m}, \sin\frac{2\pi i}{m}, 1 \right), i = 1,...,m \tag{4}$$

Stage 4 involves coloring mapped data according to specified categories. Coloring created another dimension of data visualization and can be classified into interactivity features as users are given the liberty to select a variety of colors to represent different attribute values in the dimension. It involves categorizing data based on similar factors and assigning colors to each group of factors. Thus, enabling visualization of information data distribution or clear clustering effect.

2.3. Manipulation of the interactive features phase

Initial stage of this phase involves gaining knowledge and understanding of the multidimensional data mapped into the star coordinate. This is done through comparing the star coordinate views with the quantitative views using the histogram.

The next experiment involves the manipulation of the incorporated interactive feature to produce the clustering effect. This phase involved three processes which are scaling; rotation and coloring process.

Process 1 involves scaling features where the length of the axis is changed based on the quantitative analysis using histograms.In this experiment the attribute is adjusted to 2:1 ratio of other attributes and vice versa for example. The ratios are then multiplied to mapping calculation. The data are re-mapped again according to the new scaling factor depending on the length of the axis.

Process 2 involves rotation features which it is used to determine the correlation between selected attributes. A function is created to store the rotated angle value and send the parameter to mapping function. As angle changes, recalculation is made and the data are re-mapped again.

Process 3 involves coloring features. This process is also beneficial in visualization to differentiate the required information and also can be used to identify clusters if any. In this process, the data are categorized in same factors and then colors are assigned to each other. When users select coloring feature, data are plotted in its corresponding colors.

2.4. Usability & accessibility

In this phase, the clustering outcome of the previous phase are being presented Professor Dr. Mohd Zaki Salleh and Professor Dr. Teh Lay Kek experts in the pharmacy domain. Since dataset is multi-factorial dataset, many factors need to be considered. They suggested that high correlated attributes must be positioned next to each other to highlight important information relationships and optimize analysis.

3. Result and Discussion

The results discussion is divided into two sections; comparing star coordinates with histogram visualization of the same attributes and clustering multidimensional data using the interactive features.

3.1. Comparison of star coordinates color data distribution and histogram

Comparison between star coordinates and histogram are made according to attributes clusters. Star coordinates have limitation by providing the amount of data in each cluster of numerical data compared to histogram. However, it provides a better illustration in terms of data relationship distribution and are able to categorize non-numerical data as depicted in table 1 and table 1. Table 1 depicts the visualization outcome of real numerical values. The

histogram provides the frequency of each attribute. While star coordinate visualizes clusters during the data mapping.

Analysis of age attributes (numerical data) shows various colors that represent different ranges of age during the mapping clearly shows clustered data (range age between 60-65 has the highest frequency). However, conclusion about the frequency of patients based on age are just a guess. A histogram of the age range is able to verify this guess quantitatively. The same goes with the tumor size analysis.

Table 1. Comparison between star coordinates with histogram to identify clusters for age and tumor size.

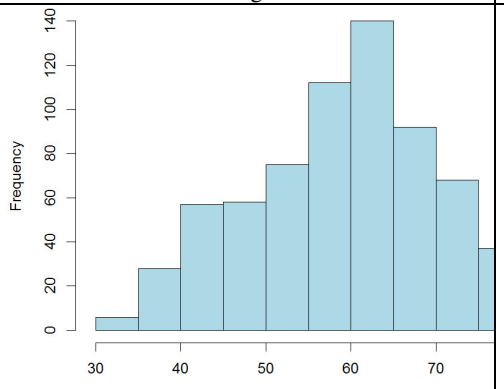
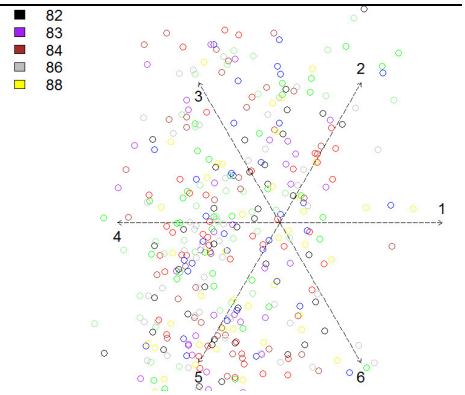
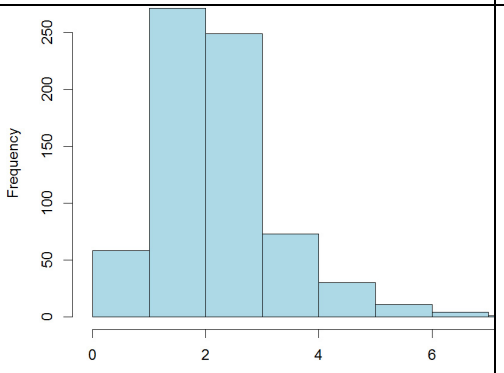
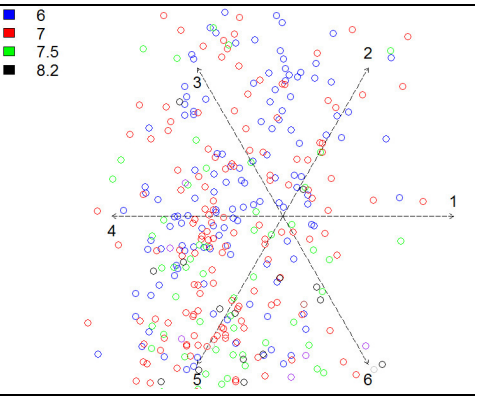
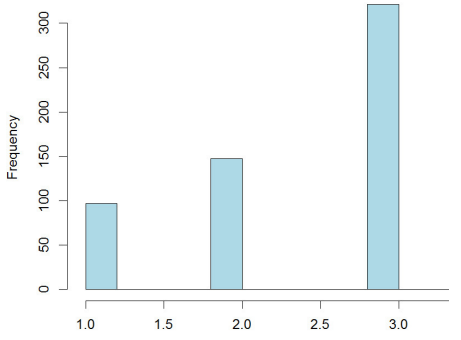
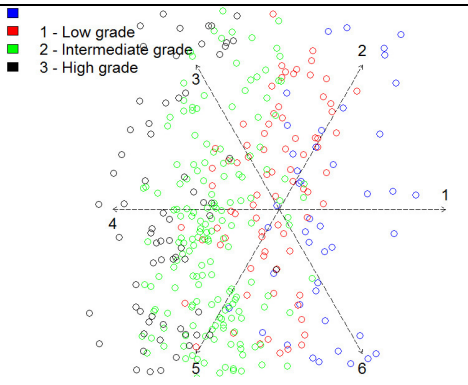
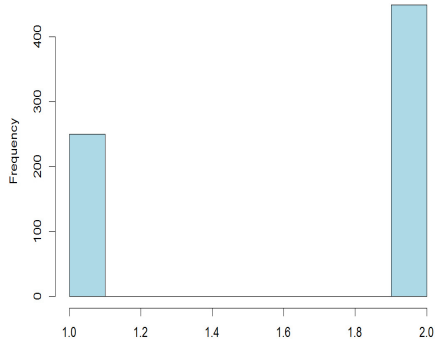
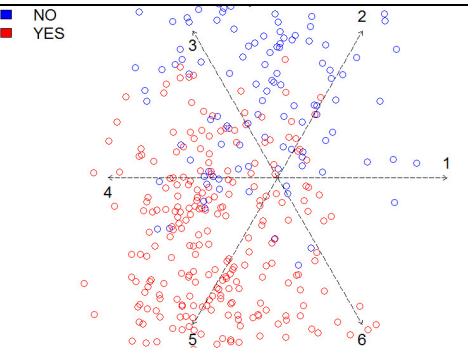
Attributes	Histogram	Star Coordinate
Age		
Tumor Size		

Table 2 depicts the visualization outcome of discrete numerical (Elston Ellis Grade) and non numerical (Tamoxifen Treatment) data. The histogram of the Elston Ellis Grade data shows the frequency of each discrete data whereas the star coordinates shows the discrete data distribution in relation with other attributes. Analysis of tamoxifen treatment attributes (categorical data) shows two clusters that represent the different colors for No and Yes classification in star coordinate. Here we can see that most patients are on medication but we did not know whether it is accurate or not. In this case, the histogram is applied to see the frequency of patients on medication. Furthermore, it is proved that majority patients are on medication for breast cancer treatment.

Table 2. Comparison between star coordinates with histogram to identify clusters for Elston Ellis Grade and Tamoxifen Treatment.

	Histogram	Star Coordinate
Elston Ellis Grade		
Tamoxifen Treatment		

This analysis shows that star coordinate is suitable for illustrating for non-numerical data (categorical data) while histogram is for numerical data (non-categorical data). Attributes for age and tumor size are examples of numerical data while Elston-Ellis Grade and tamoxifen treatment are fall under non-numerical data (categorical data). Attributes of age and tumor size clearly shows that star coordinate is not preferable when analyzing numerical data compared to the histogram.

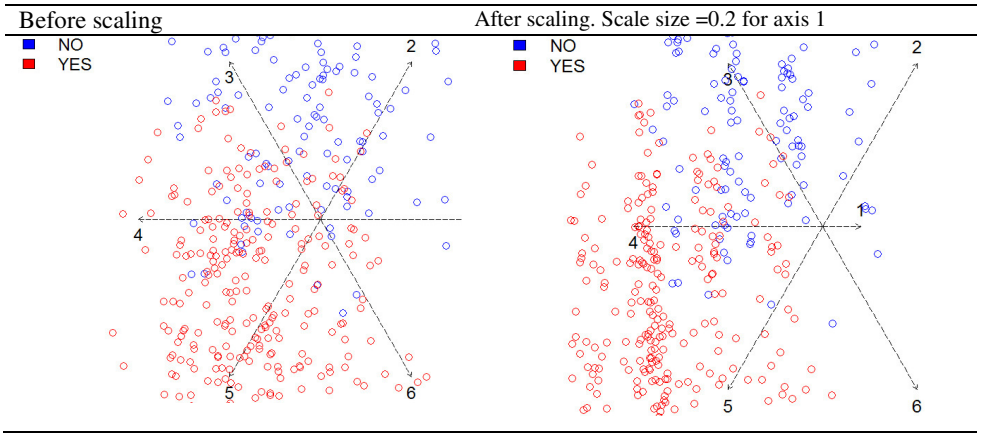
3.2. Interactive manipulation results

This section discusses the data clustering observed from interactive star coordinates using scaling , rotation and coloring interactive features.

3.2.1. Scaling results

Initially, all the axis scale size for all attributes (axis) are set to 1. The data point is observed as coarsely scattered over each attribute. Scaling transformation allows the users to change the length of the corresponding dimensional axis. Generally, scaling may be applied to one or more axis. Scaling is crucial to observe relationship between dimension in the specific range of the scale. This often result in the visualization of data similarities when data on different factors fall under same cluster. In this case, some form of clustering is revealed when the scale size is set to 0.2 for axis 1 as observed in table 3.

Table 3. Comparison of scaling result (before and after)



3.2.2. Rotation results

Users are allowed to rotate particular attribute by adjusting the angle value of the axis. As angle change, recalculation is made and the data are mapped again. Each angle represent for each attribute; axis 1 (age), axis 2 (disease free survival days), axis 3 (disease metastasis free survival days), axis 4 (Elston Ellis Grade), axis 5 (tamoxifen treatment) and axis 6 (tumor size). Rotation makes selected fields or attributes (axis 1to axis 4) more or less correlated to other attributes. Figure 6 shows the visual analysis using interactive star coordinates is able to form two clusters when four fields/attributes(axis 1 to axis 4: age, disease free survival days, disease metastasis free survival days and Elston Ellis Grade respectively) are aligned in the same angular direction. In this case the Age field/attribute plays significant role in surviving breast cancer, depending on the level of tumor grade.

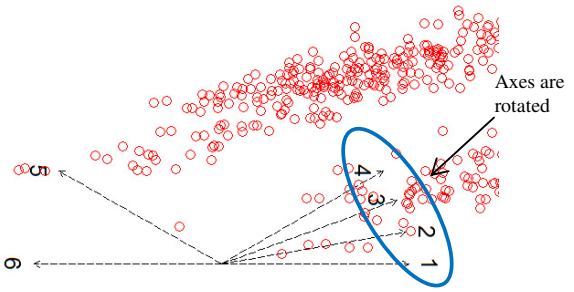


Fig. 4. Star Coordinate : After rotation.

3.2.3. Coloring Results

Colors created another dimensions in data visualization. Attributes used for this analysis are axis 1 (age), axis 2 (disease free survival days), axis 3 (disease metastasis free survival days), axis 4 (Elston Ellis Grade), axis 5 (tamoxifen treatment) and axis 6 (tumor size). Blue color represents for tamoxifen treatment (No) and red color represents for tamoxifen treatment (Yes).

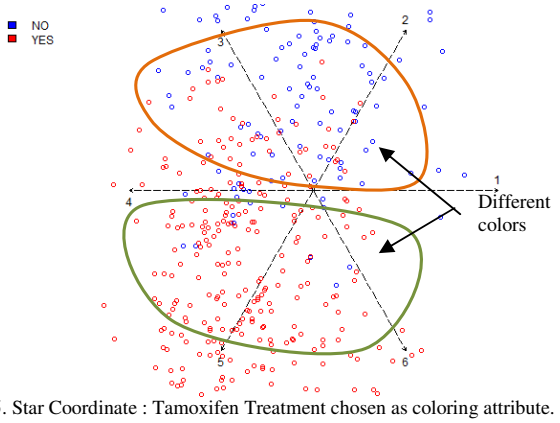


Fig. 5. Star Coordinate : Tamoxifen Treatment chosen as coloring attribute.

Fig. 5. shows that most patients treated with tamoxifen (Yes) with large tumor size in red appears on axis 6, while patients with small tumor size who are not treated with tamoxifen (No) in blue appears on axis 2. Based on this analysis, it can be concluded that when patients get treatment at an early stage have longer survival free days.

4. Conclusion

Interactive Star Coordinate technique has provided the transformation of data from tabular form into a more understandable visualization view. Furthermore, this technique provides a method that enhances user capability in gaining insight multidimensional information sets. As a result, interactive star coordinates contains various capabilities, including representing multidimensional information sets (such as Breast Cancer Data) in two dimensional spaces and assisting in obtaining intuitive information of patients clinical pattern. The interactive features incorporated into star coordinates can successfully visualize the data categories into clusters that are much easier to recognize. Thus, facilitates user understanding of large and multidimensional data that can assist in decision making. However, a more accurate quantitative value can be obtained using histograms.

Acknowledgement

The authors wish to thank to UiTM for the facilities provided. Special gratitude to Prof Dr. Mohd Zaki Salleh and Prof Teh Lay Kek from Pharmacy faculty (UiTM) for their advice, comments, guidance and sharing information throughout this research.

References

[1] J. Barkai, “Using Visual Decision Making to Optimize Manufacturing Design and Development,” 2012.
[2] N. Andrienko, G. Andrienko, S. Birlinghoven, and S. Augustin, “Informed Spatial Decision Making Using Coordinated Views,” 2003.
[3] G. Dzemyda, O. Kurasova, and J. Zilinskas, *Multidimensional Data Visualization*. Springer Optimization and Its Applications, 2013, p. 248.
[4] R. Spence, *Information Visualization*. ACM Press Book, 2001.
[5] Enrico, “How do you visualize too much data,” 2011. [Online]. Available: http://fellinlovewithdata.com/guides/how-do-you-visualize-too-much-data.
[6] M. Khan and S. S. Khan, “Data and Information Visualization Methods , and Interactive Mechanisms : A Survey,” *Int. J. Comput. Appl.*, vol. 34, no. 1, pp. 1–14, 2011.
[7] Q. V. Nguyen, G. Nelmes, M. L. Huang, S. Simoff, and D. Catchpoole, “Interactive Visualization for Patient-to-Patient Comparison.,” *Genomics Inform.*, vol. 12, no. 1, pp. 21–34, Mar. 2014.
[8] D. A. Keim, “Information Visualization and Visual Data Mining,” *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 1–8, 2002.
[9] W. W. Chan, “A Survey on Multivariate Data Visualization,” 2006.
[10] E. Kandogan, H. Road, and S. Jose, “Star Coordinates : A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions,” *Proc. IEEE Inf. Vis. Symp. Vol. 650*, 2000.
[11] S. Loi, B. Haibe-Kains, C. Desmedt, P. Wirapati, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, K. Ryder, J. F. Reid, M. G. Daidone, M. a Pierotti, E. M. Berns, M. P. Jansen, J. a Foekens, M. Delorenzi, G. Bontempi, M. J. Piccart, and C. Sotiriou, “Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen.,” *BMC Genomics*, vol. 9, p. 239, Jan. 2008.
[12] E. Kandogan and S. Jose, “Visualizing Multi-dimensional Clusters , Trends , and Outliers using Star Coordinates,” *Proc. seventh ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, pp. 107–116, 2001.