SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Big Data: Issues and Challenges in Clustering Data Visualization

Ummu Hani' Hair Zaki[1], Izyan Izzati Kamsani[1,*], Ahmad Firdaus Ahmad Fadzil[2], Zainura Idrus[3], Eser Kandogan[4]

[1] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia
[2] College of Computing, Informatics and Media, Universiti Teknologi Mara Cawangan Melaka (Kampus Jasin), 77300 Merlimau, Melaka, Malaysia
[3] Faculty of Computer and Mathematical Science, Universiti Teknologi Mara, 40450 Shah Alam, Selangor, Malaysia
[4] Megagon Labs, 444 Castro St #900, Mountain View, CA 94041, United States

ARTICLE INFO

Article history:
Received 22 June 2023
Received in revised form 10 February 2024
Accepted 15 August 2024
Available online 2 September 2024

Keywords:
Big data; Clustering visualization; Geometric projection; Star coordinate

ABSTRACT

In the era of big data, the continuous generation of data from various fields has resulted in large and complex datasets. These datasets often come in diverse formats and structures, including unstructured or semi-structured data. Despite the wide availability of big data, high dimensionality remains a significant challenge for analysing and understanding the data for various purposes. Clustering analysis plays a crucial role in data analysis and visualization by uncovering hidden patterns and structures within datasets. However, several challenges hinder the effectiveness of clustering analysis, including data dimensionality, selection of appropriate clustering algorithms, determining the optimal number of clusters, interpreting the results, and handling outliers. This paper aims to explore these challenges and presents preferable visualization techniques that aid in visualizing and interpreting clustering results. By addressing these challenges, including the difficulty of handling outliers and the struggles with high-dimensional datasets, and employing effective visualization techniques, researchers and practitioners can enhance their understanding and utilization of clustering analysis in data analysis

1. Introduction

Data visualization is a quick and simple technique to depict complicated ideas for improved comprehension and intuition graphically. It must find diverse relationships and pattern concealed by the massive of data. Still, it can be challenging to display huge amounts of data that is very diverse in format. In this paper major challenges faced by big data visualization will be discussed.

The five (5) of big data. Big data is a term that describe data too large and complex to store in traditional database. The five Visualizations (Vs) of big data are as mention [1]:

i. Volume – the amount of data generated.
ii. Velocity – the speed at which data is generated, collected and analysed.

* *Corresponding author.*
E-mail address: izyanizzati@utm.my

https://doi.org/10.37934/araset.51.1.150159

iii. Variety – the different types of structured, unstructured and semi-structured.
iv. Value – the ability to turn data into useful insights.
v. Veracity – trustworthiness in terms of quality and accuracy

This paper aims to address big data issues and challenges while proposing an effective visualization technique to enhance the interpretation and visualization of clustering results. By directly consider these challenges and utilizing appropriate visualization methods, researchers and practitioners can improve their understanding and practical application of clustering analysis in the field of data analysis. The findings of this research provide valuable insights and practical strategies for overcoming these challenges, contributing to the advancement of knowledge, and improving the overall effectiveness of data analysis approaches.

People, businesses, and devices have all become data factories that bring out vast amounts of information every day because of how fast technology changes. Internet users generate 1.7 megabytes of data per second [2]. It makes 146 gigabytes of data creation per day. Statista mentioned that by 2025, the world would have created more than 180 zettabytes of data [3]. One zettabyte is equal to a trillion gigabytes.

These resources are growing exponentially, resulting in massive high-dimensional data. The curse of dimensionality, which Richard E. Bellman came up with, is a big problem that often comes up with high-dimensional data [4]. Analysing and managing high-dimensional data presents many challenges. Dimension refers to the addition of a new variable. As variables get added, data space becomes increases. Establishing prediction accuracy becomes difficult. It also makes the available data spaces empty resulting in the curse of dimensionality.

On top of that, visualising such high-dimensional datasets is challenging too. Reducing the number of dimensions is needed to make the data easy for humans to understand visually. However, it is algorithmically unstable and expensive to do. It also causes a lag in rendering time, making visualisation patterns and trends difficult to identify [5]. It indicates that the correlation between variables (dimensions) and records in a dataset is hard to establish. One way to understand high-dimensional data is to display it in a low-dimensional plane [5]. High-dimensional data can be displayed in a clustered result through visualisation approaches. Thus, cluster analysis can be used. Clustering is finding groups of similar data based on their attributes [7]. Proper clustering is a helpful technique for statistical data analysis.

An effective clustering of the high-dimensional dataset is an important research issue in data mining [7]. Although cluster analysis results are obtained in the form of raw data, however, it is not easy to understand [8]. Humans have difficulty getting vital information from the cluster's analysis in a limited time. Hence, another method is required to transform the cluster analysis result into a low-dimensional space.

Visual cluster exploration is one of the effective ways to visualise clusters in a low-dimensional space. This method can be a tool that visualises clusters' results, which can provide helpful summaries and help improve decision-making [9]. What is needed for a user or information provider is a simple, easy-to-understand, time-saving, and efficient way to present data content and understand the meaning of the data. No matter what data visualisation technology tool is used, the goal should be to meet the needs of some users [10].

Implementing dimensionality projection techniques and plotting them in a lower dimension is a straightforward choice. Since 1996, Keim & Kriegel [11] have identified several well-known techniques for visualising high-dimensional datasets. There are pixel-oriented (PO), geometric projection (GP), and icon-based (IB). Madalena *et al.,* [12] observed that GP are more flexible, being able to represent quantitative and qualitative data. GP is a linear transformation technique that maps

high-dimensional data space into two-dimensional space. GP also provide a good overview of the data, assigning no priorities to represent its attributes.

However, when applying GP, there is a problem where dimensions are randomly plotted following the sequence of data columns from raw high-dimensional data [13]. It would produce the clusters to become cluttered and overlap. For example, one of GP visualization technique such as Star Coordinate (SC) technique is difficult to be manipulated by novice users. Although, SC technique offers the interactive features to the novice users for further data exploration, they encounter challenging phases in using it to reveal the clusters for a quick summary and decision. This problem has attained concentration from researchers in the last few years. Users can better comprehend clusters by manipulating interactive SC elements like dimension arrangement, angle of similar pattern dimensions, and dimension scaling. The arrangement of dimensions is essential as it influences the appearance of cluster patterns and how information is perceived [14].

Analysing existing techniques and understanding their focus work is necessary for developing some additional applicable technique that can be an improvement of the existing techniques to take advantages from earlier studies. This paper helps future researchers to clearly understand the recent status, needs, future requirements and to locate the loopholes responsible for inefficiency in clustering visualization. In detail, this study is going to answer below research questions such as:

i. What are the challenges in GP technique that affect high dimensional data cluster visualization formation?
ii. How does using SC's interactive features help solve GP problems?
iii. How can the performance of the proposed strategy for finding clusters be evaluated?

The remaining sections of the paper are structured as follows: Section 2 is based on some related work of researches in GP visualisation techniques. Section 3 describe the scenario of SC interactive features. Section 4 focuses on the analysis of the study. Section 5 suggests some future research areas in clustering visualisation and present the conclusion and recommendations in section 6.

2. Challenges with Data Clustering Visualization Related Work

The section provides a brief overview of related work that is most relevant to the contributions of this paper: high dimensional data, data quality, cluster analysis, and visualization techniques.

2.1 High Dimensional Data

The collection of large, complex datasets has become common across a wide variety of domains, such as text mining, security, aerospace, healthcare, and many more. In addition to analytical approaches such as statistics, data mining, and machine learning, visualizing high dimensional data increasingly plays a key role in exploring and answering complex questions about such large datasets to support precision, evidence-based decision making [15].

High-dimensional dataset and outliers pose significant challenges in data analysis and clustering. In high-dimensional data, where the number of features is large compared to the number of observations, computational complexity and visualization difficulties arise, making it harder to discern meaningful patterns. Outliers, which are data points that deviate significantly from the overall data distribution, can greatly influence clustering results, and disrupt the identification of coherent clusters. By addressing both high dimensionality and outliers, researchers and practitioners can

improve the quality and reliability of clustering analyses, leading to more insightful interpretations of the data [16].

Briefly described, high dimensional data consists of one or many related data tables. A data table is a structured format that is usually organized in rows and columns. A spreadsheet is one example of a data table. Column can also be presented as a field, dimension, attribute or even variable. Then, the number of attributes is known as the data dimensions. Rows act as an object, tuple, data case, data point, data item, data observation or record [17]. Additionally, each record corresponds to observation, measurement, and transaction. It is a simplified form of data matrix where m represents columns (dimensions) and n represents rows (records). These records are usually represented as points (vectors) in a multi-dimensional space. High dimensional data matrix formation is shown in Figure 1.

	F_1	F_2	F_3	F_4	...	F_m
R_1	V_{11}	V_{12}	V_{13}	V_{14}	...	V_{1m}
R_2	V_{21}	V_{22}	...	...	...	...
R_3	V_{31}	...	...	...	...	...
R_4	...	...	...	...	...	...
...	...	...	...	...	...	...
R_n	...	...	...	...	...	V_{nm}

F_m: Data features/ dimensions
R_n : Data records
V_{nm}: Vector points

Fig. 1. High dimensional data structure

High-dimensional data is often referred to as multi-dimensional or multi-aspect or multi-modal data throughout the literature [17]. In 2004, Santos & Brodlie [18] found that collection of data represents the relationship between data dimensions and records. The relationship of data is difficult to analyse, understand and interpret when the structure of data becomes complex. To search for hidden information and the relationship that occurs in the dataset, cluster analysis is implemented. However, interpreting the results of clusters analysis from raw data is challenging. Thus, visualization can be used as a tool to transform the results of clusters analysis into an efficient visual presentation.

2.2 Data Quality

Data that is to be displayed is an important matter that needs to be understood during visualization. The initial step in designing visualization is to examine the characteristic of data [19]. Steinbach & Kumar (2004) summarized three typical types of data features and common data scales as shown in Table 1 and Table 2 [19].

Table 1
Different feature types

Feature type		
Type	Description	Example
Binary	Two values	true or false
Discrete	A finite number of values or integers	Counts
Continuous	An effectively infinite number of real values	Weight

Table 2
Different feature scales

Feature Scale			
Scale	Sub scale	Description	Example
Qualitative	Nominal	The values are just different names	colours
	Ordinal	The values reflect an ordering	Good, better, best
Quantitative	Interval	The differences between values are meaningful	temperature (C^o)
	Ratio	The scale has an absolute zero so that ratios are meaningful	pressure

2.3 Cluster Analysis

High dimensional data require platforms of processing to enable better process optimization, insight discovery and decision-making.

Cluster analysis of a high dimensional data aims to partition a large data set into meaningful subgroups of subjects [37]. The goal is to put together groups of objects that are alike but different from other groups.

Based on a similarity measure between different subjects, data are divided according to a set of specified characteristics. Clusters analysis reveals patterns and correlation. Cluster analysis helps understand and recover high-dimensional data. Clusters that summarise a few groups of subjects help users decide quickly.

Clustering relies on the measurement of "closeness" or "similarity" between data points in order to form clusters. Consequently, clustering patterns serve to illustrate the similarities and differences within the data. The calculation of data record similarities is accomplished through the utilization of a distance measure [20]. It is important to note that the choice of distance measure significantly impacts the outcomes of cluster analysis. The selection of an appropriate distance measure is crucial to obtain accurate and meaningful clustering results.

2.4 Visualization Techniques

Cluster analysis and proper visualization enable the discovery of behavioural patterns or features of high dimensional data. Visualization is a way of presenting the results of cluster analysis so it can let viewers easily understand the clustering results and draw valuable conclusions about the dataset. This can be achieved by applying visualization techniques. The visualization techniques commonly disclose cluster patterns and the number of *k* clusters in a dataset. Among common visualization techniques is GP – as mentioned in section 1. Some of the examples of existing GP visualization techniques are Parallel Coordinates (PC), Star Coordinates (SC), and Scatter Plot (SP). Each one of them incorporates several interactive features. PC includes polygonal line, dimension manipulation, and brushing. While SC incorporates circular arrangement in term of clockwise direction, dimension rotation in angular manner, and dimension scaling. As for SP, it is composed of X-Y plot feature [21].

3. Dimension Management

High dimensional data bring an important issue to existing multi-dimensional visualization techniques - dimension management. Without effective dimension management, such as dimension ordering, spacing, and filtering, high dimensional visualizations can be cluttered and difficult for users to navigate the data space. The order of dimensions is crucial for the effectiveness of a large number of visualization techniques [22]. For example, in many high dimensional visualization techniques, such as Parallel Coordinates, Star Glyphs, Circle Segments and Recursive Pattern, the dimensions

have to be positioned in some one- or two- dimensional arrangement on the screen. This chosen arrangement of dimensions can have a major impact on the expressiveness of the visualization because the relationships among adjacent dimensions are easier to detect than relationships among dimensions positioned far from each other.

In addition, Ventocilla and M. Riveiro [22] concluded that, as data dimensionality increases, cluster quality measures are likely to produce estimates which are different from those perceived by users using any of the visualization techniques. It seems that the confidence, intuitiveness, and difficulty to estimate the number of clusters, as well as the accuracy of the estimations, are influenced by both the complexity of the patterns inherent in the data set, and in the capacity of a visualization techniques to disclose them. Furthermore, it is worth noting that perceived usability is not bound to the visual encoders.

4. Analysis of Dimension Management

The section provides an analysis of dimension management that are related to this study. The related analysis involved dimension arrangement, angle of similar patterned dimensions and dimension scaling. These will be furthered explained in the following section.

4.1 Dimension Arrangement

In such datasets, the dimension p of variables is much larger than the sample size n, but only a small number of variables are believed to be significantly relevant to the response of interest. It is imperative to perform a screening stage for relevant variables before a formal statistical model building procedure. This is done in order to extract truly useful underlying information from the data [4].

In this section, the focus is on dimension arrangement The concept of dimension arrangement involves the ability to rotate the axis in various directions, as highlighted by researchers [23-25]. This feature provides users with the flexibility to obtain the desired perspective when analysing data. By manipulating the dimensions and adjusting their orientations, users can effectively explore the data and gain valuable insights from different viewpoints. The incorporation of dimension arrangement as an interactive feature enhances the interpretability and customization of data analysis, allowing users to uncover meaningful patterns and relationships within the dataset.

There are several studies which had focused on dimension arrangement. The first study was done in year 1998 by Ankerst, Berchtold, & Keim [21,23]. They stated that the order and arrangement of dimensions play a vital role in presenting many quality visualization techniques. These techniques involve scatterplot, SC and PC as mentioned in the previous section. Dimension arrangement issue has been shown to be an N-P problem and they suggested heuristic algorithm to solve the problem. Heuristic algorithm can determine the similarity of each data dimensions. Then, data dimensions with similar behaviours will be placed next to each other. Other researchers like Yang, Peng, Ward, & Rundensteiner [26] had proposed an interactive hierarchical ordering of the dimensions. It is based on their similarities which can improve and reduce the complexity of the ordering. In the following year, Ward & Rundensteiner [27] applied the concept of clutter-based dimension ordering in various visualization techniques. The performance in reducing the visual clutter was recorded. Then, Sun, Tang, Tang, & Xiao [28] came out with their idea of designing dimension configuration strategy. Their idea is to optimize the order and angle of the dimension axes. They use diameter as the dimension axis instead of radius. In 2010, Di Caro, Frias-Martinez, & Frias-Martinez presented on understanding the relationship between the arrangement of dimensions and the quality of visualization using the

Radviz technique. Garcia *et al.,* [31,34-36] also studied how the order of data dimension can have an impact on revealing pattern and clustering, enabling users to understand them easier. Lastly, Wang *et al.,* [11] studied about determining which dimensions are relevant or irrelevant to be displayed in the SC layout which contributes to clustering. The overview of arranging dimension is depicted in Figure 2.

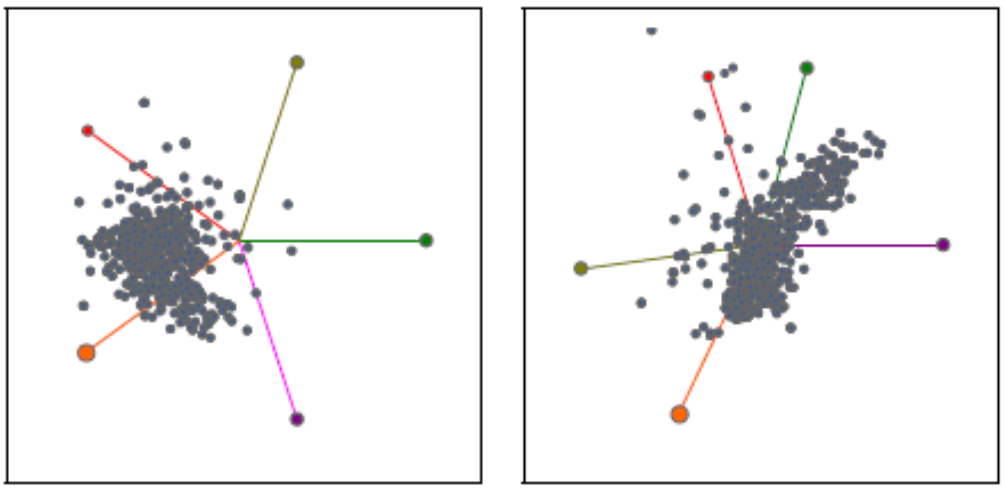

Fig. 2. Rotation operation: original mapping (left) and resulted mapping (right) [10]

Thus, it is very difficult for organizations to focus on the relevant information content. As data volumes and varieties grow, processing and consuming the insights generated becomes challenging.

4.2 Angle of Similar Patterned Dimensions

As discussed in the previous sub-section, manipulating dimension is well explored. However, positioning dimension-based angle is not widely investigated. There are numerous studies in dimension arrangement without detailing the importance of angle between each plotted dimensions [30-33]. They are focussing on users who have basic knowledge in using SC technique. Therefore, when plotting the dimensions, it will be based on their previous experience using SC technique. They manipulate the angle of each dimension to get clearer clusters appearance without telling the reason of positioning the dimension in such angle.

4.3 Dimension Scaling

Another interactive feature, dimension scaling is enriched further to enable the detection of existing cluster more clearly. As stated previously, dimension scaling can be achieved by changing the length of each dimension axis. Users can pick the end point of an axis and push or pull towards or away from the origin [35]. Dimension scaling indicates less or more the contribution of a particular data dimension on the resultant visualization [35]. Through scaling dimension, users are able to expand or collapse the clusters. Thus, clusters patterns can be distinguished clearly into a specific group as if the clusters are expended. While for collapse clusters, it can be concluded as a larger group of data that share the similarity features.

Another interaction features that are usually discussed in many papers is dimension scaling [24-26]. However, the studies on dimension scaling do not get much attention [34] they discussed this feature in a very general way. They scale the dimensions based on their experience of using SC technique and also implement the trial-and-error method [38].

Dimension scaling also contributes in obtaining obvious clusters after applying dimension arrangement. The scaling operation is illustrated in Figure 3.

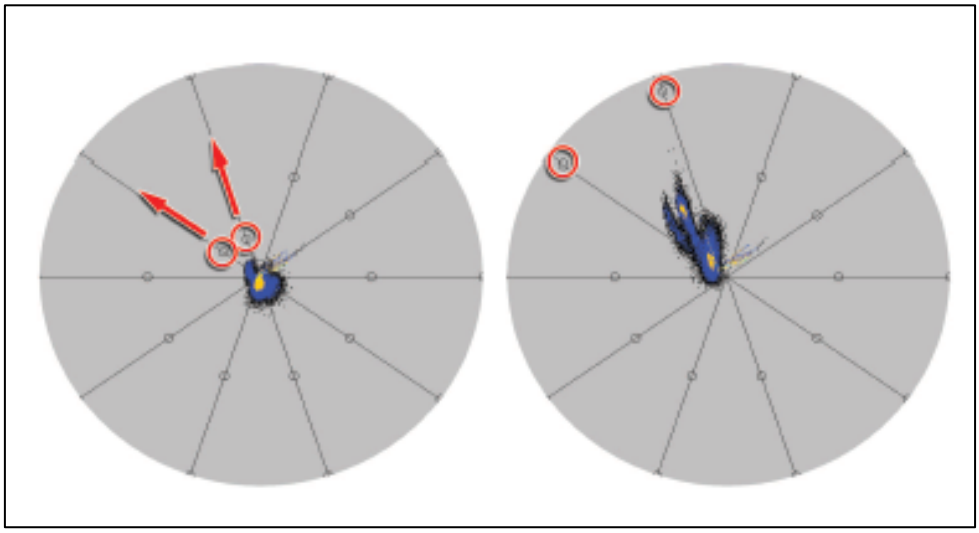

Fig. 3. Scale operation: original mapping (left) and resulted mapping (right) [12]

5. Suggested Future Works

In the future, it is imperative to consider the diverse types of data that are being generated in various forms. As data continues to evolve, it is essential to develop comprehensive guidelines for non-expert users to utilize clustering methods, such as the Silhouette Coefficient (SC) method, without requiring in-depth knowledge of the underlying technical details [39]. By providing a user-friendly interface and clear instructions, non-expert users can effectively apply the SC method to their datasets. Moreover, the SC method can be enhanced to initially display the most important and appropriately scaled dimensions during the visualization process. This approach allows users to quickly gain an overview of the data and reduces the need for trial-and-error exploration to identify distinct cluster patterns. By streamlining the clustering process, this advancement enables users to save time and facilitates more efficient and accurate data analysis [40].

6. Conclusion and Recommendations

In conclusion, as the dimensionality of data increases, it becomes increasingly challenging to effectively comprehend and describe the data. In order to properly interpret and summarize the data, it is necessary to manipulate the arrangement of dimensions. Prior to conducting a cluster analysis, it is crucial to gain valuable insights into the data by observing its similarities. Additionally, the identification of dimensions that contribute to clustering patterns is essential. This entails selecting and visualizing only the significant dimensions, thereby reducing clutter and enhancing the interpretability of the data visualization. By employing these approaches, researchers can overcome the challenges posed by high-dimensional data and gain meaningful insights from the clustering analysis.

Acknowledgement
This work was funded by the Ministry of Higher Education under Fundamental Research Grant Scheme (FRGS/1/2021/ICT03/UTM/02/2).

References

[1] Genender-Feltheimer, Amy. "Visualizing high dimensional and big data." *Procedia Computer Science* 140 (2018): 112-121. https://doi.org/10.1016/j.procs.2018.10.308
[2] Wise, Jason. "How Much Data Is Generated Every Day In 2023?(New Stats)." *Earthweb, Date Views* 3 (2023): 2023.
[3] IDC, Statista. "Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025 (in zettabytes)." *pp. Graph): Statista* (2021).
[4] Bellman, Richar E. "Adaptive Control Processes: A Guided Tour." *Princeton University Press,* (2015).
[5] Probst, Daniel, and Jean-Louis Reymond. "Visualization of very large high-dimensional data sets as minimum spanning trees." *Journal of Cheminformatics* 12, no. 1 (2020): 1-13. https://doi.org/10.1186/s13321-020-0416-x
[6] Khalid, Abdul, and Izyan Izzati Kamsani. "Star Coordinate Dimension Arrangement using Euclidean Distance and Pearson Correlation." *Indonesian Journal of Electrical Engineering and Computer Science* 12, no. 1 (2018): 348-355. https://doi.org/10.11591/ijeecs.v12.i1.pp348-355
[7] Agarwal, Parul, Shikha Mehta, and Ajith Abraham. "A meta-heuristic density-based subspace clustering algorithm for high-dimensional data." *Soft Computing* 25 (2021): 10237-10256. https://doi.org/10.1007/s00500-021-05973-1
[8] Satre-Meloy, Aven, Marina Diakonova, and Philipp Grünewald. "Cluster analysis and prediction of residential peak demand profiles using occupant activity data." *Applied Energy* 260 (2020): 114246. https://doi.org/10.1016/j.apenergy.2019.114246
[9] Mazher, Abeer. "Visualization framework for high-dimensional spatio-temporal hydrological gridded datasets using machine-learning techniques." *Water* 12, no. 2 (2020): 590. https://doi.org/10.3390/w12020590
[10] Bai, Shengyuan, Xiangyi Zhou, You Lyu, Jiali Wang, and Chengxiang Pan. "Data visualization model methods and techniques." In *IOP Conference Series: Earth and Environmental Science*, vol. 252, no. 5, p. 052063. IOP Publishing, 2019. https://doi.org/10.1088/1755-1315/252/5/052063
[11] Keim, Daniel A., and H-P. Kriegel. "Visualization techniques for mining large databases: A comparison." *IEEE Transactions on knowledge and data engineering* 8, no. 6 (1996): 923-938. https://doi.org/10.1109/69.553159
[12] Blumenschein, Michael, Xuan Zhang, David Pomerenke, Daniel A. Keim, and Johannes Fuchs. "Evaluating reordering strategies for cluster identification in parallel coordinates." In *Computer Graphics Forum*, vol. 39, no. 3, pp. 537-549. 2020. https://doi.org/10.1111/cgf.14000
[13] Karahoca, Adem, ed. *Advances in data mining knowledge discovery and applications*. BoD–Books on Demand, 2012. https://doi.org/10.5772/3349
[14] Lu, Liangfu, Wenbo Wang, and Zhiyuan Tan. "Double-arc parallel coordinates and its axes re-ordering methods." *Mobile Networks and Applications* 25 (2020): 1376-1391. https://doi.org/10.1007/s11036-019-01455-9
[15] Feng, Kang, Yunhai Wang, Ying Zhao, Chi-Wing Fu, Zhanglin Cheng, and Baoquan Chen. "Cluster aware star coordinates." *Journal of Visual Languages & Computing* 44 (2018): 28-38. https://doi.org/10.1016/j.jvlc.2017.11.003
[16] Kamalov, Firuz, and Ho Hon Leung. "Outlier detection in high dimensional data." *Journal of Information & Knowledge Management* 19, no. 01 (2020): 2040013. https://doi.org/10.1142/S0219649220400134
[17] Thudumu, Srikanth, Philip Branch, Jiong Jin, and Jugdutt Singh. "A comprehensive survey of anomaly detection techniques for high dimensional big data." *Journal of Big Data* 7 (2020): 1-30. https://doi.org/10.1186/s40537-020-00320-x
[18] Dos Santos, Selan, and Ken Brodlie. "Gaining understanding of multivariate and multidimensional data through visualization." *Computers & Graphics* 28, no. 3 (2004): 311-325. https://doi.org/10.1016/j.cag.2004.03.013
[19] Steinbach, Michael, Levent Ertöz, and Vipin Kumar. "The challenges of clustering high dimensional data." In *New directions in statistical physics: econophysics, bioinformatics, and pattern recognition*, pp. 273-309. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. https://doi.org/10.1007/978-3-662-08968-2_16
[20] Goel, Shivani. "Systematic review of clustering high-Dimensional and large datasets." (2018).
[21] Ankerst, Mihael, Stefan Berchtold, and Daniel A. Keim. "Similarity clustering of dimensions for an enhanced visualization of multidimensional data." In *Proceedings IEEE symposium on information visualization (Cat. No. 98TB100258)*, pp. 52-60. IEEE, 1998.
[22] Ventocilla, Elio, and Maria Riveiro. "A comparative user study of visualization techniques for cluster analysis of multidimensional data sets." *Information visualization* 19, no. 4 (2020): 318-338. https://doi.org/10.1177/1473871620922166
[23] Blumenschein, Michael, Xuan Zhang, David Pomerenke, Daniel A. Keim, and Johannes Fuchs. "Evaluating reordering strategies for cluster identification in parallel coordinates." In *Computer Graphics Forum*, vol. 39, no. 3, pp. 537-549. 2020. https://doi.org/10.1111/cgf.14000

[24] Kandogan, Eser. "Visualizing multi-dimensional clusters, trends, and outliers using star coordinates." In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 107-116. 2001. https://doi.org/10.1145/502512.502530
[25] Kandogan, Eser. "Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions." In *Proceedings of the IEEE information visualization symposium*, vol. 650, p. 22. Citeseer, 2000.
[26] Rusu, Adrian, Confesor Santiago, Andrew Crowell, and Eric Thomas. "Enhanced star glyphs for multiple-source data analysis." In *2009 13th International Conference Information Visualisation*, pp. 183-190. IEEE, 2009. https://doi.org/10.1109/IV.2009.65
[27] Yang, Jing, Wei Peng, Matthew O. Ward, and Elke A. Rundensteiner. "Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets." In *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*, pp. 105-112. IEEE, 2003.
[28] Peng, Wei, Matthew O. Ward, and Elke A. Rundensteiner. "Clutter reduction in multi-dimensional data visualization using dimension reordering." In *IEEE Symposium on Information Visualization*, pp. 89-96. IEEE, 2004.
[29] Sun, Yang, Jiuyang Tang, Daquan Tang, and Weidong Xiao. "Advanced star coordinates." In *2008 The Ninth International Conference on Web-Age Information Management*, pp. 165-170. IEEE, 2008. https://doi.org/10.1109/WAIM.2008.20
[30] Di Caro, Luigi, Vanessa Frias-Martinez, and Enrique Frias-Martinez. "Analyzing the role of dimension arrangement for data visualization in radviz." In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 125-132. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. https://doi.org/10.1007/978-3-642-13672-6_13
[31] Zanabria, Germain Garcia, Luis Gustavo Nonato, and Erick Gomez-Nieto. "iStar (i*): An interactive star coordinates approach for high-dimensional data exploration." *Computers & Graphics* 60 (2016): 107-118. https://doi.org/10.1016/j.cag.2016.08.007
[32] Cooprider, Nathan D., and Robert P. Burton. "Extension of star coordinates into three dimensions." In *Visualization and Data Analysis 2007*, vol. 6495, pp. 256-265. SPIE, 2007. https://doi.org/10.1117/12.703359
[33] Bordignon, Alex Laier, Rener Castro, Hélio Lopes, Thomas Lewiner, and Geovan Tavares. "Exploratory visualization based on multidimensional transfer functions and star coordinates." In *2006 19th Brazilian Symposium on Computer Graphics and Image Processing*, pp. 273-280. IEEE, 2006. https://doi.org/10.1109/SIBGRAPI.2006.17
[34] Maalej, Abdelaziz, Nancy Rodriguez, and Olivier Strauss. "Survey of multidimensional visualization techniques." In *CGVCVIP'12: Computer Graphics, Visualization, Computer Vision and Image Processing Conference*, pp. N-A. 2012.
[35] Borland, David, Wenyuan Wang, Jonathan Zhang, Joshua Shrestha, and David Gotz. "Selection bias tracking and detailed subset comparison for high-dimensional data." *IEEE Transactions on Visualization and Computer Graphics* 26, no. 1 (2019): 429-439. https://doi.org/10.1109/TVCG.2019.2934209
[36] Cutillo, Luisa. "Parametric and multivariate methods." (2019): 738-746. https://doi.org/10.1016/B978-0-12-809633-8.20335-X
[37] Zhao, Bangxin. "Analysis challenges for high dimensional data." PhD diss., The University of Western Ontario (Canada), 2018.
[38] Miller, Matthias, Xuan Zhang, Johannes Fuchs, and Michael Blumenschein. "Evaluating ordering strategies of star glyph axes." In *2019 IEEE Visualization Conference (VIS)*, pp. 91-95. IEEE, 2019. https://doi.org/10.1109/VISUAL.2019.8933656
[39] Schmidt, Johanna. "Usage of Visualization Techniques in Data Science Workflows." In *VISIGRAPP (3: IVAPP)*, pp. 309-316. 2020. https://doi.org/10.5220/0009181903090316
[40] Qin, Xuedi, Yuyu Luo, Nan Tang, and Guoliang Li. "Making data visualization more efficient and effective: a survey." *The VLDB Journal* 29 (2020): 93-117. https://doi.org/10.1007/s00778-019-00588-3