Enhancing Visualization of Multidimensional Data by Ordering Parallel Coordinates Axes

Ayman Nabil[1]
Faculty of Compuuter Science
Misr International University
Cairo,Egypt

Karim M. Mohamed[2], Yasser M. Kamal[3]
College of Computing and Information Technology
AASTMT, Cairo Branch
Cairo,Egypt

***Abstract*—Every year business is overwhelmed by the quantity and variety of data. Visualization of Multi-dimensional data is counter-intuitive using conventional graphs. Parallel coordinates are proposed as an alternative to explore multivariate data more effectively. However, it is difficult to extract relevant information through the parallel coordinates when the data are Multi-dimensional with thousands of lines overlapping. The order of the axes determines the perception of information on parallel coordinates. This paper proposes three new techniques in order to arrange the axes in the most significant relation between the datasets. The datasets used in this paper, for Egyptian patients, with many external factors and medical tests. These factors were collected by a questionnaire sheet, made by medical researchers. The first Technique calculates the correlation between all features and the age of the patient when they get diabetes disease. The second technique is based on merging different features together and arranging the coordinates based on the correlations values. The Third Technique calculates the entropy value for each feature and then arrange the parallel coordinates in descending order based on the positive or negative values. Finally based on the result graphs, we conclude that the second method was more readable and valuable than the other two methods.**

Keywords—Parallel coordinates; visualization; correlation coefficient; entropy function

I. Introduction

In the recent studies of computer science and technologies, an accelerating information explosion is being witnessed. In digital universe today about 2.7 Zeta bytes executed continuously [1]. Based on the Estimations and studies presented by the International Data Corporation (IDC), they suspect that by 2020 business transactions on the internet-business-to-business and business-to-consumer will reach 450 billion per day [2]. Moreover, analysis and knowledge are power and in order to analysis and interpret these huge amounts of data, Users have to use tools to visualize this data. These visualization tools can assist in retrieving valuable information, which may effectively help in solving many different types of problems. One of these important tools is the Parallel coordinates, method of visualizing high dimensional geometry and analyzing multidimensional data [3].

These days the data and its dimensions' increase rapidly which results too much interference in the coordinates and timelines of the parallel coordinates, lead to obstacles in analyzing. For this reason, many papers are presented to solve these difficulties and complexities to interpret this data [4] [5] [6] [7]. This interference could lead to a complexity in reading or interpreting the data.

Previous research has proposed exploratory techniques to enhance the visualization of multidimensional data. Within the last 20 years researches focused on Techniques to reduce the number of poly-lines or reducing or reordering the parallel axes [8] [9]. This paper introduces novel techniques for reordering the factors of the data based on the correlation coefficient calculations. The goal of these techniques is to facilitate the readiness and the complexity of the parallel coordinates. The paper categorized into different sections, the proposed methods, a detail explanation about the new techniques proposed. The results and discussion the comparison between the three techniques and finally the conclusion section.

II. Background and Related Work

The Parallel coordinates is an interactive visualization, and is the most used for multidimensional data visualization. It was developed and popularized by Alfred Inselberg [10]. Improving the parallel coordinates plot is a highly active research topic. There are some techniques proposed in previous research that attempted to enhance the readability of the results by applying clustering techniques or sampling polynies [11] [12] [13] [14]. Moreover the readiness and effectiveness of the parallel coordinates depends on ordering the dimensions and factors, different dimension ordering techniques were presented [15] [16] [17].

Other papers proposed new methods for interpreting the readiness of the parallel coordinate by dividing the dimensions of the datasets input into groups of lower dimensions based on the correlations calculations; the conclusion of this technique can represent various groups of correlated dimensions in high dimensional data space [8].

Furthermore, another paper proposed the automated assistance to rearrange the order of the variable; this automation was done using a system called V-miner. Motorola engineers were affected by the new powerful enhancements and also facilitate the use of the parallel coordinates [4].

Also techniques were proposed to simplify the representation of the parallel coordinate visualization, where a new study proposed using the eye tracking. The main idea is to understand whether the parallel coordinate visualizations are easy to be perceived or not.

From the results of this study, the users were able to interpret and realize the parallel coordinate easily by concentration on the correct areas for the chart [18].

III. The Proposed Methods

This section will discuss the proposed methods to enhance the visualization in the parallel coordinates. The goal of using the Parallel coordinates is one of the most important techniques to visualize dataset with multidimensional datasets, the better visualization becomes obvious, and more information can be retrieved [19]. The results of the parallel coordinate visualization always confuse the reader, and could lead to difficulties to read. Past studies proved that the correlation coefficient affects the result and the interpretation of the parallel coordinate visualization [20]. The effectiveness on the interpretation and the readiness, also has an effect on the visualization between two coordinates, for instance, the parallel coordinates plot for data that have negative 1 correlation different from the parallel coordinates for data that have 1 correlation is as follows:

In Fig. 1 and 2, the correlation affects the visualization of the parallel coordinate chart, and incase the two features are correlated or not the lines interfere or move in parallel path. For this reason, this paper proposed two of the new methods based on the correlation coefficient. In order to simplify the complexity of the intervention between lines that may lead to difficulties in tracking the parallel coordinate's graphs.

Moreover, these techniques give the user a better chance to interpret and analyze the datasets more professionally. The used datasets are for Egyptian people suffering from the diabetes disease. This data was collected by the Egyptian National Research Center and was based on standard medical questionnaire. This questionnaire was prepared by specialized doctors in the diabetic field.

The goal of implementing these two methods on the diabetic patients' dataset is to reach the most significant features that affect the health of these patients and assist in triggering the diabetic disease faster in younger ages.

Fig. 1. Parallel Coordinates Plot for Data with Correlation Coefficient of 1.

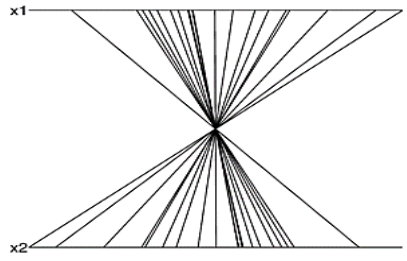

Fig. 2. Parallel Coordinates Plot for Data with Negative Correlation Coefficient.

A. Datasets

The Egyptian National Research Center compiled these Datasets based on a medical questionnaire which contains 348 patients. This questionnaire is comprised of questions regarding the risk factors that cause diabetes disease and were questions for diabetes patients. After that these forms were extracted into a statistical tool called SPSS, for doing statistical analysis on this data and finally they were exported into an excel sheet, in order to be used in experiments as shown in Fig. 3.

The datasets were collected 6-years ago. The Dataset comprised of 23 features; these features are summarized in Table I.

TABLE. I. Description Of Dataset Feature

No.	Feature name	Type	Range
1	Diabetes Age	Numeric	Real Values
2	Gender	Numeric	Categorical
3	Education	Numeric	Categorical
4	Diabetic Family member	Numeric	Categorical
5	Smoker	Numeric	Categorical
6	Cigarette number	Numeric	Real Values
7	Smoking Start Date	Date	Real Values
8	Exercising Status	Numeric	Categorical
9	Frequent Exercise per week	Numeric	Real Values
10	Exercise Type	Numeric	Categorical
11	Food Type	Numeric	Categorical
12	Healthy Food status	Numeric	Categorical
13	No of Basic Meals	Numeric	Real Values
14	Snacks Status	Numeric	Real Values
15	Snacks Number	Numeric	Real Values
16	Snack Type	Numeric	Categorical
17	Regime Status	Numeric	Categorical
18	Blood Pressure Status	Numeric	Categorical
19	Blood Fat Status	Numeric	Categorical
20	Foot Complications	Numeric	Categorical
21	Neuro Complications	Numeric	Categorical
22	Low Vision status	Numeric	Categorical
23	Wound Recovery Status	Numeric	Categorical

	DO_REGIME	FREQ_EXERCISE_PER_WEEK	NO_BASIC_MEAL	SNACK_NUMBER	FOOT_PROBLEM	EXERCISE_TYPE	SNACK_TYPE_VALUE	BLOOD
DO_REGIME	1							
FREQ_EXERCISE_PER_WEEK	0.049471547	1						
NO_BASIC_MEAL	-0.001741377	-0.097109289	1					
SNACK_NUMBER	0.021013927	-0.076879549	0.064209246	1				
FOOT_PROBLEM	-0.054684668	-0.195865106	0.23474581	0.055161172	1			
EXERCISE_TYPE	0.028923513	0.768675232	-0.068062947	-0.079239514	-0.149901169	1		
SNACK_TYPE_VALUE	0.026140543	0.005203366	0.032395066	0.335012311	0.117296333	-0.06025718	1	
BLOOD_FAT	-0.044573224	-0.017266322	0.064489211	-0.040385303	0.067797956	0.067280965	-0.144544455	
SMOKING_START_DATE	0.090459612	0.049183136	-0.152850745	-0.030259944	0.08276238	0.047886892	0.022566313	0.0769
LOW_VISION	0.055089677	0.132931471	-0.042225446	-0.040269349	0.019181563	0.153894258	0.009375393	0.0869
NEURO_DIABETES	0.059185685	0.142272781	-0.040411164	0.011691825	-0.037526699	0.294795851	-0.06545159	0.146
CIGARETTE_NO	0.027662745	-0.045657471	-0.050573461	0.029059763	0.124489857	0.034663353	0.125461118	0.1320
DIABETIC_FM_VALUE	-0.137977634	-0.054680884	0.073288527	-0.033070271	0.078070658	-0.033057412	0.024858162	0.0974
EDUCATION	-0.179526894	0.007266046	0.203857394	-0.002952057	0.205417001	0.037587988	0.09731062	0.0817
HIGH_BLOOD_PRESSURE	0.064851825	0.149703699	-0.02994862	-0.018757247	-0.000509856	0.186435257	-0.07025151	0.2703
SMOKER	0.066345604	0.097569258	-0.009107539	-0.042647419	-0.155200146	-0.040551577	-0.079195428	-0.1856
EAT_SNACKS	0.235927735	0.120180067	-0.078927561	-0.449805336	-0.109768599	0.084824091	-0.617283433	0.0894
SEX	-0.01571209	-0.075091016	-0.027877746	0.052257706	-0.086127842	-0.153473827	-0.113123887	-0.0618
WOUND	0.075880149	-0.045969185	-0.075759496	0.042787295	-0.142375358	-0.067224603	0.008152662	-0.1226
DO_EXERCISE	-0.068613467	-0.833904812	0.102927223	0.088486779	0.217020687	-0.942131273	0.066500866	-0.0364
FOOD_TYPE_VALUE	0.181435712	0.199179676	-0.17668066	-0.025509278	-0.255252988	0.129107138	-0.04574371	0.0366
VEGITABLE_FRUITS_DAILY	0.15644772	-0.039593198	-0.297781107	-0.07452418	-0.018620822	0.010087969	-0.075461071	0.0926
DIABETES_AGE	-0.019171533	-0.029319354	-0.053046479	-0.058517877	-0.055553749	-0.083544245	-0.089959581	-0.1106

Fig. 3. The Correlation Coefficient for Each Variable with Respect to the Age.

B. First Method

In the first method the datasets are categorized into independent variable and dependent variables, the dependent variable in this case is the age of the patients when they got the diabetes disease. Then calculate the correlation between all the features with the dependent variable (age). These features will be organized based on the correlation values ascending on both sides of the age variable. The positive correlation features arranged on the right hand side and the negative correlation values on the left hand side. Then a parallel coordinate chart is drawn using the TIBCO spotfire software.

The correlation was calculated based on the Pearson's correlation function; the used function is:

$$\hat{r}(x,y) = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}} \quad (1)$$

This function will measure the strength of the linear association between the two variables. n is the number of pairs data. The X and the Y variables represent the independent and the dependent variables. r is such that -1 < r < +1. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

1) Positive correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increase, values for y also increase.

2) Negative correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

3) No correlation: If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables.

In Fig. 5, it illustrates the arrangement of the positive or negative values regarding to the dependent variable.

C. Second Method

In this section a new method is proposed for rearranging the coordinates. On the first method calculated the correlations of all the features with the output value only. But the values of the most two significant correlation values are merged with the age and then calculate the correlation of these merged factors with the rest of the features to get the most two significant values to the new merged value as shown in Fig. 4.

For example, the result of the first calculation for correlation with the age factors were the high blood pressure and the smoking variable, subsequently multiply the values of these three factors, and recalculate the correlation again. Other positive and negative correlations will be resulted; hence multiply the five factors together and recalculate the correlations and so on until getting a final arrangement based on this methodology, then draw the parallel coordinate chart based on these arrangements.

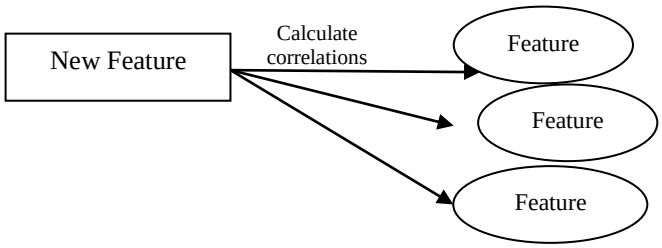

Fig. 4. Merge Features.

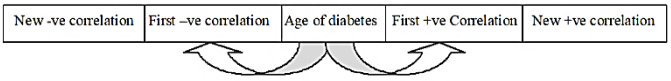

Fig. 5. Ordering Features.

D. Third Method

Third method used the entropy function, which characterizes the impurity of an arbitrary collection. The entropy always uses the information theory and is used in the decision tree algorithm to calculate the homogeneity of the datasets [21]. In this method the entropy value is being calculated for all the independent and dependent variables, then rearranging them in a descending order.

Furthermore, these features will be arranged based on the sign, where positive values on the right and the negative values on the left. Finally, plot the parallel coordinates chart with the result in ordering. The following query is used to calculate the entropy:

$$E(s) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (2)$$

Pi: the probability of class i. Compute proportion of i in the set. The higher E(s) the more information gain.

IV. Results and Discussion

The first experiment is comparing the three methods in general without using the brushing tool. The differences between the three figures are obvious. Fig. 9 is readable and easily to interpret comparing to Fig. 8 and Fig. 10, for instance in Fig. 9 there are many negative correlations easily to be tracked or analysis other than the other two figures. These negative correlations became visible after applying the 2nd new method on the parallel coordinate chart.

Moreover, some features aren't correlated in Fig. 8 and 10, where the lines crossing all around forming disorganization and complexity, for example the Neuro and cigarettes numbers factors.

Fig. 9 shows all the features are correlated to each other, this could result in enhancing the readability of the charts by reorganizing the factors or the parallel coordinates based on the correlations combination method. On the next section a comparison between two snap shots Fig. 6 represent the parallel coordinates using the 1st method and Fig. 7 represent the 2nd method. If analyzed, the lines in Fig. 7 between Cigarette, Family Diabetes History and High blood pressure features are organized and correlated.

Fig. 6, the lines are highly interventions which lead to difficulties in interpreting.

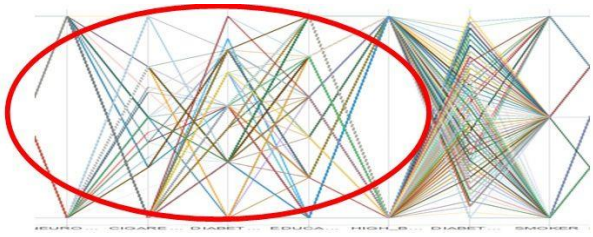

Fig. 6. 1st Method Showing Lines between Features.

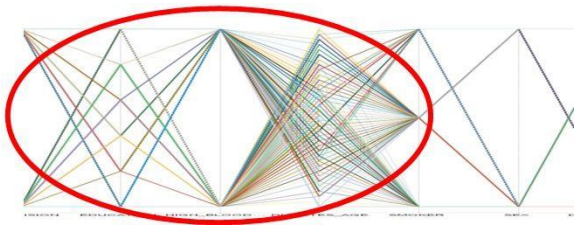

Fig. 7. 2nd Method Showing Lines between Features.

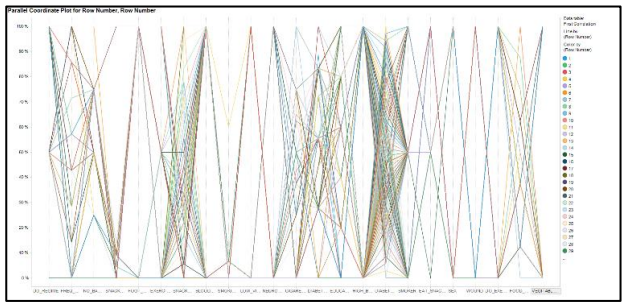

Fig. 8. Parallel Coordinates Chart based on the 1st Method Calculating the Correlation for Each Feature with Respect to Age Feature.

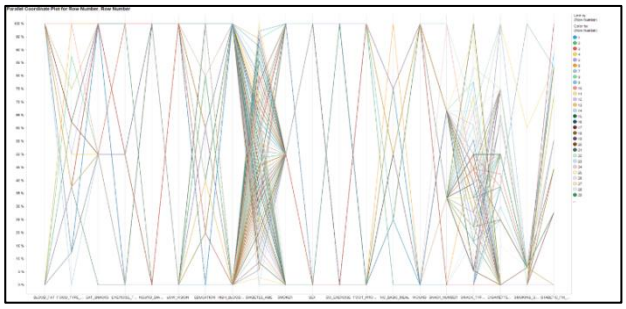

Fig. 9. Parallel Coordinates Chart based on the 2nd Method Calculating the Correlation by Merging the Previous Features with the Respect to Age Feature.

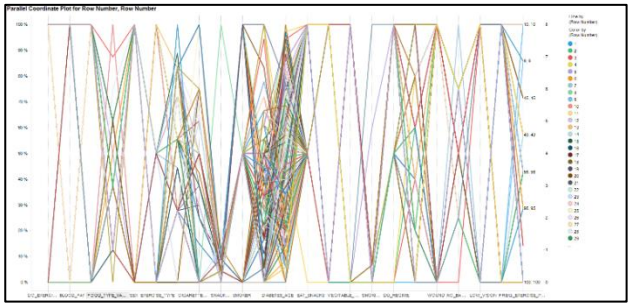

Fig. 10. Parallel Coordinates Chart based on the 3rd Method Calculating the Entropy Function for Each Feature then Rearrange the Coordinates Accordingly.

In this section, Fig. 11, 12 and 13, brushed data to focus on the Education feature and specifically the highest level of educational patients, as we can see the difference between the three graphs, where the features aren't correlated between each other in the first graph, forming random lines between different features. On the other hand, most of the features are either positive or negative correlations in the second chart.

Furthermore, as a quick notice can be reached from the second figure, all people with high level of education are much stressed and most of them suffer from high blood pressure, for this reason they probably may have a high risk to be candidate of diabetes disease at a younger age. Also the second figure is still better than the third chart.

Another example, Fig. 14, 15 and 16, when brushing the data for people who are smoking, the same result like previous charts, most of the features are significant correlated with the second method other than the first and third method. Also extracting useful information from the second chart, for instance most of the male patients are smoking and they are a positive correlation to be a candidate of the diabetes disease.

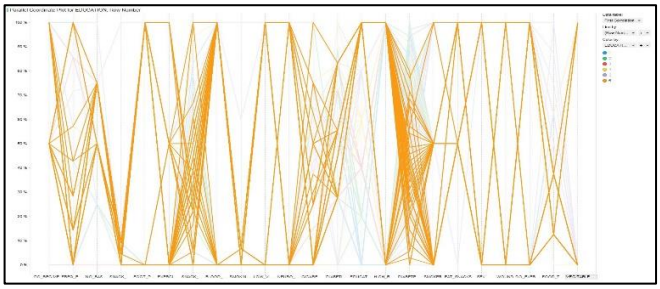

Fig. 11. 1st Method Focus on the Education Feature and Specially the Highest Level of Education.

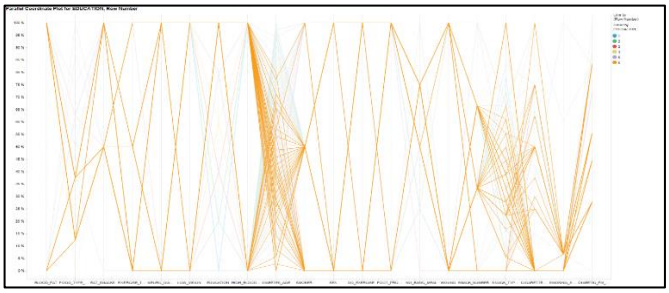

Fig. 12. 2nd Method Focus on the Education Feature and Specially the Highest Level of Education.

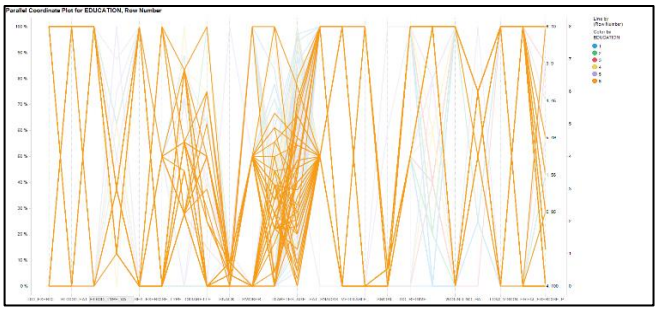

Fig. 13. 3rd Method Focus on the Education Feature and Specially the Highest Level of Education.

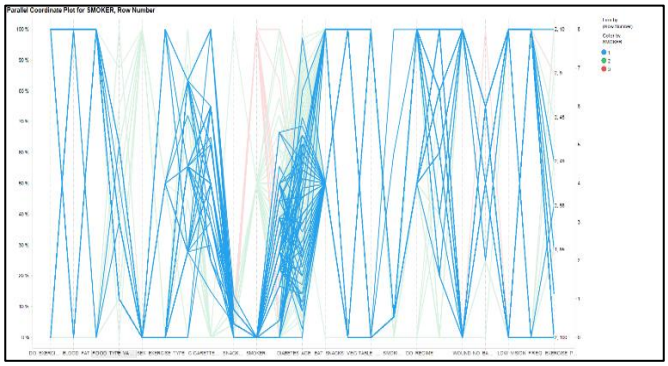

Fig. 14. 1st Method Focus on Smoking Feature and Specially the Smokers.

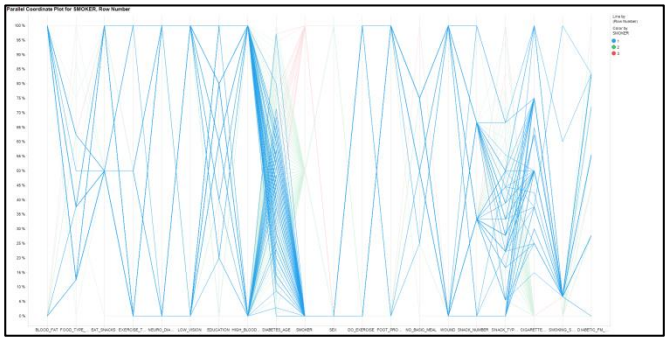

Fig. 15. 2nd Method Focus on Smoking Feature and Specially the Smokers.

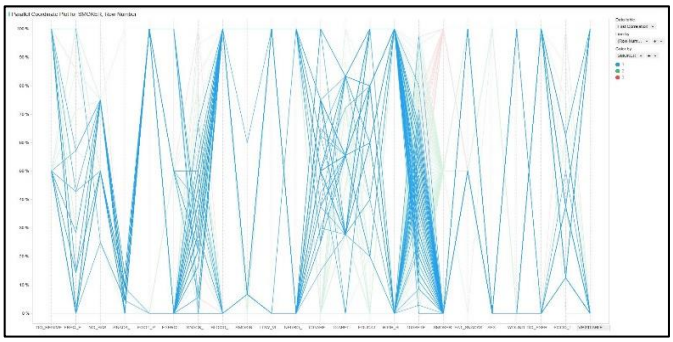

Fig. 16. 3rd Method Focus on Smoking Feature and Specially the Smokers.

V. Conclusion

In this paper, three techniques to reorder the coordinates of the charts were introduced. Two of these techniques based on the correlation coefficient and the third one based on the entropy function. The goals of these techniques to enhance the parallel coordinate visualization and facilitate the interpretation of data.

Concluding based on the analysis and by comparison, the second method results a better visualization than others. New information was interpreted and extract from the charts. In the future work a plan to merge between the three techniques with the clustering methodology. Moreover, further analysis and discussion will be held between the old and the new charts.

References

[1] G. Noseworthy, "Infographic: Managing the Big Flood of Big Data in Digital Marketing," [Online]. Available: http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digitalmarketing/.

[2] V. T. M. S. Carrie MacGillivary, "IDC's Worldwide Internet of Things Taxonomy," IDC, 2015.

[3] M. G. B. Akbar, "Data Analytics Enhanced Data Visualization and Interrogation with Parallel Coordinates Plots," in 26th International Conference on Systems Engineering, ICSEng 2018 , 2019.

[4] T. M. T. A. S. Kaidi Zhao, "Detecting Patterns of Change Using Enhanced Parallel Coordinates Visualization," in ICDM '03 Proceedings of the Third IEEE International Conference on Data Mining, ,2003.

[5] X. Y. Z. G. X. Huamin Qu, "Scattering Points in Parallel Coordinates," IEEE Transactions on Visualization & Computer Graphics, vol. 15, pp. 1001-1008,, 2009.

[6] W. Sun and S. Wang, "A new data mining method for early warning landslides based on parallel coordinate," in Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services, 2011.

[7] G. R. ,. T. J. ,. F. L. D. A. a. R. B. Joris Sansen *, "Visual Exploration of Large Multidimensional Data Using Parallel Coordinates on Big Data Infrastructure," Informatics, vol. 7, 2017.

[8] Z. Y. T. I. F. Haruka Suematsu, "Arrangement of Low-Dimensional Parallel Coordinate Plots for High-Dimensional Data Visualization," in 2013 17th International Conference on Information Visualisation, 2013.

[9] K. Zhao, B. Liu, T. Tirpak and A. Schaller, "Detecting Patterns of Change Using Enhanced Parallel Coordinates Visualization," in Third IEEE International Conference on Data Mining, 2003.

[10] B. D. Alfred Inselberg, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in Proceedings of the First IEEE Conference on Visualization: Visualization `90, 1990.

[11] M. W. E. R. Y. Fua, "Hierarchical parallel coordinates for exploration of large datasets," in Proceedings Visualization '99 (Cat. No.99CB37067), 1999.

[12] P. L. M. J. M. C. J. Johansson, "Revealing Structure within Clustered Parallel Coordinates Displays,," in INFOVIS '05 Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization, 2005.

[13] F. C. P. A. A. M. V. Elena Geanina ULARU, "Perspectives on Big Data and Big Data Analytics," Database Systems Journal, Vols. vol. III, no. 4/2012, pp. 3-14, 2012.

[14] X. Y. H. Q. W. C. B. C. H. Zhou, "Visual Clustering in parallel Coordinates," in EuroVis'08 Proceedings of the 10th Joint Eurographics / IEEE - VGTC conference on Visualization, 2008.

[15] S. T. S. J. Hemant Mekwana, "Axes Re-ordering in parallel coordinate for pattern Optimization," International Journal of Computer Applications , vol. Volume 40– No.13, pp. 42-47, 2012.

[16] L. F. Lu, M. L. Huang and T.-H. Huang, "A New Axes Re-ordering Method in Parallel Coordinates visualization," in 11th International Conference on Machine Learning and Applications, 2012.

[17] J. Z. B. H. R. Rosenbaum, "Progressive Parallel Coordinates," in IEEE Pacific Visualization Symposium, 2012.

[18] H. Siirtola, T. Laivo, T. Heimonen and K.-J. Räihä, "Visual Perception of Parallel Coordinate Visualizations," in 13th International Conference Information Visualisation, 2009.

[19] Tran Van Long, "Visualizing High-density Clusters in Multidimensional Data," Jacobs University, 2009.

[20] J.-B. M. a. J. J. V. W. J. Li, "Judging correlation from scatter plots and parallel coordinate plots," Information Visualization, vol. Volume 9 Issue 1, pp. 13-30, 2010.

[21] C. Kamath, Scientific Data Mining : A practical Perspective,, Society for Industrial and Applied Mathematics, 2009.