Specifying the visualizing type of multi-dimensional data

IOP Conf. Series: Journal of Physics: Conf. Series **1294** (2019) 032033 doi:10.1088/1742-6596/1294/3/032033

Specifying the visualizing type of multi-dimensional data

Zainab H Majeed[1], Basaeir Y Ahmed[2] and Safa A Najim[1]

1 Computer Information System Department, College of CS and IT, Basrah University
2 College of Science, Basrah University

Abstract: Knowing the type of data is widely required to make better and faster decisions. However, the process requires the user to provide information about the configuration of the data. This paper presents the first attempt to analyze data to extract it is type automatically from multi-dimensional data sets. This is useful not only for experts but also to users, also reduces manual search effort. Layers of multi-dimensional data are formed and evaluated, and the focus is on the most efficient ones. Experiments on experimental and real data demonstrate the efficiency and effectiveness of the proposed method.

1 Introduction

Recently, the possibility of generating a visulization representation of the data is popular as the best answer. Data visualization is a method of computing that gives the visible form of complex data using graphics and imaging technology for the purpose of illustrating invisible information [4] [6]. This depends on the idea that the human visual system has ability to handle and interpret large volume of data efficiently. The need for a specific data structure is necessary to make scientific processing effective, but this possibility is often missing, especially when dealing with large volume of data that generate problems for this level of data.

Most scientific experiments using huge data, so the processing depends on the analytical side to focus on the important aspects of this data and extract the most useful things [7]. Thus, data visualization with an appropriate data structure can be integrated to get faithful solution of the data fusion [2] [10] [12].

One of the ways to identify the meaning of data need the right data model. Scientific data contains many dependent and independent variables, for example: time, spectral, spatial, etc. Each of these variables needs special representation that can be different from the other. As an attempt to add something simplistic, it is possible to deal with a certain number of variables for the purpose of reaching a possible solution, as in the regression method [8] [3].

The data sets can be defined on the basis of it's set layers. Independent variables are called dimensions, and a number of independent variables are called data dimensions, which are the basic specifications of the data. Data in one dimension are the dependent variables to represent the basic parameters. Therefore, the study of variables in one dimension is necessary to extract the required information [11][9]. An important thing for data analysis is to know the relationships between those data, in terms of linear or nonlinear relationships [5] [1].

In this paper, new method to analysis data sets by using data visualization has been suggested. Studying the relationship between the data sets and trying to extract what is important in the analysis process will be one of the basics used during processing. It is possible to display a visualization

explains the data fusion, and the right decision in choosing the appropriate processing for this data sets is possible.

2 Methodology

One of the important things in processing data correctly is to determine the data fusion in terms of linear or nonlinear structures. Thus, the focus will be on this aspect by using the data visualization to overcome this challenge. The basis idea of the proposed method is to preserve the relationships between data in one dimension as well as the other dimensions. The figure 1 gives a general idea of the proposed method.

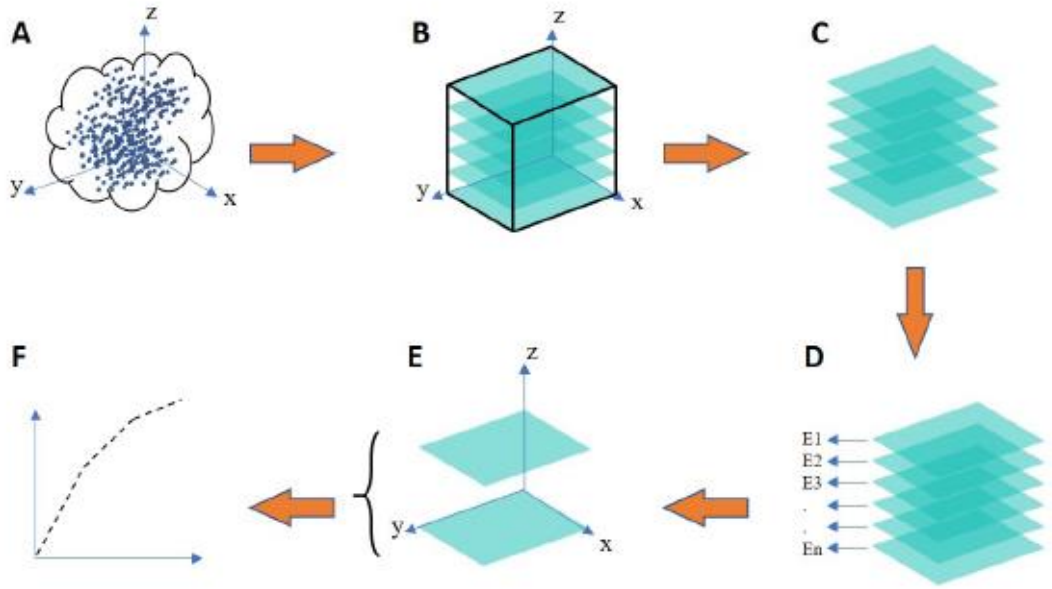


Fig. 1: *The general idea to explore the data structure of a data sets. (A) Original data sets. (B) Representing the original data sets as a set of layers. (C) The processing will be on the layers, according to their relation among them. (D) The efficiency of the all layers are computed. (E) The highest two efficiency layers are chosen. (F) linear or non-linear structure of data sets is specified.*

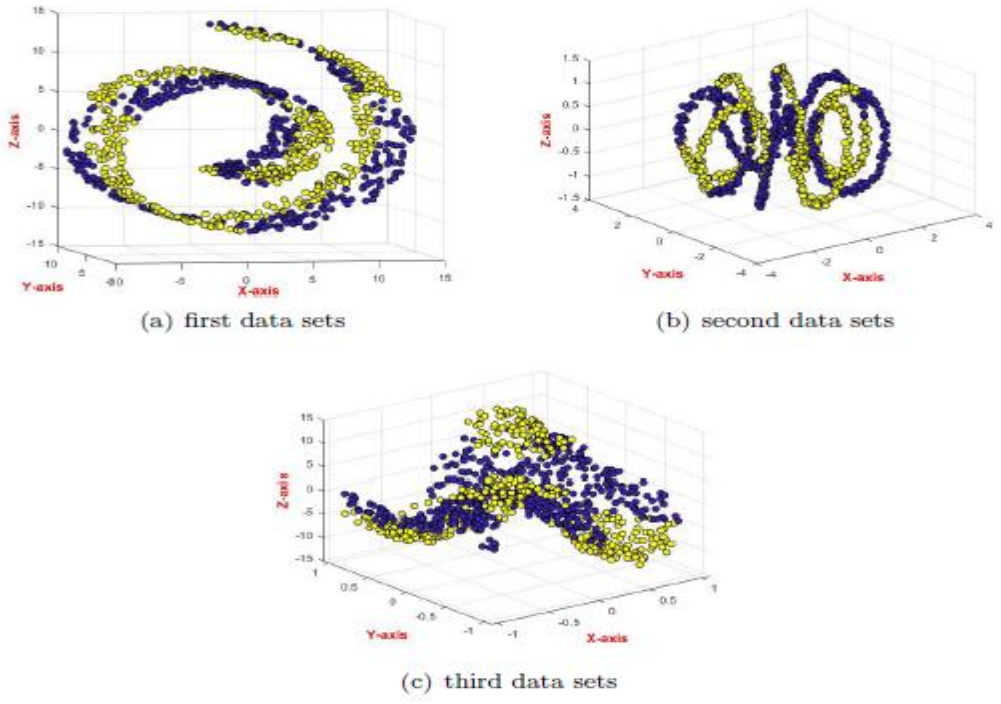
For the data sets under processing are formed in the multidimensions space, this representation is used to illustrate the dimensions (layers) in which the original data are formed. The processing will be focused of the layers rather than original data sets. Each layer is evaluated based on its information, and how closely the data in one layer relate to other. While each area in the single layer has an evaluation, the less evaluated areas indicate the weakness of the relationship between this area and the rest of the layers.

The method selects the best layers that have the highest evaluations, where it is used to determine the structure of the original data. The selection is done because the rest of the layers contain additional useful information when trying to visualize the original data, but they are excluded to focus deeply on the basic information. In the last step, the low dimension space is used for the purpose of checking linear or non-linear specifications. The preservation of neighborhood relations as in the original data sets indicates the data sets require to the linear structure, otherwise nonlinear structure is a results.

The method can deal with large volume data sets, where it represents them in a simpler structure without change. It extracts most important information that helps to achieve the goal. Two spaces of the same data sets are used, where the first one is the original data sets has n dimensions, and the second space is to represent the data sets in 2 dimensions but after excluding n−2 dimensions. The relation between data in both spaces is examined by calculating the points neighbours.

The process is done by selecting a point from the original data sets and finding its nearest neighbouring points. Then, the corresponding point in 2D space is chosen to find its nearest points. The relationship between the neighbourhood in the original space and 2D space is necessary, where preserving that relation means that data sets have linear structure. Otherwise, they have nonlinear structure. The processing can be summarized as the following step:

1. Let suppose X is a data sets in the high-dimensional space has n dimensions.
2. X is defined as set of layers X = {X1, X2, ..., Xn}.
3. Xi is a collection of points describes one particular layer Xi = (xi1, xi2, ..., xim),
where i = 1, 2, ...n.
4. Each Xi is evaluated based on their relation with other layers.
5. Xp and Xq are selected, which are the highest evaluation layers among all Xi, i = 1, 2...n. They construct the 2 dimensional space.
6. Preservation of the neighbourhood relation, between 2 dimensional space and X space, is a measure of data structure (linearity or nonlinearity) of the X.



(a) first data sets (b) second data sets

(c) third data sets

Fig. 2: *The three experimental data sets. The colours have been added to describe the relationship between points through processing.*

3 Experimental Results

3.1 Data description

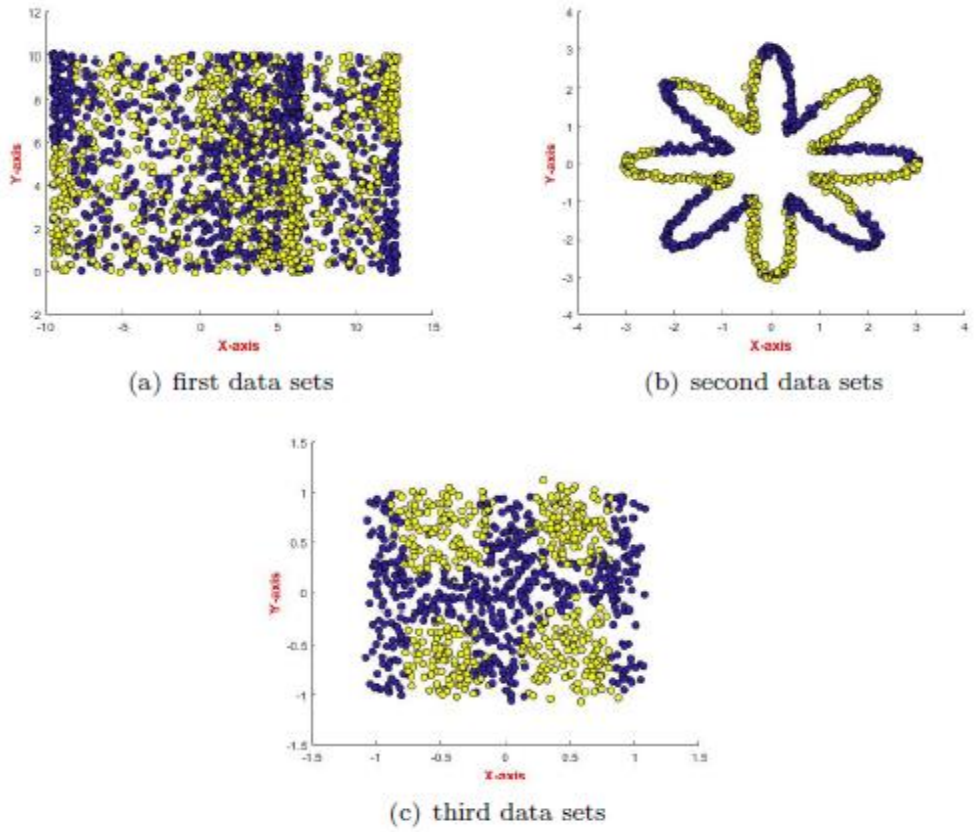
There are four experimental data sets are used; the first three data sets have 1000 points and the fourth data sets have 48800 points. The structure of neighbourhood relationships among points are explained by using colours.

1. First data sets: This data are generated as swiss-roll data sets. Figure 2(a) shows the data with their relation. As it can be seen, the structure of this set is complex, where specifying their data structure is not clear.

2. Second data sets: The topology of these data is like a wire around a cylinder, as in Figure 2(b). The data structure cannot be easily explored. There- fore, they need to simplify in order to understand the data fusion of these data.

3. Third data sets: These data sets are demonstrated in the figure 2(c). According to the topology of this data sets, they are difficult to interpret the correlations among data points.

4. Fourth data sets: These data sets are shown figure 7(a), where it shows the real colour image.



(a) first data sets (b) second data sets

(c) third data sets

Fig. 3: *The two dimensional space of the three experimental data sets.*

3.2 Results and discussion

The suggested method is applied on the four data sets in Matlab14 environment to solve their problem of the data fusion.

First data sets: Figure 3(a) shows 2 dimensional space of the original data sets, where the visualization shows neighbourhood relations between points are lost. To prove that, Figure 4 is used different neighbourhood size when using 6, 8, 10 and 12. The highest percentage is 59.7% shown on 100 random points at 12 neighbours. While the lowest percentage 47.7% shown on all 1000 points at 12 neighbours. At 100 random points the percentage average of preserving 6 nearest points is 54.5%, and the average slightly goes up at 8, 10 number of neighbour to reach 59.7% at 12 neighbour. When the random number of the selected points is increased to 500, the percentage average of preserving neighbours is less than in 100 points. Starting from 50.3% at 6 points and ending in 47.7% at 12 points. Finally, the whole 1000 point experiment shows almost as the same results as 500 random points. The percentage average ranged from 49% at 6 neighbours to 45.7% at 12 neighbour. According to this results, the percentages of preserving neighbours lead to conclude this data has nonlinear structure.

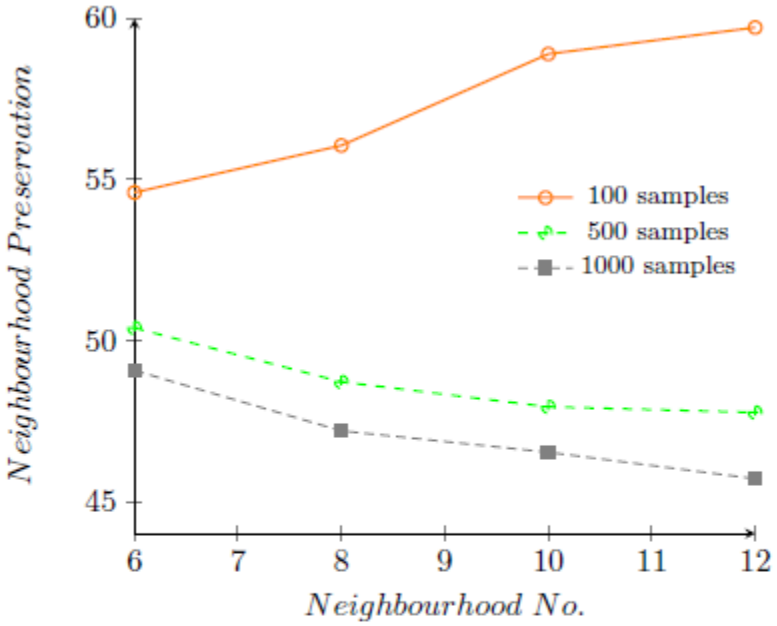


Fig. 4: *The result of applying proposed method to the first data set. The highest percentage is 59.7 shown on 100 random points at 12 neighbours. While the lowest percentage 47.7 shown on 1000 points at 12 neighbours. The results indicates this data sets have nonlinear characteristic.*

Second data sets: The results of second data sets after applying proposed method is demonstrated in Figure 3(b). The neighbourhood relation between points is preserved as in the original data sets. Figure 5 illustrates how neighbours have been preserved in each selected data points. The highest percentage is 85.5% shown on 500 random points at 8 neighbours. While the lowest percentage 65.04% shown on 100 random point at 12 neighbours. The percentages of preserving neighbours are more than 65% in all selected points at any number of neighbours. The 500 random points has the most significant percentage of neighbours preserving, which is 85.002%. This percentage could be obviously seen on the top of 8 neighbours. The percentages are still high on the

same selected data set points 83.90%, 85.0%, 84.32% at 12, 10 and 6 respectively. On the other hand, the 100 selected data set points has the lowest percentages, where they ranged from 65.04 at 12 neighbour to 67.87% at 6 neighbour. When using all 1000 points, the preserving the neighbourhood relation has good results which is 82.8%. The results of the second data sets prove that these data have linear structure. The percentages of preserving neighbours are very high and adequate to explore the linearity relationship between these data sets.

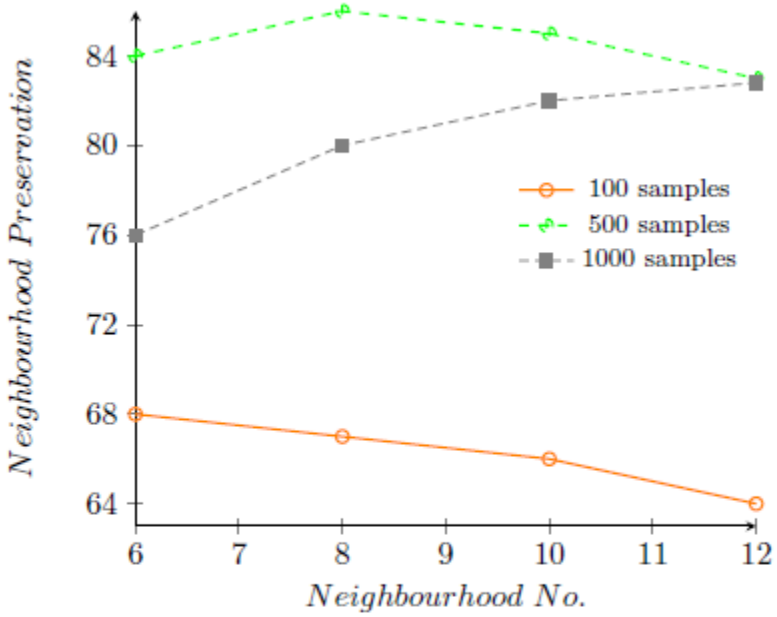


Fig. 5: *The results of applying proposed method to the second data set. The highest percentage is 85.5 shown on 500 point at 8 neighbours. While the lowest percentage 65.04 shown on 100 point at 12 neighbours.*

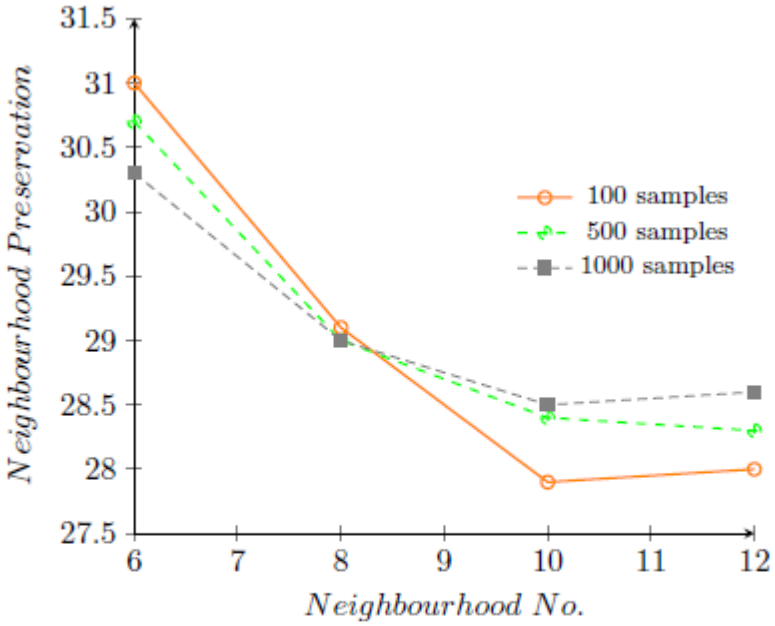


Fig. 6: *The results of applying proposed method on the third data set. The percentages of this data set are low. They do not exceed 31.5%.*

Third data sets: The visualization on the Figure 3(c) shows the third data sets lost their neighbourhood relation among points. Figure 6 shows the number of the selected points and the number of neighbours. The curves on the figure represents the percentage average of preserving neighbours. The preserving of the neighbours in the third data sets is lost, where the best percentage is 31.003% which is registered in 100 random points at 6 neighbours. However, this number of the random points has the worst preserving neighbours percentages which is 27.92% at 10 neighbours. Changing in the number of the random points and the number of neighbours did not make a more difference; where preserving neighbours keeps weak and percentages still low. On selecting 500 random points or all 1000 points, the percentages are approximately identical. They are ranged from 30% at 6 neighbours to 28% at 12 neighbours.

The experimental results conclude the weakness of neighbours preserving is obvious with low percentages in all numbers of selected points at all numbers of neighbours. None of the percentages is exceeded 31.5%. These results make the third data sets has definitely nonlinear structure.

3.3 Real world data sets

The dealing with real data sets as colour image is a good way to prove the efficiency of the suggested method. The goal of processing is to determine the linearity or non-linearity structure of the current image. While the colour image is a data sets have multi-dimensions, the processing requires to convert this high dimensional data into a simplified form.

The colour image is divided into set of clusters, where each cluster has three dimensions:

$$Image = \{C_1, C_2, ..., C_k\} \quad \ldots\ldots (1)$$

$$C_i = \{C_{ir}, C_{ig}, C_{ib}\},$$

where /C_i/ = 3D.

Each cluster C_i, processed separately through deleting one dimension.

$$C'_i = C_i - \{C_{ij}\} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

Where /C'_i/ = 2D

The specifying the data fusion of each cluster is the process to be applied in this step. Measuring the amount of neighbourhood relation for each colour is the way by which the cluster type is assigned. Then, the average of the efficiency measure for all image; clusters is the real value, that represents the type of the original image data sets being processed.

$$Image\ type = \frac{\sum_{i=k}^{n} M(C_i)}{K} \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Processing: According to the equation 3, the amount of preserving neigh- bours in each selected numbers of clusters could be measured by M(Ci)/K. The figure 7, illustrates the different amount of neighbourhood relationship in each number of clusters. The weakest amount of preserving neighbours is 30.856% registered when k = 3, while the strongest one is registered when k = 7 which is 34.83%. This indicates to inadequate amount of preserving neighbours, and all other tested number of clusters have a less percentage of preserving neighbours. The weaker neighbourhood relationship in all clusters demonstrate a low preserving neighbours. According to these results, it is easily to conclude the colour image has a nonlinear structure.

(a) Real colour image (b) Classifying into three cluster

Fig. 7: a)*The real taken image. The image is in the real world colours, where its size is* 200x244. b)*The real taken image after classifying to three clusters. The image is in three colours: white, black and gray.*

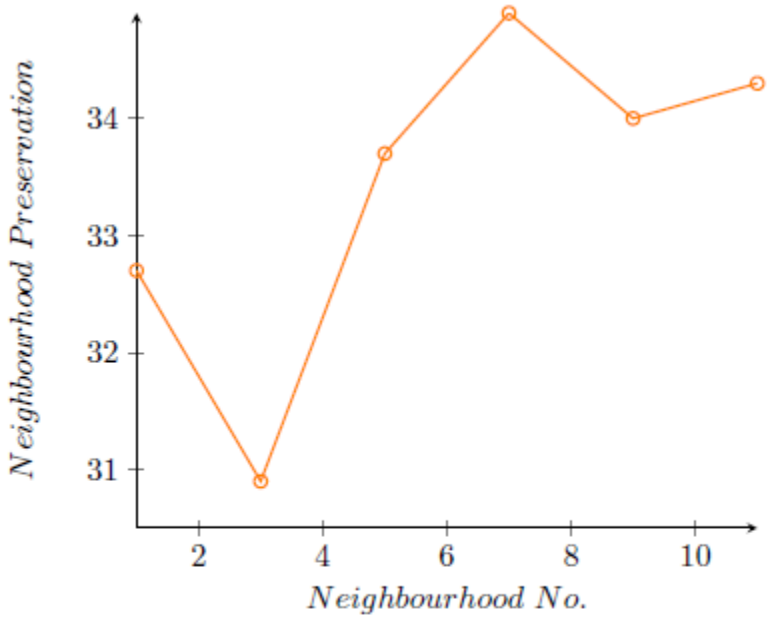


Fig. 8: *The results of applying proposed method on the third data sets. The percentages of this data set are low. They do not exceed 31.5%.*

4 Conclusion

Data visualization of scientific data has been used in this paper to overcome the problem of the data fusion. The proposed method focused on the basic information contained in the data sets for the purpose of facilitating the analysis process. The method relied heavily on extracting relationships between data, which had a major role in getting right conclusions. The process of extracting the appropriate data structure has been characterized by being automated to facilitate the task and focus on the direction of proper processing.

References

[1] Ahmed, B., Najim, S.A., Mansour, W.A.: Role of software engineering in visualizing large volume of hyberspectral and medical data sets. International Journal of Computer Application 176(9), 1–5 (2017)
[2] Keahey, T.: Using visualization to understand big data. IBM Corporation (2013)
[3] LILJA, D.J.: Linear Regression using R an Introduction to Data. University of Minnesota Libraries Publishing Minneapolis, Minnesota, USA (2016)
[4] Liu, S., Maljovec, D., Wang, B., Bremer, P.T., Pascucci, V.: Visualizing high-dimensional data: Advances in the past decade. IEEE Transactions on Visualization and Computer Graphics 23, 1249 – 1268 (2017)
[5] Najim, S.A.: Information visualization by dimensionality reduction: A review. Journal of Advanced Computer Science and Technology 3(2), 101–112 (2014)
[6] Najim, S.A., Lim, I.S.: Trustworthy dimension reduction for visualization different data sets. Information Science 278, 206–220 (2014)

[7] Najim, S.A., Lim, I.S., Wittek, P., Jones, M.: FSPE: Visualisation of hyperspectral imagery using faithful stochastic proximity embedding. IEEE Geoscience and Remote Science Letters 12(1), 18–22 (2015).
[8] Pek, J., Sterba, S.K., Kok, B.E., Bauer, D.J.: Estimating and visualizing nonlinear relations among latent variables: A semiparametric approach. Multivariate Behavioral Research 44(4), 407436 (2009).
[9] Pena, M., Barbakh, W., Fyfe, C.: Topology-preserving mappings for data visualisation. Lecture Notes in Computational Science and Enginee 58, 132–152 (2008).
[10] Silva, R., Rauber, P.E., Telea, A.C.: Beyond the third dimension: Visualizing high_dimensional data with projections. Computing in Science and Engineering 18(5), 98–107 (2016).
[11] Tingting Mu, J.Y.G., Ananiadou, S.: Data visualization with structural control of global cohort and local data neighborhoods. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(6), 1323–1337 (2018).
[12] Yan, W., Pablo, T., Kun, Z.: Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. bioRxiv (2018). DOI 10.1101/276261. URL https://www.biorxiv.org/content/early/2018/06/22/276261.