Star Coordinate Dimension Arrangement using Euclidean Distance and Pearson Correlation

Noor Elaiza Abdul Khalid, Izyan Izzati Kamsani
Universiti Teknologi Mara, 40450, Shah Alam, Selangor Darul Ehsan, +60355442000, Malaysia

Article Info

Article history:

Received May 18, 2018
Revised Jun 19, 2018
Accepted Jul 3, 2018

Keywords:

Data dimension arrangement
Euclidean distance
Pearson correlation
Star coordinate
Visualization

ABSTRACT

Star Coordinate (SC) is a circular visualization technique that maps k-dimensional data. Its interactive features allow user to manipulate projections to search for hidden information. Without prior knowledge of relationship between dimensions users will be blindly searching for clusters. This paper proposes dimension rearrangement using Euclidean Distance and Pearson Correlations to reveal the clusters in SC. The methodology consists of four phases; Calculate the distance between individual attributes against a dependent attribute using Euclidean distance; Pearson correlation is used to obtain the correlation data attributes; Sort the correlation values in ascending order; finally, attributes table are reordered with the positive values to the right and negative values to the left according to the correlation value. The resulting tables are applied to produce the SC. This method is successful in producing clusters that makes it easier for the users to further manipulate the SC for their data analysis.

Corresponding Author:

Noor Elaiza Abdul Khalid,
Universiti Teknologi Mara,
40450, Shah Alam, Selangor Darul Ehsan, +60355442000, Malaysia.
Email: elaiza@tmsk.uitm.edu.my

1. INTRODUCTION

Recent advancement of high performance technologies has resulted in collections of large high dimensional data. This consists of data with large number of records and attributes. Extracting meaningful information from raw data could be a difficult task. One way to understand high dimensional data is to display it in a low-dimensional plane [1].

The main motivation for domain experts in analyzing their multidimensional data is to detect and interpret cluster separation and outliers [2]. Prior to that, we need to study and analyze high dimensional data to understand and interpret the relationship between cluster and data attributes. Few (2007) stated that, handling growing high dimensional data causes difficulties especially in clustering elements into groups including visualization problems due to data clutters or distracting results. Proper clustering is a useful technique for statistical data analysis [4]. It is a process of grouping data based on the similarity of their properties.

High dimensional data can be displayed in a clustered result through visualization approaches in which there are many techniques that can be used. One of it is the Star Coordinate (SC) technique. SC technique is able to reveal patterns and groups from high dimensional data while still showing the impact of data attributes in the formation of its patterns and groups [5]. SC technique can also reveal the clusters through manipulation of the axes by trial-and-error. Thus, the critical question here is which feature or dimensions best separates the classes and allow cluster-based data classification. The main problem is without prior knowledge finding the right ones is trivial [2].

Advances of high performance technologies in the field of medicine, engineering, science and business has resulted in the production of huge amounts of data, which is known as high dimensional data.

This consists of information related to multiple data records, attributes and the relationship between them. High dimensional data is difficult to interpret and understand. One way to get meaningful information from high dimensional data is through visualization technique. Visualization is a technique which transforms raw data into a graphical form. SC technique is one of the familiar techniques used to visualize high dimensional data. The good of using SC technique is, data distribution and data dimension (attributes) is plotted in a single window. Thus, it eases users to do a comparison or understanding the pattern in visualization. Compared to other technique, it needs a more space to place the results of visualization. High dimensional data consist of information in many tables that are related to multiple data records and attributes. Furthermore, there is a relationship between records and attributes. Data records represent the data point while data dimensions (attributes) involved with the axis position in SC environment. The arrangement of data dimension in SC is vital since it affects the appearance of cluster in future. Initially, as a laymen user, they will apply a trial-and-error method to produce visualization. They will arrange the data dimension randomly which are extracted from a data table. However, random arrangement does not reveal good clustering.

The main motivation in exploring clustering analysis is to determine whether the arrangement of data dimensions influences the appearance of clustering, and why it is important to know the correct positions of data attributes in each axis. There are two reasons why this is important. The first is that users with little knowledge of SC techniques may lack guidance in exploring the data, having to resort to trial-and-error, and would subsequently feel discouraged Feng et al. (2018). Second, users need to know the importance of attribute points in the data, and why they are arranged that way. Knowing this beforehand would enable them to form a good summary and make a faster decision. SC is limited when it comes to a high number of data dimensions and will clutter the data formation. Third, it is important to eliminate the irrelevant data dimensions from being displayed [5] that would not affect the formation of data and avoid cluster. From this, a set of questions arise which motivates our work: How to motivate layman users without pre-knowledge in SC on the first step in clustering analysis? Where is the correct, proper placement of data dimensions? Why it is important to know the right position of data attributes? Which data dimensions is irrelevant to be display in an SC layout? In the next paragraph, related works to this study is reviewed.

Previous literature related to dimension arrangement in SC will be discussed in detail. As introduced by Kandogan, Road, & Jose (2000), SC is a simple and efficient technique for visualizing multidimensional data. SC works by presenting data points using vector sum of attributes values along the axis. In this paper, they provided the users with the ability to view clusters, trends and outliers in the distribution of data. Cluster analysis is often one of the first steps in the analysis of data. However, there are some weaknesses in exposing the clusters pattern, the largest of which is the arrangement of dimensions.

There are several studies which focus on dimension arrangement. The first paper is found in year 1998 by Ankerst, Berchtold, & Keim. They stated that the order and arrangement of dimensions plays a significant role in presenting many high-quality visualization techniques such as parallel coordinate, scatterplot and more. In their paper, dimension arrangement issue has been shown to be an N-P problem and they suggested using heuristic algorithm to determine the similarity of each data dimension. Data dimension with similar behaviors are placed next to each other. Yang, Peng, Ward, & Rundensteiner (2003) proposed an interactive hierarchical ordering of the dimensions based on their similarities, thus improving the manageability of high-dimensional datasets and reducing the complexity of the ordering. Ward & Rundensteiner (2004) applied the concept of clutter-based dimension ordering in various visualization techniques to reduce the visual clutter. Then, Sun, Tang, Tang, & Xiao (2008) came out with their idea on designing dimension configuration strategy 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 relation between the arrangement of dimensions and the quality of visualization using the Radviz technique. Garcia et al. (2016) proposed an interactive Star Coordinate (iStar) which can handle a large number of data dimensions. They also studied how the order of data dimension can have an impact on revealing pattern and clustering, enabling users to understand them easier. Wang et al. (2017) studied about determining which dimensions are relevant or irrelevant to be displayed in the SC layout which contributes to clustering.

This paper presents the study of dimension arrangement in SC environment to reveal the clusters using *Pearson Correlation* technique with basic knowledge of SC technique. The proposed method which not only improves the efficiency of axes manipulation with higher cluster quality, but also enables users to learn the relations between clusters and data attributes. Dimensions are arranged based on the sorted correlation value within the same length of axis and angle. The correlation values which show similar behavior are placed next to each other and would benefit first-time users using SC technique by serving as guidance when observing the cluster’s appearance. Firstly, the distance of each data attributes are calculated. Then the correlations between data attributes were determined. The correlation values from small to large value were sorted. The correlation values would produce negative and positive values. When plotting the data in SC, the negative correlation values for each data attributes will be positioned on the left side while positive value would be placed on the right side.

The remaining part of this paper is organized in the following manner: Section 2 describes the methodology used. Result and discussion will be explained in section 3. Lastly, the conclusion will be discussed in section 4.

2. RESEARCH METHOD

In this section, all the processes and experiments that have been done will be discussed in detail. SC layout consists of circularly arranged vectors vi with a common origin, each vector corresponding to a data attribute. Data instances are mapped to the layout as a linear combination of the vectors vi. To improve the user experience, SC methods enable interactive features that allow users to rotate the angle of the vectors vi to find configurations where patterns and groups are more clearly revealed.

2.1. Data Collection

To test the method, an automobile dataset will be used in the experiment. This dataset is a benchmark data that is widely used by various researchers and consist of 395 automobiles from the 1970's until 1980's. The attributes measured here are fuel efficiency-miles per gallon (MPG), name of the cars (name), origin of the car (origin), year of the car (year), acceleration, weight, horsepower, engine displacement (displacement) and number of cylinders (cylinders).

2.2. Process and procedures

The figure below shows the process on arranging the data dimension in right position. There are six steps in this process. Details are discussed below:

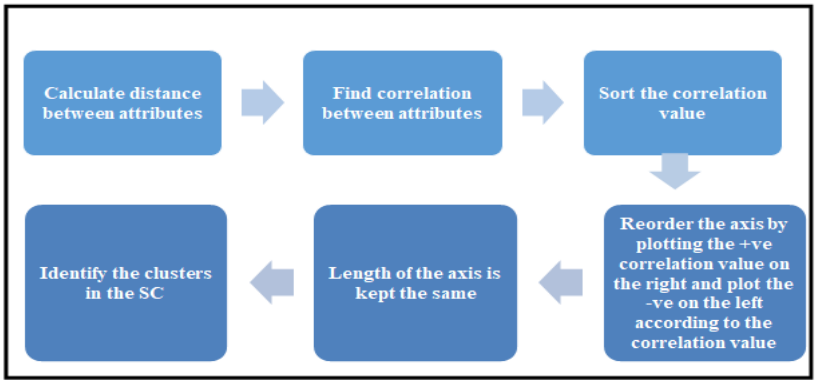

Calculate distance between attributes
Find correlation between attributes
Sort the correlation value
Reorder the axis by plotting the +ve correlation value on the right and plot the -ve on the left according to the correlation value
Length of the axis is kept the same
Identify the clusters in the SC

Figure 1. The process of dimension arrangement in SC environment

Step 1: Calculate distance. Method used in calculating distance is Euclidean distance.

Eq. $d_{i,j} = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

To determine their similarity, a Euclidean distance measure is used. Results are used to determine the similarity arrangement of dimensions as in Step 2.

Step 2: Pearson correlation, r.

Eq. $r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$

In order to determine the correlation between axis, the Pearson formula is used. This step is crucial to ensure the position of each axis is correct so that the appearance of each cluster can be identified.

Step 3: Sort the correlation values

In this step, each axis will be sorted from the highest to the lowest value. The sorted axis can be referred as in the Table 1. However, not all axes will be selected, if the correlation value is out of range [-0.5, 0.5]. The attribute of Origin and Acceleration are not chosen to be displayed in the visualization since it doesn’t meet the range.

Table 1. An example of sorted attributes with correlation value using MPG attribute as an anchor

Anchor: MPG	Correlation Value	Chosen Axes
MPG with Year	0.580384	√
MPG with Origin	0.479549	×
MPG with Acceleration	0.420574	×
MPG with Cylinder	-0.77714	√
MPG with Horsepower	-0.77843	√
MPG with Displacement	-0.80525	√
MPG with Weight	-0.83228	√

Step 4: Plot the data dimension (attributes) with same angle between each axis

1. Negative on the left.
2. Positive on the right.

As can be seen, vector additions within the SC space must be valid, in order to project all the data points correctly on an SC. It is a clockwise angle between axes. The positive correlation value is placed on the right side, while negative correlation value is placed on the left side. The outline of axis positioned is shown below in Figure 2.

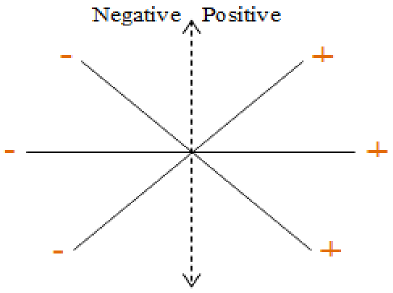


Figure 2. The outline of axis position based on obtains correlation value

Based on Table 1, the first axis to be positioned on the right side after an anchor, MPG with the positive correlation value is Year, followed by negative correlation value which is Cylinder, Horsepower, Displacement and Weight.

Step 5: Identify clusters

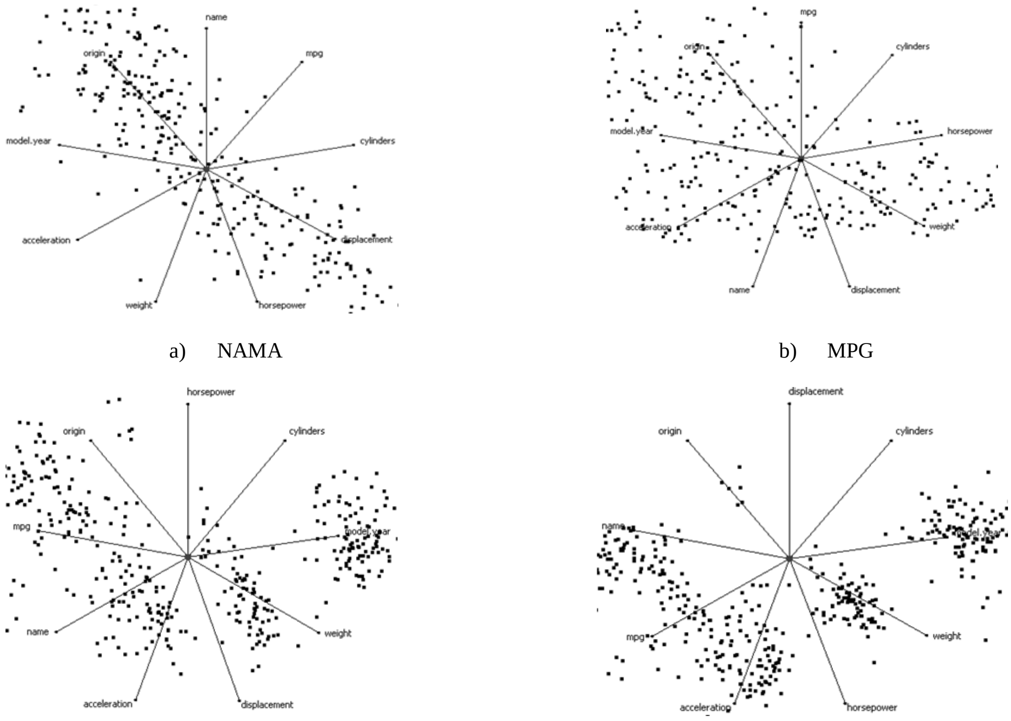
After plotting the axis, an early cluster appearance is formed. This cluster can be an initial guideline to the laymen to do more data exploration.

3. RESULTS AND ANALYSIS

Before implementing proposed method, we came out with random plotting. This is to show whether plotting randomly can produce cluster results or not. Results are shown in Figure 3.

Figure 3 shows randomly plotted data dimensions. As illustrated from this figure, we are unable to both extract any useful information and define the relationship between data dimensions. The result only shows the distribution of data that represents high dimensional data visualization in (a). In (b) it shows that mpg is plotted as an anchor to examine the compromise between other data attributes. However, the positions of data attributes are placed regardless of interrelatedness between attribute data. Cylinder, horsepower, and weight are positioned on the right side, while origin, year and acceleration are located to the opposite. Based on (b), the data is plotted in a trial-and-error method to get useful information. Unfortunately, without knowing the correlation between data attributes it is meaningless. Figure in (c) shows the appearance of clusters, but without knowing the reason why it is positioned in that way. Here, horsepower data are plotted as an anchor to see the effect on other data attributes. Lastly in (d) we are able to reveal the cluster, with minimal previous knowledge on the importance in data dimension arrangement, but through trial-and-error process.

Step 1: <u>Random Plotting</u>



a) NAMA

b) MPG

c) HORSEPOWER

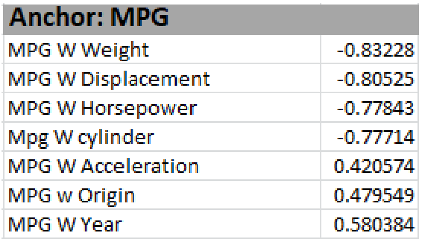
d) DISPLACEMENT

Figure 3. Random plotting with different anchors along with all data dimensions using the same angles in SC

All of the above will take more time to get an accurate result. Therefore to avoid this obstacle, we applied the proposed method.

Step2: <u>Remove value in range [-0.5, 0.5]</u>

This experiment will be tested on all the data attributes that act as an anchor. Prior to that, data attributes with a correlation value between -0.5 and 0.5 will be abolished. This is because, the data attributes within that range will be considered as less correlated with the other data attributes as it does not help in performing clustering pattern. Figure 4 shows an example that takes MPG as an anchor and removes acceleration and origin data dimension.



Anchor: MPG	
MPG W Weight	-0.83228
MPG W Displacement	-0.80525
MPG W Horsepower	-0.77843
Mpg W cylinder	-0.77714
MPG W Acceleration	0.420574
MPG w Origin	0.479549
MPG W Year	0.580384

Figure 4. Acceleration and Origin data dimension will be removed as it is in range [-0.5, 0.5]

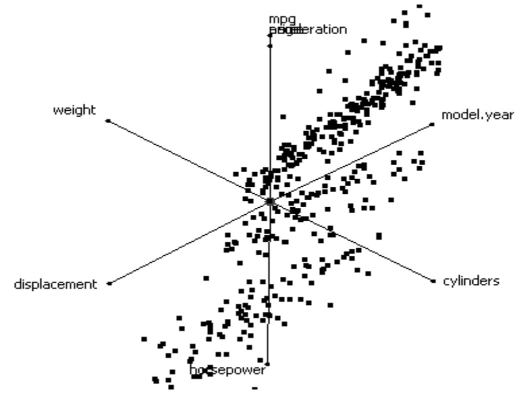
Step 3: <u>Plot the selected data attributes</u>

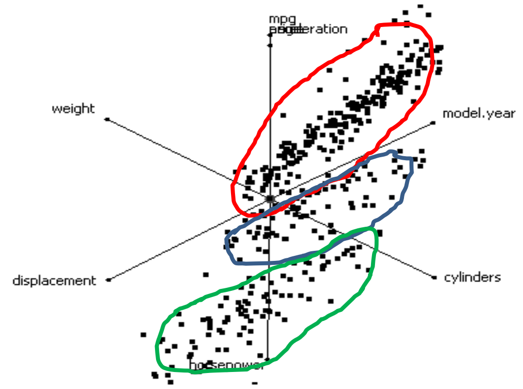
Then, selected data attributes that have negative value will be positioned on the left side, while positive value will be positioned on the right side. Each data attributes are presented in a different axis, divided within the same angle and arranged in clockwise. The same experiments goes to different anchor; cylinder, displacement and year. These results are illustrated in Table 2.

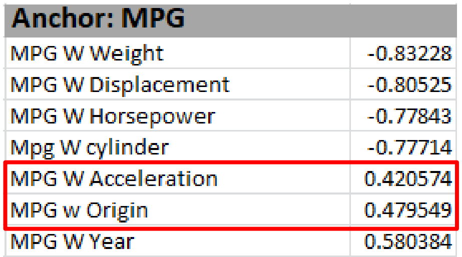
Table 2. Results for different anchors

SC | **Description**

Anchor: MPG







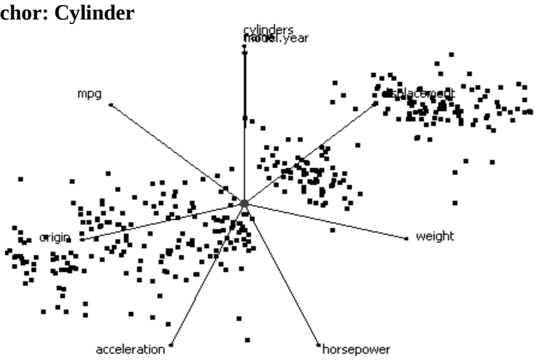
Anchor: MPG	
MPG W Weight	-0.83228
MPG W Displacement	-0.80525
MPG W Horsepower	-0.77843
Mpg W cylinder	-0.77714
MPG W Acceleration	0.420574
MPG w Origin	0.479549
MPG W Year	0.580384

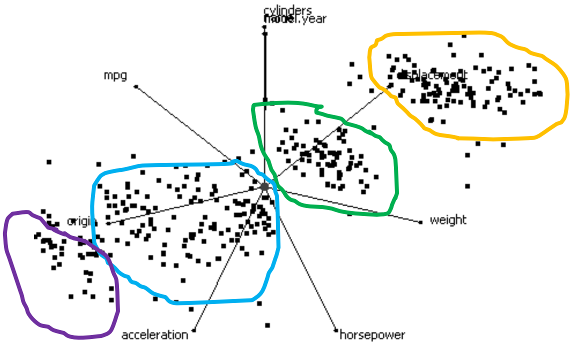
In this experiment, all correlation that is > -0.5 and < 0.5 is eliminated. Which means acceleration and origin has the correlation less than 0.5. Indicating low correlation.

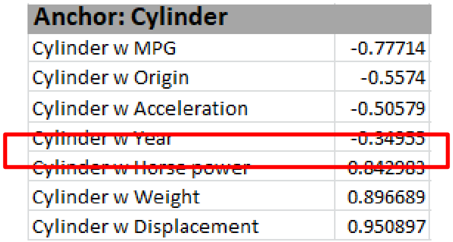
The axis that are selected are year, cylinder, horsepower, displacement and weight. The selected axes are placed in sorted order and located evenly throughout the 360 angle.

The resulting image is a vague formation of 3 clusters.

Anchor: Cylinder







Anchor: Cylinder	
Cylinder w MPG	-0.77714
Cylinder w Origin	-0.5574
Cylinder w Acceleration	-0.50579
Cylinder w Year	-0.34955
Cylinder w Horse power	0.842983
Cylinder w Weight	0.896689
Cylinder w Displacement	0.950897

As the above correlation value, it shows that Year will be eliminated in the display because its value is greater than - 0.5.

The selected axes are displacement, weight, horsepower, acceleration, origin and MPG. The length of the axes will be kept the same and they are evenly located throughout the 360 angle.

Attributes which has positive value will be placed on the right side while negative values will be placed on the left side. However, which axis should come first is determined by its correlation value.

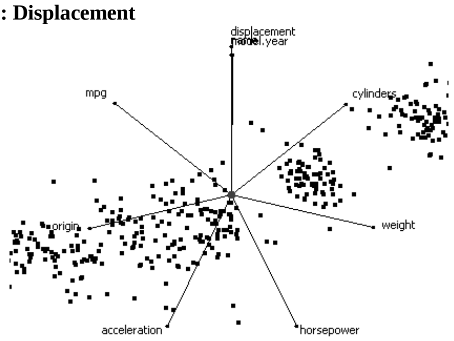
Attribute with higher correlation value like displacement will be located after its anchor, cylinder. Next, lower correlation value such as weight will be placed next to displacement followed by horsepower attribute.

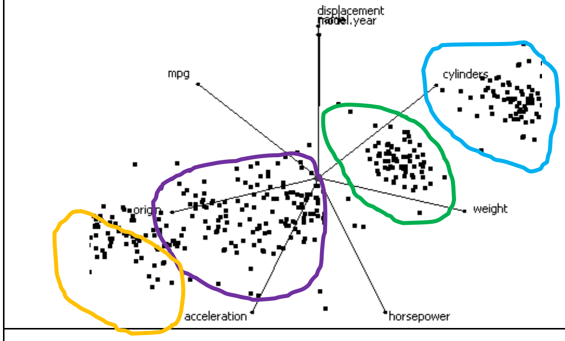
While for negative correlation value, the smaller value like MPG will be placed next to its anchor, cylinder.

All of the dimension position is arranged clockwise.
The resulting image shows formation of 4 clusters.

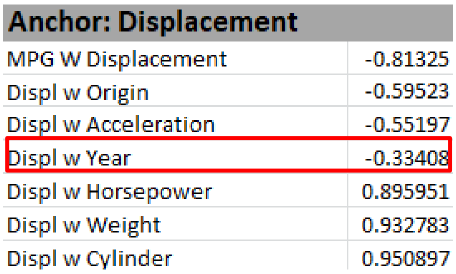
<table>
<tr><th>SC</th><th>Description</th></tr>
<tr><td>

Anchor: Displacement





</td><td>



<table>
<tr><th colspan="2">Anchor: Displacement</th></tr>
<tr><td>MPG W Displacement</td><td>-0.81325</td></tr>
<tr><td>Displ w Origin</td><td>-0.59523</td></tr>
<tr><td>Displ w Acceleration</td><td>-0.55197</td></tr>
<tr><td>Displ w Year</td><td>-0.33408</td></tr>
<tr><td>Displ w Horsepower</td><td>0.895951</td></tr>
<tr><td>Displ w Weight</td><td>0.932783</td></tr>
<tr><td>Displ w Cylinder</td><td>0.950897</td></tr>
</table>

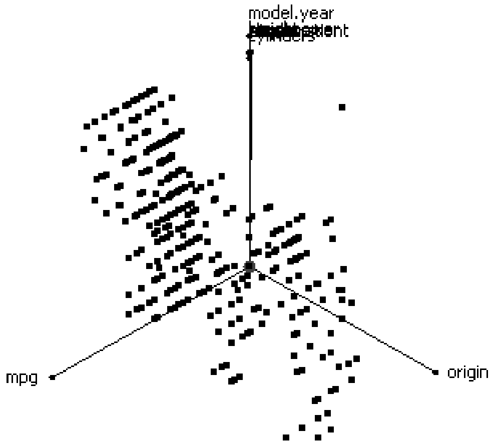
Attribute Year is eliminated from the display since it shows correlation value greater than -0.5. The positions of attributes are same as the previous anchor; cylinder except for the displacement which swaps it places with the cylinder.

The display of this figure shows the similarity data pattern with anchor cylinder. Though displacement and cylinder are swapped the pattern of data shows the similarity as in the previous figure. It indicates that the correlation between displacement and cylinder has higher correlation compared to other attributes.

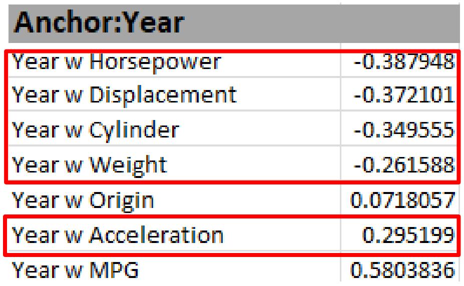
The resulting image shows formation of 4 clusters.

</td></tr>
<tr><td>

Anchor: Year



</td><td>



<table>
<tr><th colspan="2">Anchor:Year</th></tr>
<tr><td>Year w Horsepower</td><td>-0.387948</td></tr>
<tr><td>Year w Displacement</td><td>-0.372101</td></tr>
<tr><td>Year w Cylinder</td><td>-0.349555</td></tr>
<tr><td>Year w Weight</td><td>-0.261588</td></tr>
<tr><td>Year w Origin</td><td>0.0718057</td></tr>
<tr><td>Year w Acceleration</td><td>0.295199</td></tr>
<tr><td>Year w MPG</td><td>0.5803836</td></tr>
</table>

The anchor in this figure is year. All of the attributes show that their correlation value is greater than -0.5 and less than 0.5 except origin and MPG. This means that origin and MPG are not eliminated in the display.

Based on the figure, there is no strong correlation among horsepower, displacement, cylinder, weight and acceleration. Thus, no cluster is formed in this presentation.

Thus, year attribute should be eliminated during the display since there is no obvious correlation between origin and mpg.

</td></tr>
</table>

4. CONCLUSION

Determining the correlation between data dimension is vital in performing clustering analysis. Using the Pearson method makes it easy to know the correct positioning of data dimensions in an SC. Consequently, giving laymen with this information helps them in their first step towards exploring analysis of these clusters, which thus saves them from doing a time-consuming trial-and-error process. Thus, the proposed technique could assist users to reveal the cluster relationship between data dimension through dimension correlation.

REFERENCES

[1] Sanchez A, Lehmann DJ. *Adaptable Radial Axes Plots for Improved Multivariate Data Visualization*. *Comput Graph Forum*. 2017;36(3):389–99.

[2] Wang Y, Li J, Nie F, Theisel H, Gong M, Lehmann DJ. *Linear Discriminative Star Coordinates for Exploring Class and Cluster Separation of High Dimensional Data*. *Comput Graph Forum*. 2017;36(3):401–10.

[3] Few S. *Data Visualization Past, Present and Future*. Percept edge. 2007.

[4] Mor M. A *Review on Various Clustering Techniques in Data Mining*. *Int J Comput Sci Commun Networks*. 2016;6(July):138–42.

[5] Garcia G, Gustavo L, Gomez-nieto E. *iStar (i*): An interactive star coordinates approach for high-dimensional data exploration*. Comput Graph [Internet]. *Elsevier*; 2016;60:107–18. Available from: http://dx.doi.org/10.1016/j.cag.2016.08.007

[6] Feng K, Wang Y, Zhao Y, Fu C, Cheng Z, Chen B. *Cluster aware Star Coordinates*. J Vis Lang Comput [Internet]. *Elsevier* Ltd; 2017;44:28–38. Available from: https://doi.org/10.1016/j.jvlc.2017.11.003

[7] Kandogan E, Road H, Jose S. *Star Coordinates : A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions*. *Proc IEEE Inf Vis Symp*. 2000;(650):22.

[8] Ankerst M, Berchtold S, Keim D a. *Similarity clustering of dimensions for an enhanced visualization\nof multidimensional data*. *Proc IEEE Symp Inf Vis*. 1998;98:52–60.

[9] Yang J, Peng W, Ward MO, Rundensteiner EA. *Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets*. *Proceedings - IEEE Symposium on Information Visualization*, INFO VIS. 2003. 105-112 p.

[10] Ward MO, Rundensteiner E a. *Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering*. *IEEE Symp Inf Vis*. Ieee; 2004;89–96.

[11] Sun Y, Tang J, Tang D, Xiao W. *Advanced Star Coordinates*. *Proceedings - The 9th International Conference on Web-Age Information Management, WAIM 2008*. 2008. p. 165–70.

[12] Di Caro L, Frias-Martinez V, Frias-Martinez E. *Analyzing the role of dimension arrangement for data visualization in Radviz*. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2010. p. 125–32.