Data Mining

Seminar 02

Visualizing Multi-dimensional Clusters,Trends, and Outliers using Star Coordinates

Department of Computer Science
Kristianstad University
Course: DT584C
Master's in computer science

Student: Hazrat Ali <react.dev.se@gmail.com>
Teacher: Dawit Mengistu

Date: 15 January 2020

Table of contents

Introduction

This is a critique of the article “Visualization of multidimensional Trends and Outliers using Star Coordinates [1]” authored by Eser Kandogan in 2001 and published in “ Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - August 2001”. This article is concerned about visualization of multidimensional data in a simple form using a technique proposed called Start Coordinates.

The author belongs to the University of Maryland and has a total of 26/76 publications during 1996 to 2011. His work has been cited over 600/1200 times in various other research articles[2]/[3].

Visualization of in the fields of scientific engineering and business data plays an important role in knowledge discovery. Visualization is important in initial knowledge discovery as there is very little info in the start about the data.

The article has presented a very nice solution with user interactions to get insight into the data by scaling, rotating and marking etc.

Body

Visualization techniques helps in decision making activities. Since both of the scientific, engineering and business data is multi-dimensional but we live in three dimensional space and therefore it is challenging to visualize it.

To achieve this a new technique of star coordinates have been proposed by the author in which multidimensional data is first arranged on two dimensional surface. Where each point represents a set of encoded attributes.

On top of it various interactions have been added to stimulate visual thinking, e.g integrate and separate dimensions, analyze correlations of multiple dimensions, view clusters, trends, and outliers in the distribution of data.

According to author the start coordinates system is useful in analysis of multiple factors using hierarchical clustering.

The motivation of this task is to reduce the dimensionality curse and provide simple data representations. With simple representation it is still possible to grasp the distribution of data and clusters.

Star coordinates

In star coordinates each axis share the same origin point and each multi-dimensional data element is represented a point where each attribute contributes to its location through uniform encoding.

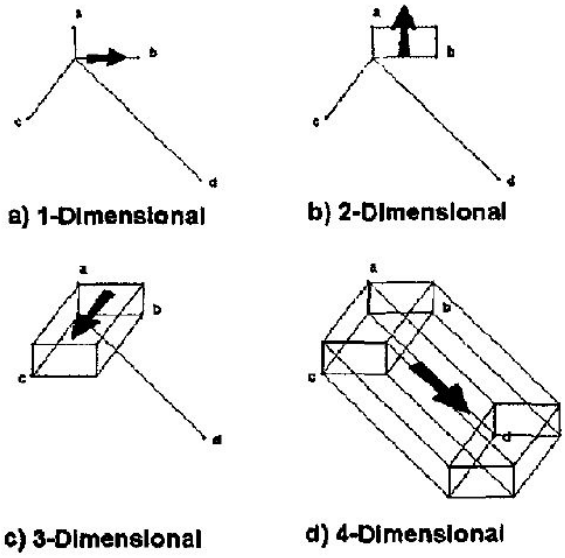

Figure 1 - n^{th} dimensional space is represented by a sweep of n-1 dimensional space

The n dimensions space can be represented by a sweep of n-1 dimensional space. In this system the coordinates are first arranged in two dimensional space. Minimum value is mapped at the origin and maximum at the other end of coordinate system.

Each point simply represents n-dimensional data, Start coordinates extend this idea to higher coordinates . Overlapping introduces ambiguities when data points are encoded (two same value). For example x = 2, y = 2 into one dimensional 4 but 1+3 is also 4. It is solved with the aid of interactive dynamic transformations such as rotations and translations.

Interaction techniques

The author suggest these techniques to the users for better understanding of their datasets. To make a better sense of data distribution the user can apply a set of transformations such as scaling, rotation, marking, range selection and footprints.

The author has given an example which uses a dataset of 400 auto specs (mpg, cylinders, weight, acceleration etc). Figure 2 below shows the initial histogram with settings (not visible)

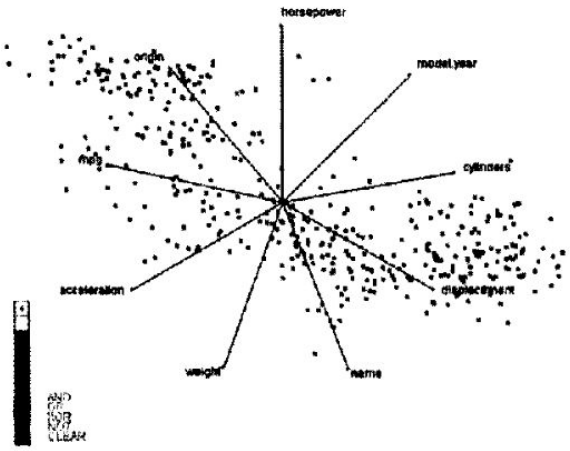

Figure 2 - Star coordinates - visualization of 400 auto specs

Scaling

This transformation enables users to change the length of an axis by increasing or decreasing the contributing attributes in the visualization. In Figure 3 below the author scales downs the name attribute to show (5) clusters independent of the car vendor name.

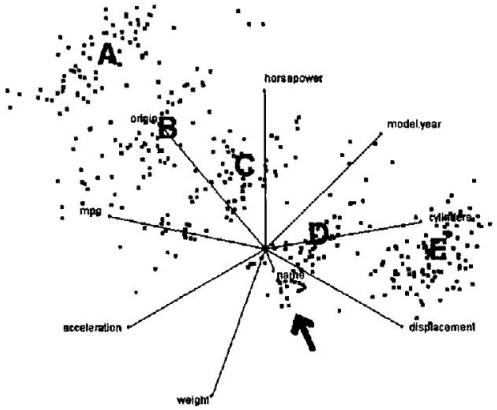

Figure 3 - Scaling down of name attribute from datasets reduces clusters to 5

Subsequent scaling down of attribute ‘origin’ (continents) from the datasets merges the clusters A, B and C as shown in Figure 4.

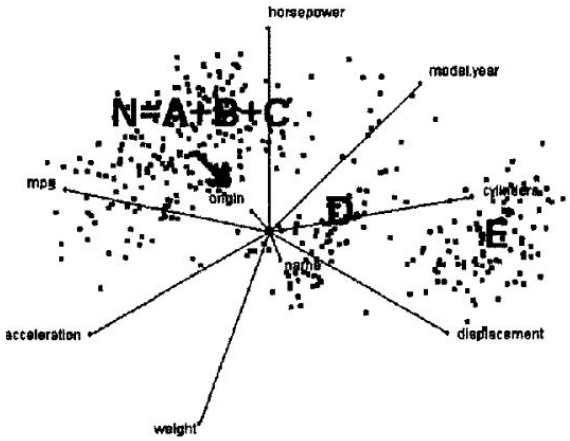

Figure 4 - Subsequent scaling down of origin attribute from datasets merges cluster A,B,C

Rotation

In this transformation the author wants to rotate the unit vector of axis. This makes a data attribute more or less correlated to the rest. Attributes of same interest are rotated in the same direction and vice-versa. Rotation helps in resolving ambiguities according the the author.

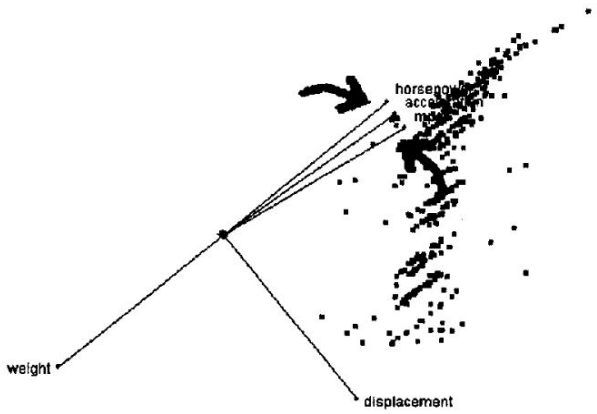

Figure 5 - Rotation restructures the data points based on desirable (e.g. 'horsepower', 'acceleration', and 'mpg')

Marking and Range selection

Here the author facilitates the user to select a set or range of points which would be visualized with different colours which helps in subsequent transformations (Figure 6).

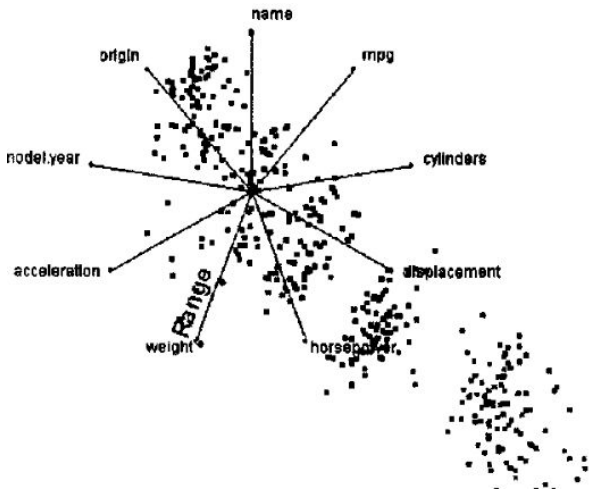

Figure 6 - Points corresponding to a set/range are marked with different colors(not visible)

Footprints

When several transformations are applied to the data for better visualization, the data points get repositioned. However it is sometimes more useful to look at the path followed by these data points. Such visualization is achieved in the article using Footprints as shown in Figure 7.

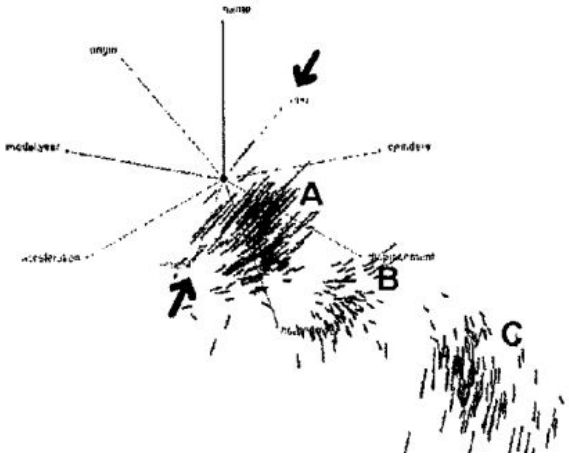

Figure 7 -Footprints on multi-dimensional scaling of mpg and weight

Churn analysis

Churn means the cancellation of services of the customers. The author has given a practical example of a telecom company. Telecom companies collect data from different sources like billing, call information, subscriptions to improve their services and create promotions for customer's loyalty.

The 'churn' dataset examined by the author contains information like account length, phone number, local and international minutes, call and voice mail etc. Fig 6 show overview of the dataset where churned attributes are turned off to remove its contribution to the location of the data points.

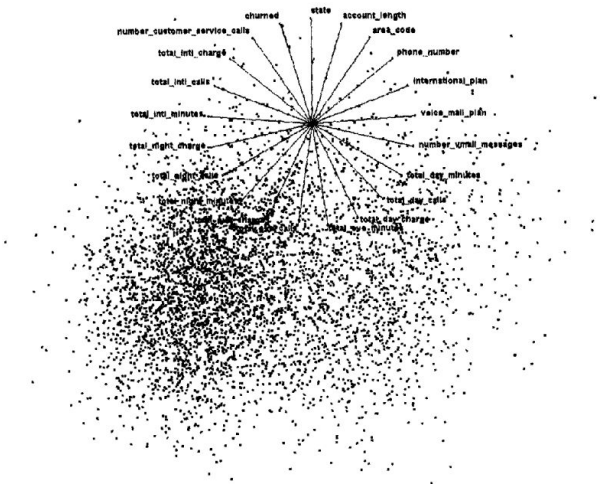

Figure 8 - Overview of the churn dataset

To amplify the analysis the author then turn off attributes for **total minutes** and calls for day, evening, night and international calls while examining only the **charge amounts** for total day, evening, night and international charges. It was observed that there is no discrepancies in these attributes that leads to churn as indicated by the footprint which are parallel in Figure 9.

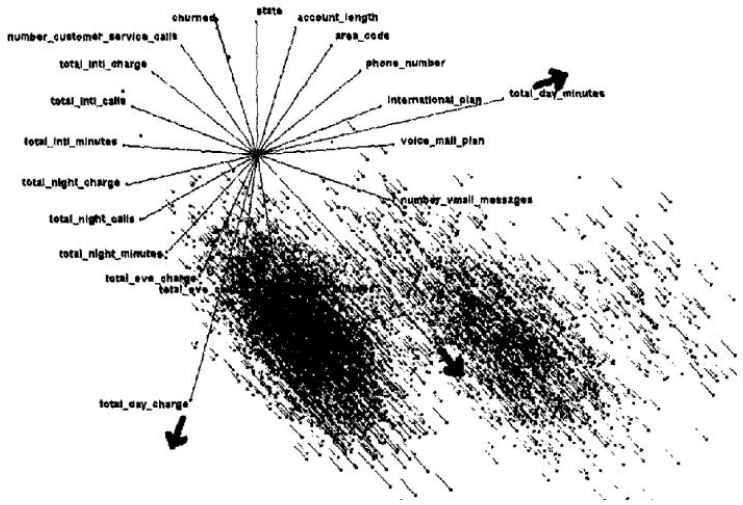

Figure 9 - Effect of total day, minute and charge attributes on churn

After removing the attributes stated earlier and scaling up international plan indicates that customers with international plan churns more (cluster C & D). However customers (cluster B) with voice mail are less likely to churn as shown in Figure 9.

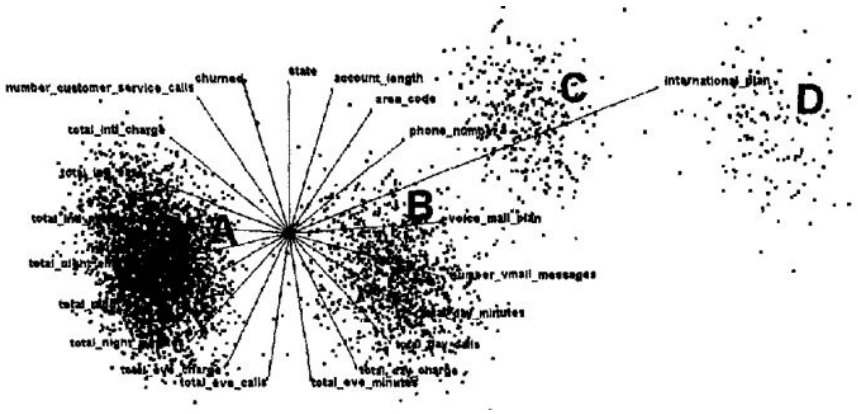

Figure 9 - Four clusters (A..D) based on international and voice plans

It was observed in this article that after scaling total day charge and number of customer service calls which indicates that these attributes play the most significant role in churn of customers (without international and voice plans i.e cluster A) as shown in Figure 10.

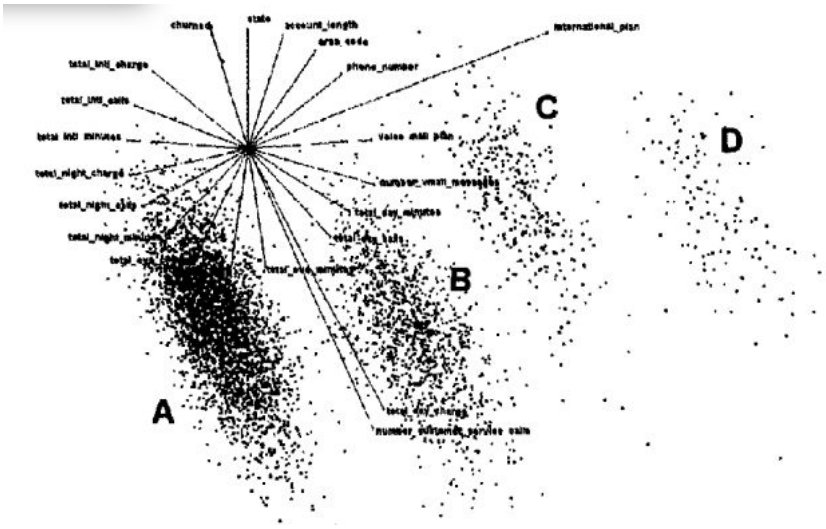

Figure 10 - Total day charge and number of customer service play the most significant role in churn

Strengths

- Very short and concise description in different sections of the article is given
- The analysis part is given is elaborated very nicely and precisely. Each observation with logical reasoning is written.
- Very fluent and scientific language has been used in the whole article, there are no noticeable grammatical, spelling errors.
- Not too short or too long sentences in the article. No unneeded complicated language has been used, simple and clear language for all explanations has been used.

Weaknesses

- Figures poorly presented and not eligible. It would be nice to visualize graphs with different markers (triangles, circles, squares etc)
- The observation for interaction techniques lack mathematical formulas (models) that has been used for scaling, rotation etc. How were the attributes summarized or the different scaling techniques were applied ? has not enough explanations.
- The analysis method has not been explained step wise and it makes it hard to reproduce the research work.
- The author has not mentioned other visualization techniques
- No reference to datasets are provided

Conclusion

We reviewed this article using analysis and evaluation method based on critical review criteria[4].

Researchers have proposed several ways of visualizing multi dimensional data with rich features of interactions to mine the data. Each of these methods and approaches have their own pros and cons and it is worthwhile to find a suitable approach for a specific problem.

In star coordinates each row represents a point (i.e two attributes) but the transformation function takes n parameters (n dimensions dataset) where each dimension contributes.

The author states that a point is a simpler representation compared to a set of connected lines and this avoids screen cluttering. Scatter plots does that same but the transformation function only takes two parameters at a time.

Star coordinates provide a promising approach to gain insight into the data, however higher dimensions > 10 are inevitable to scale this approach.

References

[1] E. Kandogan, “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*, New York, NY, USA, 2001, pp. 107–116, doi: 10.1145/502512.502530.

[2] “ACM Digital Library,” *Eser Kandogan*. [Online]. Available: https://dl.acm.org/profile/81100204008. [Accessed: 14-Jan-2020].

[3] “Research Gate.” [Online]. Available: https://www.researchgate.net/profile/Eser_Kandogan2. [Accessed: 14-Jan-2020].

[4] “Writing a critical review.” [Online]. Available: http://wwwdocs.fce.unsw.edu.au/fce/EDU/eduwritingcritreview.pdf. [Accessed: 12-Jan-2020].