Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Cluster Analysis for Multidimensional Data Using Visualization Technique

Swagati Julka[1]

Computer Science & Suresh Gyan Vihar University, Jaipur

Rajasthan, India

Abstract: Cluster analysis is one of the most important technique used in Data mining. The major drawback, of cluster analysis is that it provides numerical feedback making it difficult for users to understand, and most of the clustering algorithms are not well suited for dealing with arbitrarily shaped data distributions of datasets. The visualization techniques have been proven to be more effective in data mining; their use is cluster analysis is still a challenge, especially in applications with huge and high dimensional datasets. This paper introduces a unique approach, Hypothesis Oriented Verification and Validation by Visualization, named HOV^3, it assembles datasets on a hypothesis by visualization in 2D space. The HOV^{3}technique is goal oriented, it can reside the user to discover more cluster information from high dimensionl data sets effectively and efficiently.

Keywords: Cluster analysis, Visual Data mining, High-dimensional data Visualization.

I. Introduction

There are several clustering algorithms which are proposed in research on data mining [7]. While, most of them favor clustering spherical shaped or regular datasets, they are not very effective to deal with arbitrarily shaped clusters. The approaches reported in the literature [13, 4, 11, 5, 1, 9] attempt to overcome these problems. However they still have certain drawbacks in handling irregular shaped clusters. For example, CURE [5] and BIRCH [13] perform well in low dimensional datasets they suffer from a high computational complexity. DBSCAN [4], Wave Cluster [11] FAÇADE [9] and OPTICS [1] try to distinguish arbitrarily shaped clusters, but their non-linear complexity often them unsuitable in the analysis of very large datasets. In high-dimensional spaces, traditional clustering algorithms tend to break down in terms of efficiency as well as accuracy because data do not cluster well anymore. As a complementary technique, visualization can provide data miners with intuitive feedback on data analysis and support decision-making activities. In addition, visual presentations can be very powerful in revealing trends, highlighting outliers, showing clusters, and exposing gaps in data [12]. Many studies [2, 6] have been performed on high-dimensional data visualization, but most of them have difficulty in dealing with high dimensional and very large datasets. In applications of cluster analysis, many visualization techniques have been employed to study the structure of datasets [10], but most of them are provided as information rendering systems, because they do not focus on studying how data behavior changes along with different parameters of algorithms dynamically or interactively. In practice, those visualization techniques take the problem of cluster visualization simply as a layout problem. The approaches that are most relevant to our research are star coordinates [8] and its extensions such as VISTA [3]. We give a more detailed discussion on star coordinates in contrast with our model in the next section.

II. Background & Our Aproach

Data mining approaches are roughly categorized into discovery driven and verification driven [10]. Discovery driven methods can be regarded as discovering information by exploration, and the verification driven approach. Star coordinates [8] is a good choice as an exploration discovery tool for cluster analysis in a high-dimensional setting. Star coordinates technique and its salient features are briefly presented below.

A. Star Coordinates

Star Coordinates [8] arranges values of n-attributes of a database to n-dimensional coordinates on a two-dimensional plane. The minimum data value on each dimension is mapped to the origin, and the maximum value, is mapped to the other end of the coordinate axis. Then unit vectors on each coordinate axis are calculated accordingly to allow scaling of data values to the length of the coordinate axes. Finally the values on n-dimensional coordinates are mapped to the orthogonal coordinates. Star Coordinates uses x-y values to represent a set of points on the two-dimensional surface, as shown in figure 1.

Formula (1) states the mathematical description of Star Coordinates.

$$p_j(x,y) = \left(\sum_{i=1}^{n} \vec{u}_{xi}(d_{ji} - \min_i), \sum_{i=1}^{n} \vec{u}_{yi}(d_{ji} - \min_i)\right) \quad (1)$$

$P_j(x,y)$ is the location of D_j, which is located by the vector sum of all unit vectors (u_{xi} , u_{yi}) on each coordinate Ci; and $u_j = C_j / \max_j - \min_j$ (in which $\min_j = \min(d_{ji}, 0 \le j , n$ and $\max_j = \max(d_{ji}, 0 \le j , n)$; where n is the number of elements in dataset.

Due to mapping high-dimensional data into two-dimensional space, Star coordinates inevitably produces data overlapping and ambiguities in visual form. For mitigating these drawbacks, Star coordinates established visual adjustment mechanisms, such as scaling the weight of attributes of a particular axis, rotating angles between axes, marking data points in a certain area by coloring etc. However, Star coordinates is a typical method of exploration discovery.

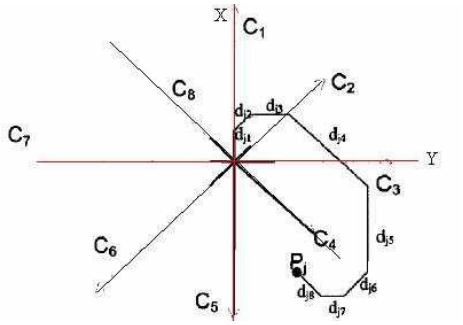

Figure 1. Positioning a point by an 8 attribute vector in Star Coordinates. [8]

1) Axis Scaling: The axis scaling in Star coordinates is to randomly adjust the weight value of each axis so that the user can see the changes dynamically. For examples, the diagrams in Fig.2 shows the original data distribution of Iris (Iris has 4 numeric attributes and 150 instances) with the clustering indices produced by the K-means clustering algorithm in iVIBRATE, where clusters overlap (K=3).

A well separated cluster distribution of Iris is described in Fig. 3, where clusters are much easier to be recognized than those of the original distribution in Fig 2.

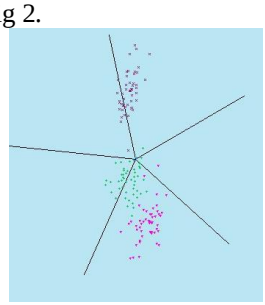

Fig. 2. The initial data distribution of clusters of Iris produced by K-means in iVIBRATE

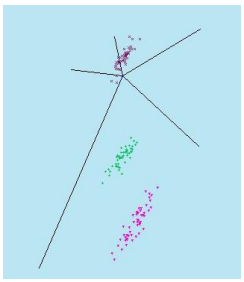

Fig. 3. The separated version of the Iris data distribution in iVIBRATE

2) Footprint: Now we use another data set auto-mpg to demonstrate the footprint feature. The data set auto-mpg has 8 attributes and 398 items. Fig. 3 presents the footprints of axis tuning of attributes "weight" and "mpg", where we find some points with longer traces and some with shorter footprints. Since its computational complexity is only in linear time. This makes them very suitable to be employed as a visual tool for interactive interpretation and exploration in cluster analysis.

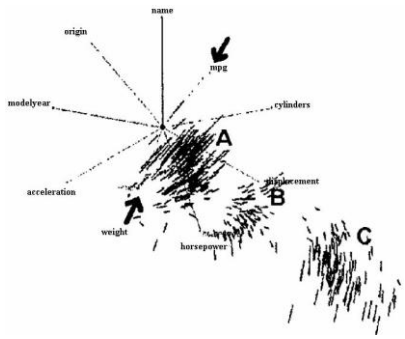

Fig. 4. Footprints of axis scaling of "weight" and "mpg" attributes in Star Coordinates [8]

However, the cluster exploration and refinement based on the users suddenly introduces randomness and subjectiveness into visual cluster analysis therefore, sometimes the adjustments of Star Coordinates and iVIBRATE could be arbitrary and time consuming. Using numerical supported cluster analysis (qualitative) is time consuming and inefficient, while using visual clustering, stochastic, and less of preciseness. To solve the problem of precision of visual cluster analysis, we introduce a new approach in the following.

B. Our approach –HOV^3

Exploration discovery (qualitative analysis) is regarded as the preprocessing of verification discovery (quantitative analysis), which is mainly used for building user hypothesis based and cluster detection. However, the way in which the qualitative analysis done by visualization mostly depends on each individual user experience. Thus subjectivity, randomness and lack of precision may be introduces in exploration-discovery, As a result, the quantitative analysis based on the result of imprecise qualitative analysis may be inefficient and time consuming.

To fill the gap between the imprecise visual cluster analysis and the unintuitive numerical cluster analysis, we propose a new approach, Hypothesis Oriented Verification and validation by Visualization, called HOV^3 synthesizes the feedback from exploration measures, and then projects test dataset against those measures.

In fact, the Star Coordinates model can be mathematically depicted by the Euler formula. According to the Euler formula: $e^{ix} = \cos x + i \sin x$, where $z = x + i.y$, and i is the imaginary unit. Let $z_0 = e^{2\pi i/n}$, such that $z_0^1, z_0^2, z_0^3, \ldots, z_0^{n-1}, z_0^n$ (where $z_0^n = 1$) divide the unit circle on the complex 2D plane into n equal sectors. Thus, Star Coordinates can be written as:

$$P_j(z_o) = \sum_{k=1}^{n} [(d_{jk} - \min d_k)/(\max d_k - \min d_k) \cdot z_o^k] \quad (2)$$

Where mind_k and maxd_k represents the minimal and maximal values of the k^{th} coordinate respectively. The idea of HOV^3 is that, in analytical geometry, the differences of a data set (a matrix) D_j and a measure vector M with the same number of variables as D_j can be represented by their inner product, $D_j.M$. HOV^3 uses a measure vector M to represent the corresponding axis, weight values. Then given a non-zero measure vector M in R^n, and a family of vectors Pj, the projection of Pj against M, according to formula (2), the HOV3 model is presented as:

$$P_j(z_o) = \sum_{k=1}^{n} [(d_{jk} - \min_k(d_k))/(\max_k(d_k) - \min_k(d_k)) \cdot z_o^k \cdot m_k] \quad (3)$$

Where m_k is the k^{th} attribute of measure M.

The aim of interactive adjustments of Star Coordinates and its extensions is to have some separated groups or full-separated clustering result of data by tuning the weight value of each axis, but their arbitrary and random adjustments limit their applicability. As shown in formula (3), HOV^3 summarizes these adjustments as a coefficient/measure vector. Comparing the formulas (2) and (3), it can be observed that HOV^3 subsumes the Star Coordinates model [14]. Thus the HOV^3 model provides the user a mechanism to quantify a prediction about a data set as a measure vector of HOV^3 for precisely exploring grouping information.

Moreover, not only does HOV^3 support quantifies domain knowledge verification and validation, it can also be directly utilize rich statistical analysis tools, such as mean, median, standard deviation, etc.

III. EXPERIMENTS WITH HOV^3

There are several statistical measurements, such as median, mean, standard deviation, and etc. can be directly introduced into HOV^3 as predictions to explore data distributions. In fact, it gives an easier interpretation of data distribution. We use the Iris dataset as an example. The datasets we used in the examples are available from the UCI machine learning website. Iris has 4 numeric attributes and 150 instances with the clustering indices produced by the K-means clustering algorithm in iVIBRATE, where clusters overlap (K=3).

As shown in fig. 3, by random axis scaling the user can divide the Iris data into several 3 groups. In this example we will explain the cluster exploration based on random adjustment may expose data grouping information, but sometimes it is hard to interpret such groupings. We take standard deviation of Iris M= [0.2302, 0.1806, 0.2982, 0.3172, 0.4089] as a prediction to project Iris by HOV^3 in iVIBRATE. The result is shown in fig. 5, where 3 groups clearly exist. We can see that in fig. 5, there is a blue point in the pink-colored cluster and a pink point in the green-colored cluster, resulting from the K-means clustering algorithm with K=3. Randomly, they have been wrongly clustered. We re-clustered them by their distributions, as shown in fig. 6.

The summarized results of both clusters (C_K) produced by the K-means clustering algorithm and new clustering result (C_H) projected by HOV^3 is summarized in Table 1. We can clearly see that the quality of the new clustering result of Iris is far much better that that obtained by K-mean according to the "Variance" comparison.

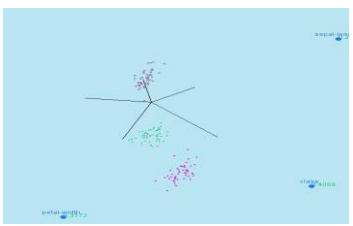

Fig. 5. Data distribution projected by HOV^3 in iVIBRATE of Iris with cluster indices make by K-means.

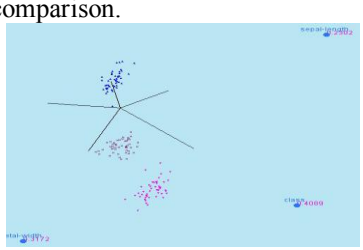

Fig. 6. Data distribution projected by HOV^3 in iVIBRATE of Iris with the new clustering indices by the user's intuition

Each cluster projected by HOV^3 has a higher similarity than that produced by K-means. By analyzing the new grouping data points of Iris, it is noticed that they are distinguished by the "class" attribute of Iris, i.e. Iris-setosa, Iris-versicolor and Iris-virginica. The cluster 1 that is generated by K-means is an outlier.

C_k	%	Radius	Variance	MaxDis	C_H	%	Radius	Variance	MaxDis
1	1.333	1.653	2.338	3.306					
2	32.667	5.754	0.153	6.115	*1*	*33.333*	*5.753*	*0.152*	*6.113*
3	33.333	8.196	0.215	8.717	*2*	*33.333*	*8.210*	*0.207*	*8.736*
4	33.333	7.092	0.198	7.582	*3*	*33.333*	*7.112*	*0.180*	*7.517*

Table 1. The statistics of the cluster in Iris produced by HOV^3 with predictive measure

With the statistical predictions ion HOV^3 the user may even expose the clusters clues that are not easy to be found by random adjustments. In addition, HOV^3 can repeat the results of VISTA, if the user can record each weight scaling and quantified them, since HOV^3 model covers Star Coordinates based techniques. Experiments on the Iris dataset also show that HOV^3 has the capacity to provide users an efficient and effective method to verify their hypothesis by visualization. We also performed more experiments on other well- known datasets, such as Shuttle, Autompg, Wine etc. But, due to space limitations, it is unable to discuss them here.

IV. Related Work

Visualization is typically employed as an observational mechanism to assist the users with intuitive comparisons and better understanding of the studied data. Instead of quantitative focusing on clustering results, most of the visualization techniques in cluster analysis focus on providing users with an easy and understanding clustering structure.

Therefore, Multidimensional Scaling, MDS [15] and Principal Component Analysis [16] are two commonly used for multivariate analysis techniques. However, the relative high computational cost of MDS (polynomial time $O(N^2)$) limits its usability in very large datasets, and PCA first has to find the correlated variables for reducing the dimensionality, which makes it not suitable for unknown data exploration.

OPTICS [1] uses a density-based technique to detect cluster structure and visualizes cluster in "Gaussian bumps", but its non-linear time complexity makes it neither suitable for dealing with very large data sets. H-BLOB visualizes clusters into blob manner in a 3D hierarchical structure [19]. It is an intuitive cluster rendering technique, but its 3D and two stages expression restricts it from interactively investigating cluster structures apart from existing clusters.

Self-organizing maps (SOM) [18] are used to project high-dimensional data sets to 2D space for matching visual models. However, the SOM technique is based on a single projection strategy and it is not powerful enough to discover all the interesting features from the original data set. Huang et. Al [17] proposed the approaches based on FastMap [21] to assist users in identifying and verifying the validity of clusters in visual form. Their techniques work well in cluster identification, but are unable to evaluate the cluster quality very well. On the other hand, these techniques are not well suited to the interactive investigation of data distributions of high-dimensional data sets. A recent survey of visualization techniques in cluster analysis can be found in the literature [20].

V. Conclusions

In this paper we have proposed a new approach called HOV^3 to assist users in visual cluster analysis in high-dimensional datasets. HOV^3 employs hypothesis-oriented measures to project data in two-dimensional space and allows users to iteratively adjust the measures for optimizing the results of clusters. HOV^3 can be seen as a bridging process between qualitative analysis and quantitative analysis. Experiments show that HOV^3 can improve the effectiveness of the cluster analysis by visualization and provide a better, intuitive understanding of the results.

Acknowledgement

Swagati Julka wishes to acknowledge my guide Mr. Sandeep Bhargava and other contributors for developing the "Cluster Analysis for Multidimensional Data using Visualization Technique" which is based on a visual approach called HOV^3 , *Hypothesis Oriented Verification and Validation by Visualization,* to assist data miners in cluster analysis.

References

[1] Ankerst M., Breunig MM., Kriegel HP., Sander J. OPTICS: Ordering points to indentify the clustering structure. Proc.of ACMSIGMOD Conference, 1999.

[2] Ankerst M., and Keim D. Visual Data Mining and Exploration of Large Database, 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg Germany, September 2001.

[3] Chen K. and Liu L. VISTA: Validating and Refining Clusters via Visualization. Journal of Information Visualization. Vol3 (4) 257-270, 2004.

[4] Ester M., Kriegel HP., Sander J., Xu. X., A density-based algorithm for discovering clusters in large spatial databases with noises. Second International Conference on Knowledge Discovery and Data Mining, 1996.

[5] Guha S., Rastogi R., Shim K. CURE : An efficient clustering algorithm for large databases. Proc. of ACM SIGMOD Conference, 1998.

[6] Hoffman P.E. and Grinstein G., A survey of visualizations for high-dimensional data mining, Information visualization in data mining and knowledge and discovery, Morgan Kaufmann Publishers Inc. August 2001.

[7] Jain A., Murty M.N., Flynn PJ., Data Clustering: A Review. ACM Computing Surveys, 31(3), 264-323,1999.

[8] Kandogan E., Visualizing multi-dimensional clusters, trends and outliers using star coordinates. Proc. of ACM SIGKDD Conference, 107-116, 2001.

[9] Qian Y., Zhang G., and Zhang K.: FAÇADE: A Fast and Effective Approach to the Discovery of Dense Clusters in Noisy Spatial Data, In Proc, ACM SIGMOD 2004 Conference, Paris, France,13-18 June 2004, ACM Press, 921-922,2004.

[10] Ribarsky W., Katz J., Holland A., Discovery visualization using fast clustering, Computer Graphics and Applications, IEEE, Volume 19(5) 32-39,1999.

[11] Sheikholeslami G., Chatterjee S., Zhang A., WaveCluster: A multi-resolution clustering approach for very large spatial databases. Proc. Of Very Large Databases Conferences (VLDB), 1998.

[12] Shneiderman B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. Discovery Science 17-28, 2001 Proc. Lecture Notes in Computer Science 2226 Springer 2001.

[13] Zhang T ., Ramakrishnan R. and Livny M., BIRCH: An efficient data clustering method for very large datasets, In Proc. Of SIGMOD96, Montreal, Canada, 103-114-1996.

[14] Zhang, K-B., Orgun, M.A., Zhang, K.: HOV^3, An Approach for cluster analysis. In: Li, X., Zaiane, O.R., Li, Z. (eds.) ADMA 2006. LNCS (LNAI), vol.4093, pp. 317-328. Springer, Heidelberg (2006)

[15] Kruskal, J.B.., Wish, M.: Multidimensional Scaling, SAGE university paper series on quantitive applications in the social sciences, pp. 7-11. Sage Publications, CA (1978)

[16] Jolliffe Ian, T.: Principal Component Analysis. Springer Press, Heidelberg (2002)

[17] Huang, Z., Cheung, D.W., Ng, M.K.: An Empirical Study on the Visual Cluster Validation Method with Fastmap. In: Proc. Of DASFAA01, pp. 84-91 (2001)

[18] Kaski, S., Sinkkonen, J., Peltonen, J.: Data Visualization and Analysis with Self Organising Maps in Learning Matrics. In: Kambayashi, Y., Winiwater, W., Arikawa, M.9eds) DaWaK 2001. LNCS, vol. 2114, pp.162-173. Springer, Heidelberg (2001)

[19] Sprenger, T.C, Brunella, R., Gross, M.H.: H-BLOB: A Hierarchical Visual Clustering Method Using Implicit Surfaces. In: Proc. Of the conference on visualization '00, pp. 61-68. IEEE Computer Society Press, Los Alamitos (2000)

[20] Seo, J., Shneiderman, B.: From Integrated Publication And Information Systems to virtual Information and Knowledge Environments. LNCS, vol 3379, Springer, Heidelberg (2005)

[21] Faloutsos, C., Lin, K.: Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia data sets. In: Proc. Of ACM-SIGMOD, pp. 163-174 (1995)

