International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Analysis of High Dimension Clustered Data using Visualization Technique

Abhimanyu Kumar[1], Archie Jain[2]

[1]Suresh GyanVihar University, Jaipur

[2]PDM College Of Engineering, Haryana

Abstract: ***In Data mining, One of the most significant and important technique is Cluster Analysis. The major stumbling block of cluster analysis is that it provides numerical feedback making it Onerous for users to understand, and most of the clustering algorithms are not pertinent for dealing with arbitrarily shaped data distributions of datasets. In Data mining, the visualization techniques have been proven to be more effectual; their accomplishment in cluster analysis is still a hurdle, especially in applications with immense and high dimensional datasets. In This paper I have introduced a distinct approach, Hypothesis Oriented Verification and Validation by Visualization, named HOV3; it gathers datasets on a hypothesis by visualization in 2D space. The HOV3technique is goal oriented, it can reside the user to discover more cluster information from high dimensional data sets in a productive and organized way.***

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

1. Introduction

For Research on Data mining there are numerous clustering algorithms have been proposed. [7]. among all those research, most of them favor clustering spherical shaped or regular datasets; they have not much potent to deal with arbitrarily shaped clusters. To conquer these problems there are approaches proclaimed in the literature [13, 4, 11, 5, 1, 9]. Still in handling irregular shaped clusters they have certain deterrents. For example, CURE [5] and BIRCH [13] endure from a high computational complexity but they accomplish well in low dimensional datasets. Arbitrarily shaped clusters is distinguished by DBSCAN [4], Wave Cluster [11] FAÇADE [9] and OPTICS [1], but their non-linear complexity often prove them ill-suited in the analysis of very large datasets. Non-clustering of data tends to breaks down the algorithms in terms of effectiveness as well as preciseness in high dimensional spaces. By considering it as a complementary technique, Visualization can be useful for data miners as it can provide instinctive feedback on data analysis and moreover also can be supportive in decision-making activities. Further, in revealing trends, highlighting outliers, showing clusters and revealing gaps in data, visual representation can be very effective[12].Numerous studies [2, 6] have been accomplished on high-dimensional data visualization, but nearly all ofthem struggle in dealing with high dimensional and very large datasets. To study the structure of the datasets [10], numerous visualization methodologies have been used in application of cluster analysis, but almost all of them are presented as information rendering systems, the reason is they never concentrate on doing the analysis that in what way the data behavior changes along with different parameters of algorithms dynamically.In Practice, the problem of cluster visualization is simply considered as a layout problem by those visualization techniques. Star coordinates [8] and its extensions such as VISTA [3] are the approaches that are most pertinent to our research. In the next section a more detailed discussion on star coordinates in contrast with our model is given.

2. Background & Our Approach

Discovery driven and verification driven are the two roughly categorized approaches of data mining [10].Discovery driven technique is mainly the idea of discovering information by exploration method, and the verification driven approach.As an exploration discovery tool for cluster analysis in a high dimensional setting.Star coordinates [8] is a good choice. Star coordinates and its most important features are concisely explained below

a) Star Coordinates

On a two-dimensional plane Star coordinates [8] arranges value of n-attributes of a database to n-dimensional coordinates. On each dimension the maximum data value is mapped to the other end of the coordinate axis and the minimum data value is mapped to the origin. To permit scaling of data values to the length of the coordinate‟s axis, the unit vectors on each coordinates axis are calculated accordingly. Lastly the mapping of the values on n-dimensional coordinates to the orthogonal coordinates happens. As shown in the figure 1, to represent a set of points on the two-dimensional surface, Star Coordinates uses x-y values. The mathematical illustration of Star coordinates is stated by the Formula (1).

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

Pj (x,y) is the location of Dj, which is located by the vector sum of all unit vectors (uxi , uyi) on each coordinate Ci; and uj= Cj / maxj – minj (in which minj =min(dji, $0 \leq j$, n and maxj = max(dji, $0 \leq j$, n); where n is the number of elements in dataset.

Star Coordinates unavoidably produces data overlapping and uncertainty in visual form because to mapping high-dimensional data into two-dimensional space. To alleviate these snags, visual adjustment mechanisms is established by Star coordinates, such as rotating angles between axes, scaling the weight of attributes of a particular axis, marking

data points in a certain area by coloring etc. However, Star coordinates is a quintessential method of exploration discovery.

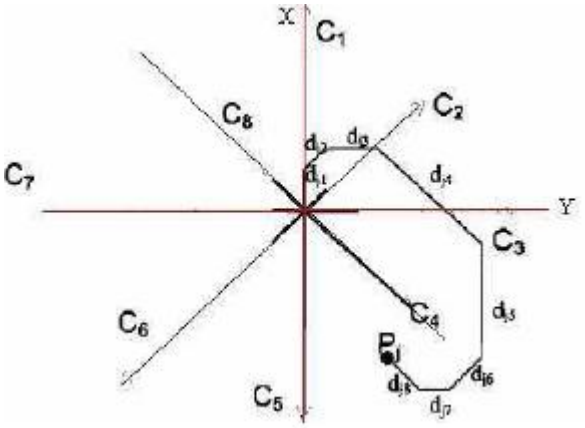

Figure 1: Positioning a point by an 8 attribute vector in Star Coordinates. [8]

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

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.

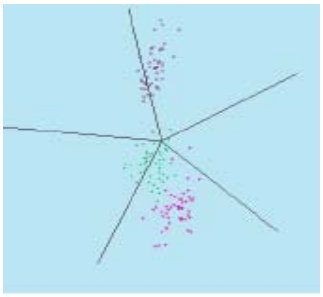

Figure 2: The initial data distribution of Clusters of Iris produced by K-means in iVIBRATE

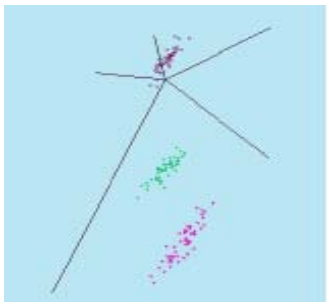

Figure 3: The separated version of the Iris data distribution in iVIBRATE

*2) Footprint:*To display the footprinttraitnow we use another dataset auto-mpg. There are 8 attributes and 398 items that the data set auto-mpg has. Fig. 3 demonstrates the footprints of axis tuning of attributes "weight" and "mpg", where we can find some point with shorter footprints and some with longer traces. Because, it"s computational complexity is only in linear time. In cluster analysis, they are very appropriate to be employed as a visual tool for interactive interpretation and exploration.

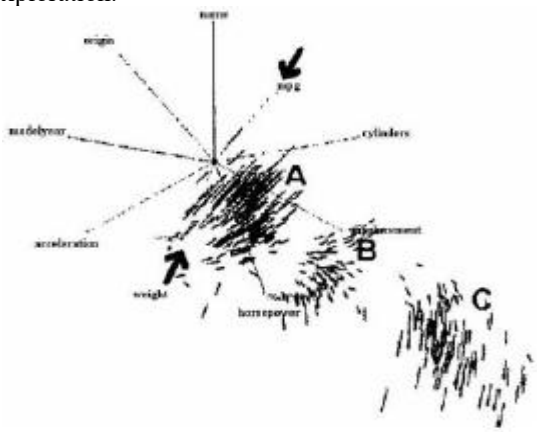

Figure 4: Footprints of axis scaling of "mpg" and "weight" attributes in Star Coordinates [8]

However, As the randomness and subjectiveness into visual cluster analysis is suddenly introduced by the cluster exploration and refinement based on the users. So, Some times the adjustment of star coordinates and iVIBRATE could be arbitrary and prolonged. Using visual clustering is stochastic and less of preciseness, while using numerical supported cluster analysis (qualitative) is time consuming and ineffectual. We have introduced a new approach to overcome the obstacle of visual cluster analysis.

A. Our approach –HOV3

For building user hypothesis based on cluster detection, Exploration discovery (qualitative analysis) which is considered as the preprocessing of verification discovery (quantitative analysis) is primarily used. However, the process of qualitative analysis done by visualization mostly depends on each individual"s user experience. As a result of the introduction of Lack of precision, randomness and subjectivity in exploration – discovery, It makes quantitative analysis inefficient and time consuming which is based on the result of imprecise qualitative analysis method.

To lessen the difference between the unintuitive cluster analysis and the imprecise visual cluster analysis, we have introduced a new method which is Hypothesis Oriented Verification and validation by Visualization which is also called HOV3 which synthesizes the response from exploration measures, and then forecasts test datasets against those measures.

In fact Euler formula can be used to delineate the Star Coordinates model mathematically. And as per Euler formulaeix = cosx+isinx, where z= x + i.y, and i is the imaginary unit. Let z0=e2πi/n, such that z01, z02, z03, …., z0n-1, z0n (where z0n=1) divide the unit circle on the complex 2D plane into n equal sectors. Thus, Star Coordinates can be represented as:-

$$P_j(z_0) = \sum_{k=7}^{n} [(d_{jk} - \min d_k)/(\max d_k - \min d_k) \cdot z_0^k] \quad (2)$$

Where the minimal and maximum values of the kth coordinate is represented by the values min dk and max dk respectively. the differences of a data set (a matrix) Dj and a measure vector M with the same number of variables as Dj can be represented by their inner product, Dj.M as an idea of HOV3 in analytical geometry. For the representation of the corresponding axis, weight values a measure vector M is used by HOV3. Then given a family of vectors Pj, a non-zero measure vector M in Rn, the projection of Pj against M, according to formula (2), the representation of HOV3 model can be done as below:-

$$P_j(z_0) = \sum_{k=1}^{n} [(d_{jk} - \min_k(d_k))/(\max_k(d_k) - \min_k(d_k)) \cdot z_0^k \cdot m_k] \quad (3)$$

The kth attribute of measure M is represented here with m_k.

To have some separated groups or full-separated clustering result of data by tuning the weight value of each axis is the main purpose of interactive adjustments of Star Coordinates,but their arbitrary and random adjustments restrict their applicability.HOV3 abridged these adjustments as a coefficient/measure vector as shown in formula (3).It can be perceived that HOV3 subsumes the Star Coordinates model [14] by comparing the formulas (2) and (3).To measure the quantity of a prediction about a data set as a measure vector of HOV3 for precisely exploring grouping information, HOV3 model provides a mechanism to users.

HOV3 is not only limited by supporting quantifying domain knowledge verification and validation but it can also be directly utilize rich statistical analysis tools, such as mean, median, standard deviation, etc.

3. Experiments with HOV3

In HOV3 there are several statistical measurements that can be directly introduced as prediction to explore data distributions such as median, mean, standard deviation, and etc. In Fact, it provides an easier elucidation of data distribution. Iris dataset we use as an example. UCI machine learning website has all the datasets which we used in the examples. In iVIBRATE, where cluster overlap (K=3), Iris has 4 attributes and 150 instances with the cluster indices produced by the K-means clustering algorithm.

Iris data can be divided into several 3 groups by the user using random axis scaling which is shown in fig. 3.Cluster exploration based on random adjustments may expose data grouping information which we will explain in this example, but such groupings are sometimes hard to interpret. To project Iris by HOV3 in iVIBRATE we take standard deviation of Iris M= [0.2302, 0.1806, 0.2982, 0.3172, 0.4089] as a prediction. The outcome of the example has been demonstrated in fig.5, where group 3 clearly exists. In fig. 5, we can see that as a result from K-means clustering algorithm with K=3, there is a pink point in the green-colored cluster and a blue point in the pink- colored cluster. But we can see that randomly, they have been wrongly clustered which we have re-clustered them again by their distribution which can be seen in fig. 6.

Figure 5: Data distribution projected by HOV3 in with cluster indicesmake by K-means.

Figure 6: Data distribution projected by HOV3 in iVIBRATE of iVIBRATE of Iris Iris with the new clustering indices by user‟s institution

If we compare cluster projected by HOV3 and by K-means then we can see that each cluster projected by Hov3 has a higher similarity than that which is produced by K-means.After analyzing the new grouping data point of Iris, it is observed that theyare distinguished by the „‟class” attribute of Iris, i.e. Iris-virginica, Iris-setosa and Iris-versicolor. The cluster 1 is an outlier which is generated by K-means.

Table 1: The statistics of the cluster in Iris produced by HOV3 with predictive measure

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

The user may even reveal the cluster clues that are not easy to be found by random adjustments with the statistical predictions in HOV3. Since HOV3 model covers Star Coordinates based techniques. So, HOV3 can repeat the results of VISTA, Additionally; VISTA‟s result can be repeated by HOV3, if the user can record each weight scaling and quantified them. HOV3 has the capacity to bestow users an efficient and effectual method to verify their hypothesis by visualization and it can be observed by the experiments on the Iris dataset. Due to space limitations, we are unable to discuss some of our more experiments which we also have performed on other well-known datasets such as Autompg, Wine, shuttle etc.

4. Related Work

To assist the users with the instinctive comparisons and better understanding of the studies data, Visualization is typically employed as an observational mechanism. Most of the visualization techniques in cluster analysis focus on providing users with an easy and understanding clustering structure instead of quantitative focusing on clustering results.

Therefore, the two commonly used multivariate analysis techniques are Principal Component Analysis [16] and Multidimensional Scaling, MDS [15]. Although, In very large datasets the relative high computational cost of MDS (polynomial time O(N2)) limits its usability, and for reducing the dimensionality PCA first has to find correlated variables, and this is the reason that makes it not suitable for unknown data exploration.
A density-based technique is used by OPTICS [1] to detect visualization cluster and cluster structure in Gaussian bumps. But because of its non-linear complexity it is not appropriate for dealing with very large data sets. H-BLOB visualizes clusters into blob manner in a 3D hierarchical structure [19].Although it"s an instinctive cluster rendering technique, but it is restricted from interactively investigating cluster structures apart from existing clusters by its own 3D and two stages expression.

For matching visual models, Self-Organizing maps (SOM) [18] are considered to project high-dimensional dataset to 2D space for matching visual models. However, it is not sufficient enough to discover all the interesting features from the original data sets. Also, SOM technique is based on a single projection strategy. To help users in acknowledging and verifying the validity of clusters in visual form, Huang et. Al [17] introduced the approaches based on FastMap [21]. Their approaches are not able to assess the cluster quality very well but they work well in cluster identification. Also, these methods are not a good approach for interactive investigation of data distributions of high- dimensional data sets. A recent survey is listed in the literature [20] of visualization techniques in cluster analysis.

5. Conclusions

In this paper, to assist users in visual clusters in high-dimensional datasets, we have proposed a new approach called HOV3. To project data in two-dimensional space and grant users to iteratively adjust the measures for optimizing the results of cluster, HOV3 employs hypothesis oriented measures. It can also be noticed that between qualitative analysis and quantitative analysis HOV3 act as a bridging process. As a result of experiments it has been also noticed that the effectiveness of the cluster analysis by visualization. Also it provides a better, intuitive understanding of the results.

6. Acknowledgement

I wish to acknowledge my Co-author Archie Jain and my guide Mr. Dinesh Goyal and other who contributed their time for developing the "Cluster Analysisof Multidimensional Data by using the Visualization Technique" and whichis based on a visual approach called HOV3, 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., KriegelHP., 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.: HOV3, 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)

Author Profile

Abhimanyu Singh completed his graduation in the field of „Computer Science‟ from Suresh Gyan Vihar University, Jaipur and is currently working with Xerox Business Services, India.

Archie Jain completed her graduation in the field of „Information Technology‟ and is currently working with Xerox Business Services, India