Hypothesis Oriented Cluster Analysis in Data Mining by Visualization

Ke-Bing Zhang Mehmet A. Orgun
Department of Computing
Macquarie University
Sydney, NSW 2109, Australia
612- 9850 9590, 612- 9850 9570
{kebing, mehmet}@ics.mq.edu.au

Kang Zhang
Department of Computer Science
The University of Texas at Dallas
Richardson, TX 75083-0688
USA 1-972-883 6351
kzhang@utdallas.edu

Yihao Zhang
Department of Computing
Macquarie University
Sydney, NSW 2109, Australia
612- 9850 9590
yihao@ics.mq.edu.au

ABSTRACT
Cluster analysis is an important technique that has been used in data mining. However, cluster analysis provides numerical feedback making it hard for users to understand the results better; and also most of the clustering algorithms are not suitable for dealing with arbitrarily shaped data distributions of datasets. While visualization techniques have been proven to be effective in data mining, their use in cluster analysis is still a major challenge, especially in data mining applications with high-dimensional and huge datasets. This paper introduces a novel approach, ***H**ypothesis **O**riented **V**erification and **V**alidation by **V**isualization*, named HOV^3, which projects datasets based on given hypotheses by visualization in 2D space. Since HOV^3 approach is more goal-oriented, it can assist the user in discovering more precise cluster information from high-dimensional datasets efficiently and effectively.

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

1. INTRODUCTION
Many clustering algorithms have been proposed in research on data mining [7]. While, most of them favor clustering spherical shaped or regular shaped 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, but when dealing with irregular cluster distributions of high-dimensional datasets they suffer from a high computational complexity. DBSCAN [4], WaveCluster [11] FAÇADE [9] and OPTICS [1] try to distinguish arbitrarily shaped clusters, but their non-linear complexity often makes them unsuitable in the analysis of very large datasets.

AVI '06, May 23-26, 2006, Venezia, Italy.

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 behaviour 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.

2. RELATED WORK & OUR APPROACH
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 can be thought of as discovering information by verification. 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.

2.1 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 X and Y, which share the origin point with n-dimensional coordinates. Star Coordinates uses x-y values to represent a set of points on the two-dimensional surface, as shown in Figure 1.

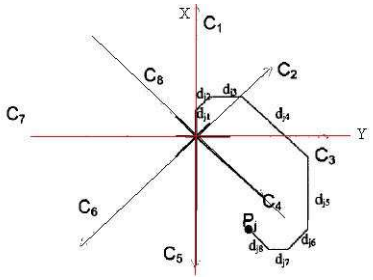

Figure 1. Positioning a point by an 8-attribute vector in Star Coordinates

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 Dj, which is located by the vector sum of all unit vectors ($\vec{u}_{xi}$, $\vec{u}_{yi}$) on each coordinate Ci; and,

$\vec{u}_j = \frac{\vec{C}_j}{man_j - \min_j}$, in which, min**j**=min{dji, 0 ≤ j < n},

max**j**=max{dji, 0 ≤ j < n}; where n is the number of elements in the 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.

Using numerical supported cluster analysis (qualitative) is time consuming and inefficient, while using visual clustering approaches (quantitative), such as Star Coordinates is subjective, stochastic, and less of preciseness. To solve the problem of precision of visual cluster analysis, we introduce a new approach in the following.

2.2 Our approach –HOV³

Exploration discovery (qualitative analysis) is regarded as the pre-processing of verification discovery (quantitative analysis), which is mainly used for building user hypotheses based on cluster detection, or other techniques. 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 introduced 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, **H**ypothesis **O**riented **V**erification and **V**alidation by **Vi**sualization, called HOV³, which is a quantified knowledge based analysis and also a bridging process between qualitative analysis and quantitative analysis. HOV³ synthesizes the feedback from exploration discovery and user domain knowledge to produce quantified measures, and then projects test dataset against those measures.

- **Mathematic Model of HOV³**

To project a high dimensional dataset to two-dimensional surface, we adopt the Polar Coordinates representation. Thus any vector can be easily transformed to the orthogonal coordinates X and Y.

In analytic geometry, the difference of two vectors A and B can be presented by their inner/dot product, A.B. Let A=(a_1, a_2, ...,a_n) and B=(b_1, b_2, ..., b_n), then their inner product can be written as:

$$\langle B, A\rangle = b_1.a_1 + b_2.a_2 + \ldots + b_n.a_n = \sum_{k=1}^{n} b_k a_k \quad (2)$$

Then we have the equation: $\cos(\theta) = \frac{\langle A,\ B\rangle}{|A||B|}$, where θ is the angle between A and B, and |A| and |B| are the length of A and B correspondingly, as shown below:

$$|A| = \sqrt{a_1^2 + a_2^2 + \ldots + a_n^2}\ ,\quad |B| = \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}\ .$$

Let A be a unit vector; the geometry of <A, B> in Polar Coordinates presents the gap from point B (d_b, θ) to point A, as demonstrated in Figure 2, where A and B are in 8 dimensional space.

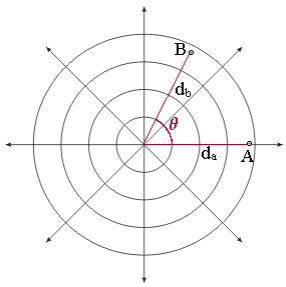

Figure 2. Vector *B* projected against vector *A* in Polar Coordinates.

- **Mapping to Measures**

In the same way, a matrix Dj, a set of vectors (dataset) can also be mapped to a measure vector M. As the result, it projects the matrix Dj distribution based on the vector M.

Let Dj=(d_{j1}, d_{j2}, ..., d_{jn}) and M=(m_1, m_2, ..., m_n); then the inner product of each vector dji, (i =1, ..., n) of Dj with M has same equation as (2) and written as:

$$\langle d_{ji}, M\rangle = m_1.d_{j1} + m_2.d_{j2} + \ldots + a_n.d_{jn} = \sum_{k=1}^{n} m_i \cdot d_{jk} \quad (3)$$

So from n-dimensional dataset to one measure (dimension) mapping F: $R^n \rightarrow R^2$ can be defined as:

$$F(Dj, M) = (\langle Dj, M\rangle) = \begin{pmatrix} m_1 d_{11}, m_2 d_{12}, \ldots, m_n d_{1n} \\ m_1 d_{21}, m_2 d_{22}, \ldots, m_n d_{2n} \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ m_1 d_{m1}, m_2 d_{m2}, \ldots, m_n d_{mn} \end{pmatrix} \quad (4)$$

Where Dj is a dataset with n attributes, and M is a quantified measure.

- **In Complex Number System**

Since our experiments are conducted with MATLAB (MATLAB®, The MathWorks, Inc), for understanding our approach better, we adopt complex number system in our study.

Let z = x + *i* y, where i is the imaginary unit.

According to the Euler formula: $e^{ix} = \cos x + i \sin x$

Let $z_0 = e^{2\pi i/n}$; we see that $z_0^1, z_0^2, z_0^3, \ldots\ldots, z_0^{n-1}, z_0^n$ (with $z_0^n = 1$) divide the unit circle on the complex plane into n-1 equal sectors. Then mapping in Star Coordinates (1) now can be simply written as:

$$p_j(z_0) = \sum_{k=1}^{n} [(d_{jk} - \min_k d_{jk}) / (\max_k d_{kj} - \min_k d_{jk})] z_0^k \quad (5)$$

where, $\min_k d_{jk}$ and $\max_k d_{kj}$ represent minimal and maximal values of the kth attribute/coordinate respectively.

This is the case of equal-divided circle surface. Then the more general form can be defined as:

$$p_j(z_k) = \sum_{k=1}^{n} [(d_{jk} - \min_k d_{jk}) / (\max_k d_{kj} - \min_k d_{jk})] z_k \quad (6)$$

where $z_k = e^{i*\theta k}$; Θ is the angle of neighbouring axes; and

$$\sum_{k=1}^{n} \theta_k = 2\pi$$

While, the part of $(d_{jk} - \min_k d_{jk}) / (\max_k d_{kj} - \min_k d_{jk})$ in equations (5) and (6) is a normalized one of original d_{jk}, we write it as $d_{N\,jk}$.

Thus formula (6) is written as:

$$p_j(z_k) = \sum_{k=1}^{n} d_{N\,jk} * z_k \quad (7)$$

In any case these can be viewed as mappings from R^n to C - the complex plane, i.e., $R^n \rightarrow C^2$.

Given a non-zero measure vector **m** in R_n, and a family of vectors P_j, then the projections of P_j against **m** according to formulas (4) and (7), we present our model HOV^3 as the following equation (8):

$$p_j(z_k) = \sum_{k=1}^{n} d_{N\,jk} * m_k * z_k \quad (8)$$

where m_k is the kth attribute of measure $\boldsymbol{m}$.

2.3 Discussion

In Star Coordinates [8] or VISTA [3], the purpose of scaling the weights of attributes of a particular axis is for adjusting the contribution of the attribute laid on a specific coordinate by the interactive actions, so that users might gain interesting cluster information automated clustering algorithms cannot provide.

Thus, comparing the models of Star Coordinates, in equation (7), and HOV^3 in equation (8), we observe that our model is more general than that of Star Coordinates. This is because, any change of weights in Star Coordinates model can be viewed as changing one or more values of m_k (k=1,..., n) in measure vector $\boldsymbol{m}$ in equation (4) or (8). As a special case, when all values in $\boldsymbol{m}$ are set to 1, HOV^3 is transformed into Star Coordinates model (7), i.e., no measure case. In addition, either moving a coordinate axis in opposite direction or scaling up the adjustment interval of axis, for example, from [0,1] to [-1, 1] in VISTA, is also regarded as setting the measure value as minus as its original one.

Moreover, not only does HOV^3 support quantified domain knowledge verification and validation, it can also directly utilize rich statistical analysis tools, such as mean, median, standard deviation, etc. as measures and guide users obtaining more incisive cluster information.

3. EXPERIMENTS WITH HOV^3

This section presents the results of our experiments with HOV^3, in comparison with those of VISTA [3]. We implemented our approach in MATLAB running under Windows 2000 Professional. The datasets we used in the examples are available from the UCI machine learning website: http://www.ics.uci.edu/~mlearn/Machine-Learning.html.

- Shuttle

Shuttle dataset has 10 attributes and 15,000 instances. Figures 3 and 4 show the initial data distribution in Star Coordinates produced by the VISTA system and the one in MATLAB produced by HOV^3 without any adopted measures respectively. It can be observed that the shapes of data distribution are almost exactly the same in the two approaches. The only difference is that VISTA shifted the appearance of data by 30 degrees in counter-clockwise direction.

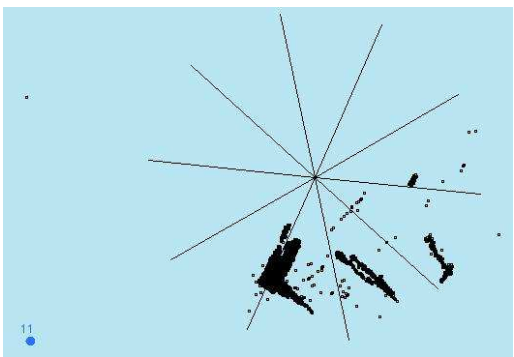

Figure 3. The original data distribution of shuttle in VISTA system

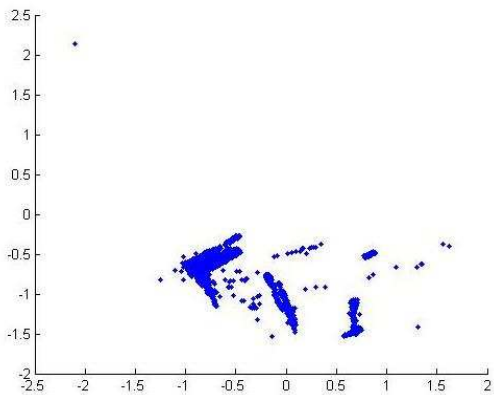

Figure 4. The original data distribution of shuttle by HOV^3 in MATLAB

These initial data distributions cannot provide users a clear idea about the clusters, since there are no criteria to cluster the dataset. Thus, in VISTA the user may verify them by further interactive actions, such as weight scaling and/or angle changing of axes. Figure 5 illustrates the results after several random weight adjustment steps, which can be observed very clearly that there are three colored data groups (clusters). Although sometimes there are better results appearing, as

shown in Figure 5, users do not even know from where those results came, because this adjustment process is pretty stochastic and not easily repeatable.

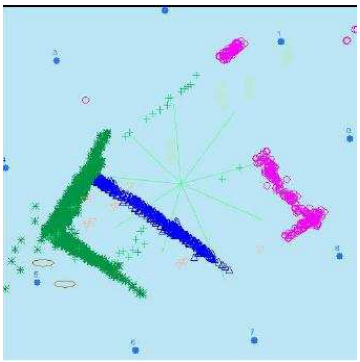

Figure 5. Post adjustment of the Shuttle data with colored labels in VISTA

We may use simple statistical methods on the Shuttle dataset as measures to detect cluster information. Figure 6 shows the data projections based on the median of Shuttle, where HOV^{3}also provides three data groups, and even more, for example, sub-clusters and outliers. Moreover, the user can clearly understand from where the results came, and repeat the experiments with the same (or different) measures.

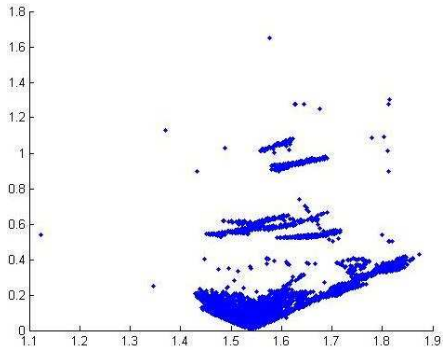

Figure 6. . Mapping shuttle dataset against to its median by HOV^3

We used the covariance matrix of Shuttle to detect the gaps of the Shuttle dataset. The results are shown in Figure 7, in which HOV^3 provides the user different cluster information as in VISTA.

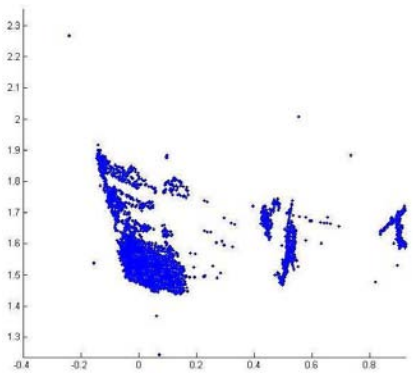

Figure 7. Mapping Shuttle dataset against its covariance matrix

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 Shuttle dataset also show that HOV^3 has the capability to provide users an efficient and effective method to verify their hypotheses by visualization. We also performed more experiments on other well-known datasets which can be accessed from UCI machine learning website, such as Iris, Autompg, Wine, etc. However, due to space limitations, we cannot discuss them here. The results of the experiments also showed that HOV^3 can reveal more precise visual exploration of data distribution to the user.

4. CONCLUDING REMARKS

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 result 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.

5. REFERENCES

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

[2] Ankerst M., and Keim D. Visual Data Mining and Exploration of Large Databases, *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 noise. *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 discovery*, Morgan Kaufmann Publishers Inc. August 2001.

[7] Jain A., Murty M. N., and Flynn P.J., 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., Jiang F., 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 Conference (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 databases, *In Proc. of SIGMOD96*, Montreal, Canada, 103-114, 1996.