A Prediction-Based Visual Approach for Cluster Exploration and Cluster Validation by HOV^3*

Ke-Bing Zhang[1], Mehmet A. Orgun[1], and Kang Zhang[2]

[1] Department of Computing, ICS, Macquarie University, Sydney, NSW 2109, Australia
{kebing,mehmet}@ics.mq.edu.au

[2] Department of Computer Science, University of Texas at Dallas
Richardson, TX 75083-0688, USA
kzhang@utdallas.edu

Abstract. Predictive knowledge discovery is an important knowledge acquisition method. It is also used in the clustering process of data mining. Visualization is very helpful for high dimensional data analysis, but not precise and this limits its usability in quantitative cluster analysis. In this paper, we adopt a visual technique called HOV^3 to explore and verify clustering results with quantified measurements. With the quantified contrast between grouped data distributions produced by HOV^3, users can detect clusters and verify their validity efficiently.

Keywords: predictive knowledge discovery, visualization, cluster analysis.

1 Introduction

Predictive knowledge discovery utilizes the existing knowledge to deduce, reason and establish predictions, and verify the validity of the predictions. By the validation processing, the knowledge may be revised and enriched with new knowledge [20]. The methodology of predictive knowledge discovery is also used in the clustering process [3]. Clustering is regarded as an unsupervised learning process to find group patterns within datasets. It is a widely applied technique in data mining. To achieve different application purposes, a large number of clustering algorithms have been developed [3, 9]. However, most existing clustering algorithms cannot handle arbitrarily shaped data distributions within extremely large and high-dimensional databases very well. The very high computational cost of statistics-based cluster validation methods in cluster analysis also prevents clustering algorithms from being used in practice.

Visualization is very powerful and effective in revealing trends, highlighting outliers, showing clusters, and exposing gaps in high-dimensional data analysis [19]. Many studies have been proposed to visualize the cluster structure of databases [15, 19]. However, most of them focus on information rendering, rather than investigating on how data behavior changes with the parameters variation of the algorithms.

* The datasets used in this paper are available from http://www.ics.uci.edu/~mlearn/Machine-Learning.html.

J.N. Kok et al. (Eds.): PKDD 2007, LNAI 4702, pp. 336–349, 2007.

In this paper we adopt HOV^3 (*Hypothesis Oriented Verification and Validation by Visualization*) to project high dimensional data onto a 2D complex space [22]. By applying predictive measures (quantified domain knowledge) to the studied data, users can detect grouping information precisely, and employ the clustered patterns as predictive classes to verify the consistency between the clustered subset and unclustered subsets.

The rest of this paper is organized as follows. Section 2 briefly introduces the current issues of cluster analysis, and the HOV^3 technique as the background of this research. Section 3 presents our prediction-based visual cluster analysis approach with examples to demonstrate its effectiveness on cluster exploration and cluster validation. A short review of the related work in visual cluster analysis is provided in Section 4. Finally, Section 5 summarizes the contributions of this paper.

2 Background

The approach reported in this paper has been developed based on the projection of HOV^3 [22], which was inspired from the Star Coordinates technique. For a better understanding of our work, we briefly describe Star Coordinates and HOV^3.

2.1 Visual Cluster Analysis

Cluster analysis includes two major aspects: clustering and cluster validation. Clustering aims at identifying objects into groups, named clusters, where the similarity of objects is high within clusters and low between clusters. Hundreds of clustering algorithms have been proposed [3, 9]. Since there are no general-purpose clustering algorithms that fit all kinds of applications, the evaluation of the quality of clustering results becomes the critical issue of cluster analysis, i.e., cluster validation. Cluster validation aims to assess the quality of clustering results and find a fit cluster scheme for a given specific application.

The user's initial estimation of the cluster number is important for choosing the parameters of clustering algorithms for the pre-processing stage of clustering. Also, the user's clear understanding on cluster distribution is helpful for assessing the quality of clustering results in the post-processing of clustering. The user's visual perception of the data distribution plays a critical role in these processing stages. Using visualization techniques to explore and understand high dimensional datasets is becoming an efficient way to combine human intelligence with the immense brute force computation power available nowadays [16].

Visual cluster analysis is a combination of visualization and cluster analysis. As an indispensable aid for human-participation, visualization is involved in almost every step of cluster analysis. Many studies have been performed on high dimensional data visualization [2, 15], but most of them do not visualize clusters well in high dimensional and very large data. Section 4 discusses several studies that have focused on visual cluster analysis [1, 7, 8, 10, 13, 14, 17, 18] as the related work of this research. Star Coordinates is a good choice for visual cluster analysis with its interactive adjustment features [11].

2.2 Star Coordinates

The idea of Star Coordinates technique is intuitive, which extends the perspective of traditional orthogonal X-Y 2D and X-Y-Z 3D coordinates technique to a higher dimensional space [11]. Technically, Star Coordinates plots a 2D plane into *n* equal sectors with *n* coordinate axes, where each axis represents a dimension and all axes share the initials at the centre of a circle on the 2D space. First, data in each dimension are normalized into [0, 1] or [-1, 1] interval. Then the values of all axes are mapped to orthogonal X-Y coordinates which share the initial point with Star Coordinates on the 2D space. Thus, an *n*-dimensional data item is expressed as a point in the X-Y 2D plane. Fig.1 illustrates the mapping from 8 Star Coordinates to X-Y coordinates.

In practice, projecting high dimensional data onto 2D space inevitably introduces overlapping and ambiguities, even bias. To mitigate the problem, Star Coordinates and its extension iVIBRATE [4] provide several visual adjustment mechanisms, such as axis scaling, axis angle rotating, data point filtering, etc. to change the data distribution of a dataset interactively in order to detect cluster characteristics and render clustering results effectively. Below we briefly introduce the two relevant adjustment features with this research.

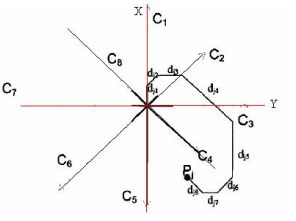

Fig. 1. Positioning a point by an 8-attribute vector in Star Coordinates [11]

- **Axis scaling**

The purpose of the *axis scaling* in Star Coordinates (called *α-adjustment* in iVIBRATE) is to interactively adjust the weight value of each axis so that users can observe the data distribution changes dynamically. For example, the diagram 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 (here k=3).

A well-separated cluster distribution of Iris is illustrated in Fig. 3 by a series of random *α-adjustments*, where clusters are much easier to be recognized than those of the original distribution in Fig 2.

For tracing data points changing in a certain period time, the *footprint* function is provided by Star Coordinates. It is discussed below.

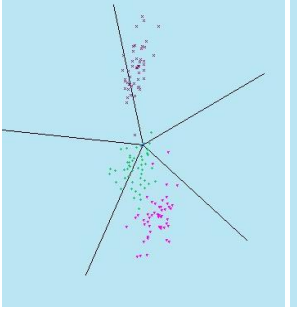

Fig. 2. The initial data distribution of clusters of Iris produced by k-means in iVIBRATE

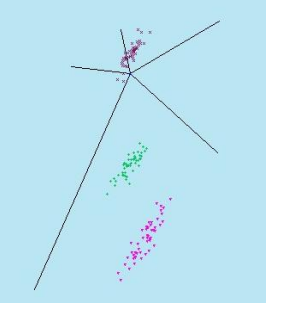

Fig. 3. The separated version of the Iris data distribution in iVIBRATE

- **Footprint**

We use another data set *auto-mpg* to demonstrate the *footprint* feature. The data set *auto-mpg* has 8 attributes and 398 items. Fig. 4 presents the footprints of axis tuning of attributes "weight" and "mpg", where we may find some points with longer traces, and some with shorter footprints.

The most prominent feature of Star Coordinates and its extensions such as iVIBRATE is that their 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.

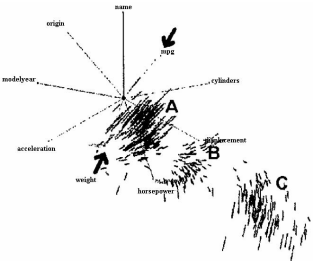

Fig. 4. Footprints of axis scaling of "weight" and "mpg" attributes in Star Coordinates [11]

However, the cluster exploration and refinement based on the user's intuition inevitably introduces randomness and subjectiveness into visual cluster analysis, and as a result, sometimes the adjustments of Star Coordinates and iVIBRATE could be arbitrary and time consuming.

2.3 HOV^3

In fact, the Star Coordinates model can be mathematically depicted by the Euler formula. According to the Eular 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 $z_0^1, z_0^2, z_0^3, \ldots, z_0^{n-1}, z_0^n$ (with $z_0^n = 1$) divide the unit circle on the complex 2D plane into n equal sectors. Thus, Star Coordinates can be simply written as:

$$P_j(z_0) = \sum_{k=1}^{n} [(d_{jk} - \min d_k)/(\max d_k - \min d_k) \cdot z_0^k] \quad (1)$$

where mind_k and maxd_k represent the minimal and maximal values of the kth coordinate respectively. In any case equation (1) can be viewed as mapping from $\mathbb{R}^n \to \mathbb{C}^2$.

To overcome the arbitrary and random adjustments of Star Coordinates and iVIBRATE, Zhang et al proposed a hypothesis-oriented visual approach called HOV^3 to detect clusters [22]. The idea of HOV^3 is that, in analytical geometry, the difference 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, $D_j \cdot M$. HOV^3 uses a measure vector M to represent the corresponding axes' weight values. Then given a non-zero measure vector M in $\mathbb{R}^n$, and a family of vectors P_j, the projection of P_j against M, according to formula (1), the HOV^3 model is presented as:

$$P_j(z_0) = \sum_{k=1}^{n} [(d_{jk} - \min d_k)/(\max d_k - \min d_k) \cdot z_0^k \cdot m_k] \quad (2)$$

where m_k is the kth attribute of measure M.

The aim of interactive adjustments of Star Coordinates and iVIBRATE 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 (2), HOV^3 summarizes these adjustments as a coefficient/measure vector. Comparing the formulas (1) and (2), it can be observed that

HOV^3 subsumes the Star Coordinates model [22]. Thus the HOV^3 model provides users a mechanism to quantify a prediction about a data set as a measure vector of HOV^3 for precisely exploring grouping information.

Equation (2) is a standard form of linear transformation of n variables, where m_k is the coefficient of kth variable of P_j. In principle, any measure vectors, even in complex number form, can be introduced into the linear transformation of HOV^3 if it can distinguish a data set into groups or have well separated clusters visually. Thus the rich statistical methods of reflecting the characteristics of data set can be also introduced as predictions in the HOV^3 projection, such that users may discover more clustering patterns. The detailed explanation of this approach is presented next.

3 Predictive Visual Cluster Analysis by HOV^3

Predictive exploration is a mathematical description of future behavior based on historical exploration of patterns. The goal of predictive visual exploration by HOV^3 is that by applying a prediction (measure vector) to a dataset, the user may identify the groups from the result of visualization. Thus the key issue of applying HOV^3 to detect grouping information is how to quantify historical patterns (or users' domain knowledge) as a measure vector to achieve this goal.

3.1 Multiple HOV^3 Projection (M-HOV^3)

In practice, it is not easy to synthesize historical knowledge about a data set into one vector; rather than using a single measure to implement a prediction test, it is more suitable to apply several predictions (measure vectors) together to the data set, we call this process *multiple HOV^3 projection*, M-HOV^3 in short. Now, we provide the detailed description of M-HOV^3 and its feature of enhanced group separation. For simplifying the discussion of the M-HOV^3 model, we give a definition first.

Definition 1. (poly-multiply vectors to a matrix) The inner product of multiplying a series of non-zero measure vectors M_1, M_2,…,M_S to a matrix $\boldsymbol{A}$ is denoted as $A \cdot * \prod_{i=1}^{s} M_i = A \cdot * M_1 \cdot * M_2 \cdot * \ldots \cdot * M_S$.

Zhang *et al* [23] gave a simple notation of HOV^3 projection as $\mathcal{D}_p = \mathcal{H}_C(\mathcal{P}, M)$, where $\mathcal{P}$ is a data set; $\mathcal{D}_p$ is the data distribution of $\mathcal{P}$ by applying a measure vector M. Then the projection of M-HOV^3 is denoted as $\mathcal{D}_p = \mathcal{H}_C(\mathcal{P}, \prod_{i=1}^{s} M_i)$. Based on equation (2), we formulate M-HOV^3 as:

$$P_j(z_o) = \sum_{k=1}^{n} [(d_{jk} - \min d_k)/(\max d_k - \min d_k) \cdot z_o^k \cdot \prod_{i=1}^{s} m_{ik}] \quad (3)$$

where m_{ik} is the kth attribute (dimension) of the ith measure vector M_i, and $s \geq 1$. When $s=1$, the formula (3) is transformed to formula (2).

We may observe that instead of using a single multiplication of m_k in formula (2), it is replaced by a poly-multiplication of $\prod_{i=1}^{s} m_{ik}$ in formula (3). Formula (3) is more

general and also closer to the real procedure of cluster detection, because it introduces several aspects of domain knowledge together into the cluster detection.

In addition, the effect of applying M-HOV^3 to datasets with the same measure vector can enhance the separation of grouped data points under certain conditions.

3.2 The Enhanced Separation Feature of M-HOV^3

To explain the geometrical meaning of M-HOV^3 projection, we use the real number system. According to equation (2), the general form of the distance σ (i.e., weighed Minkowski distance) between two points a and b in HOV^3 plane can be represented as:

$$\sigma(a,b,m) = \sqrt[q]{\sum_{k=1}^{n} | m_k (a_k - b_k) |^q} \quad (q>0) \tag{4}$$

If q = 1, σ is Manhattan (city block) distance; and if q = 2, σ is Euclidean distance. To simplify the discussion of our idea, we adopt the Manhattan metric for the explanation. Note that there exists an equivalent mapping (bijection) of distance calculation between the Manhattan and Euclidean metrics [6]. For example, if the distance between points a and b is longer than the distance between points a' and b' in then Manhattan metric, it is also true in the Euclidean metric, and vice versa.

Then the Manhattan distance between points a and b is calculated as in formula (5).

$$\sigma(a,b,m) = \sum_{k=1}^{n} \left| m_k (a_k - b_k) \right| \tag{5}$$

According to formulas (2), (3) and (5), we can present the distance of M-HOV^3 in Manhattan distance as follows:

$$\sigma(a,b,\prod_{i=1}^{s} m_i) = \sum_{k=1}^{n} | \prod_{i=1}^{s} m_{ki} (a_k - b_k) | \tag{6}$$

Definition 2. (the distance representation of M- HOV^3) The distance between two data points a and b projected by M- HOV^3 is denoted as $\overset{s}{\underset{i=1}{M}}\sigma ab$. In particular, if the measure vectors in an M-HOV^3 are the same, $\overset{s}{\underset{i=1}{M}}\sigma ab$ can be simply written as $M^s\sigma ab$; if each attribute of M is 1 (no measure case), the distance between points a and b is denoted as σab.

Thus, we have $\overset{s}{\underset{i=1}{M}}\sigma ab = \mathcal{H}_C((a,b), \prod_{i=1}^{s} M_i)$. For example, the distance between two points a and b projected by M-HOV^3 with the same two measures can be represented as $M^2\sigma ab$. Thus the projection of HOV^3 of a and b can be written as $M\sigma ab$.

We now give several important properties of M- HOV^3 as follows.

Lemma 1. In Star Coordinates space, if $\sigma ab \neq 0$ and $M \neq 0$ ($\exists m_k \in M \mid 0<|m_k|<1$), then $\sigma ab > M\sigma ab$.

Proof

$$\sigma ab = \sum_{k=1}^{n} \left| (a_k - b_k) \right| \text{ and } M\sigma ab = \sum_{k=1}^{n} \left| m_k (a_k - b_k) \right|$$

$$\sigma ab - M\sigma ab = \sum_{k=1}^{n} \left| (a_k - b_k) \right| - \sum_{k=1}^{n} \left| m_k (a_k - b_k) \right| = \sum_{k=1}^{n} \left| (a_k - b_k) \right| (1 - | m_k |)$$

$M\neq 0 \Rightarrow \{\exists m_k\neq 0 \wedge m_k\in M \mid 0<|m_k|<1, k=1\ldots n\} \Rightarrow (1-|m_k|)>0$

$\sigma ab\neq 0 \Rightarrow \sigma ab > (M\sigma ab)$ □

This result shows that the distance $M\sigma ab$ between points *a* and *b* projected by HOV[3] with a non-zero M is less than the original distance σab between *a* and *b*.

Lemma 2. In Star Coordinates space, if $\sigma ab\neq 0$ and $M\neq 0$ ($\forall m_k\in M \mid 0<|m_k|<1$), then $M^n\sigma ab > M^{n+1}\sigma ab$, $n\in \mathbf{N}$.

Proof

Let $M^n\sigma ab=\sigma' ab$

Definition 1 $\Rightarrow M^{n+1}\sigma ab= M\sigma' ab$

Lemma 1 $\Rightarrow \sigma' ab > M\sigma' ab \Rightarrow M^n\sigma ab > M^{n+1}\sigma ab$ □

In general, it can be proved that in Star Coordinates space, if $\sigma ab\neq 0$ and $M\neq 0$ ($\forall m_k\in M \mid |m_k|<1$), then $M^m\sigma ab > M^n\sigma ab$, $n\in \mathbf{N}$, $m\in \mathbf{N}$ and $m<n$.

Theorem 1. If the measure vector is changed from M to M', ($|m_k|\leq 1, |m_t+\Delta_t|<1$) and $|M\sigma ab - M\sigma ac| < |M'\sigma ab - M'\sigma ac|$ then

$$\frac{|M'\sigma ab - M'\sigma ac| - |M'^2\sigma ab - M'^2\sigma ac|}{|M'\sigma ab - M'\sigma ac|} > \frac{|M\sigma ab - M\sigma ac| - |M'\sigma ab - M'\sigma ac|}{|M\sigma ab - M\sigma ac|}$$

Proof

$$M'\sigma ab = \sum_{k=1}^{n} \left| m'_k (a_k - b_k) \right| \text{ and } M'\sigma ac = \sum_{k=1}^{n} \left| m'_k (a_k - c_k) \right|$$

$$\Rightarrow M'\sigma ab - M'\sigma ac = \sum_{k=1}^{n} |m'_k| [|(a_k - b_k)| - |(a_k - c_k)|]$$

$$M'^2\sigma ac - M'^2\sigma ab = \sum_{k=1}^{n} |m'^2_k| [|(a_k - b_k)| - |(a_k - c_k)|]$$

Let $|a_k-b_k|=x_k$ and $|a_k-c_k|=y_k$

$$\Rightarrow M'\sigma ac - M'\sigma ab = \sum_{k=1}^{n} |m'_k| [|(a_k - b_k)| - |(a_k - c_k)|] = \sum_{k=1}^{n} |m'_k| (x_k - y_k)$$

$$\Rightarrow M'^2\sigma ac - M'^2\sigma ab = \sum_{k=1}^{n} |m'^2_k| (x_k - y_k)$$

$$\Rightarrow |M'\sigma ac - M'\sigma ab| = M'\sigma xy$$

$$\Rightarrow |M'^2\sigma ac - M'^2\sigma ab| = M'^2\sigma xy$$

Lemma 2 $\Rightarrow M'^2\sigma xy < M'\sigma xy \Rightarrow \frac{|M'^2\sigma xy|}{|M'\sigma xy|} < 1 \Rightarrow \frac{|M'^2\sigma ab - M'^2\sigma ac|}{|M'\sigma ab - M'\sigma ac|} < 1$

$$\Rightarrow |M'^2\sigma ab - M'^2\sigma ac| < |M'\sigma ab - M'\sigma ac|$$

$$|M\sigma ab - M\sigma ac| < |M'\sigma ab - M'\sigma ac|$$

$$\Rightarrow |M'^2\sigma ab - M'^2\sigma ac| . |M\sigma ab - M\sigma ac| < |M'\sigma cab - M'\sigma ac|^2$$

$$\Rightarrow \frac{|M'^2\sigma ab - M'^2\sigma ac|}{|M'\sigma ab - M'\sigma ac|} < \frac{|M'\sigma ab - M'\sigma ac|}{|M\sigma ab - M\sigma ac|}$$

$$\Rightarrow 1 - \frac{|M'^2\sigma ab - M'^2\sigma ac|}{|M'\sigma ab - M'\sigma ac|} > 1 - \frac{|M'\sigma ab - M'\sigma ac|}{|M\sigma ab - M\sigma ac|}$$

$$\Rightarrow \frac{|M'\sigma ab - M'\sigma ac| - |M'^2\sigma ab - M'^2\sigma ac|}{|M'\sigma ab - M'\sigma ac|} > \frac{|M\sigma ab - M\sigma ac| - |M'\sigma ab - M'\sigma ac|}{|M\sigma ab - M\sigma ac|}$$

□

Theorem 1 shows that if the user observes that the difference of the distance between *a* and *b* and the distance between *a* and *c* are increased relatively (it can be observed by the footprints of points *a*, *b* and *c*, as shown in Fig 4) by tuning weight values of axes from M to M', then after applying M-HOV^3 to *a, b* and *c,* the distance variation rate of the distances between pairs of points *a, b* and *a, c* is enhanced, as presented in Fig 5.

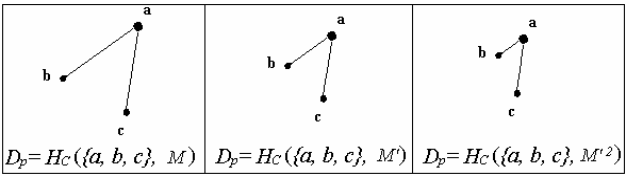

Fig. 5. The contraction and separation effect of M-HOV^3

In other words, if it is observed that several data point groups can be roughly separated visually (there may exist ambiguous points between groups) by projecting a measure vector in HOV^3 to a data set, then applying M-HOV^3 with the same measure vector to the data set would lead to the groups being more condensed, i.e., have a good separation of the groups.

3.3 Predictive Cluster Exploration by M-HOV^3

According to the notation of HOV^3 projection of a dataset $\mathcal{P}$ as $\mathcal{D}_p = \mathcal{H}_C(\mathcal{P}, M)$, the M-HOV^3 is denoted as $\mathcal{D}_p = \mathcal{H}_C(\mathcal{P}, M^n)$ where $n \in \mathbb{N}$.

We use the *auto-mpg* dataset again as an example to demonstrate predictive cluster exploration by M-HOV^3. Fig. 6a illustrates the original data distribution of *auto-mpg* produced by HOV^3 in MATLAB, where it is not possible to recognize any group information. Then we tuned each axis manually and had roughly distinguished three groups, as shown in Fig 6b. The weight values of axes were recorded as a vector M=[0.10, 0, 0.25, 0.2, 0.8, 0.85, 0.1, 0.95]. Fig. 6b shows that there exist several ambiguous data points between groups. Then we employed M^2 (inner dot) as a predictive measure vector and applied it to data set *auto-mpg*. The projected distribution $\mathcal{D}_{p2}$ of *auto-mpg* is presented in Fig 6c. It is much easier to identify 3 groups of *auto-mpg* in Fig 6c than in Fig 6b. To show the contrast between these two diagrams $\mathcal{D}_{p1}$ and $\mathcal{D}_{p2}$, we overlap them in Fig. 6d.

By analyzing the data of these 3 groups, we have found that, group 1 contains 70 items and with "original" value 2 (sourcing Europe); group 2 has 79 instances and with "original" 3 (Japanese product); and group 3 includes 249 records with "original" 1 (from USA). Actually this "natural" grouping based on the user's intuition serendipitously clustered the data set according to the "original" attribute of *auto-mpg*. In the same way, the user may find more grouping information from the interactive cluster exploration by applying predictive measurement.

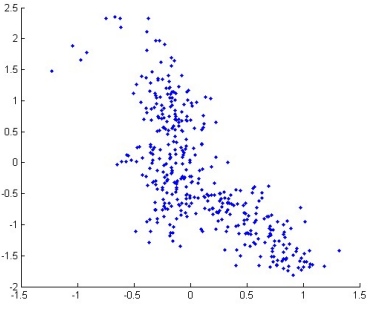

Fig. 6a. The original data distribution of auto-mpg

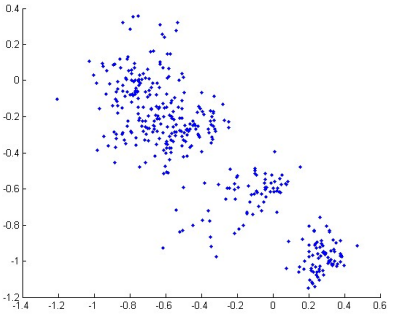

Fig. 6b. $\mathcal{D}_{p1}=\mathcal{H}_C$ (auto-mpg, M)

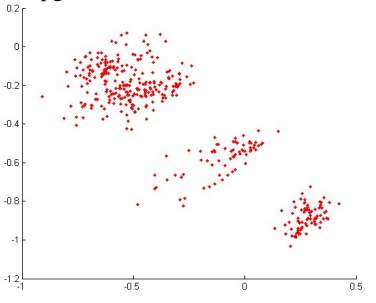

Fig. 6c. $\mathcal{D}_{p2}=\mathcal{H}_C$ (auto-mpg, M^2)

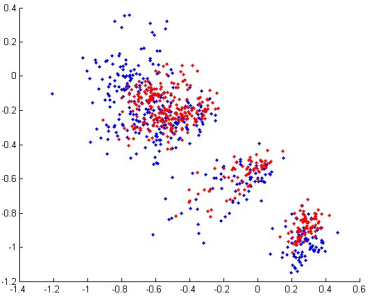

Fig. 6d. The overlapping diagram of $\mathcal{D}_{p1}$ and $\mathcal{D}_{p2}$

Fig. 6. Diagrams of data set *auto-mpg* projected by HOV^3 in MATLAB

3.4 Predictive Cluster Exploration by HOV^3 with Statistical Measurements

Many statistical measurements, such as mean, median, standard deviation and etc. can be directly introduced into HOV^3 as predictions to explore data distributions. In fact, prediction based on statistical measurements is more purposefully cluster exploration, and give an easier geometrical interpretation of the data distribution.

We use the Iris dataset as an example. As shown in Fig. 3, by random axis scaling, the user can divide the Iris data in 3 groups. This example exhibits that cluster exploration based on random adjustment may expose data groping information, but sometimes, it is hard to interpret such groupings.

We employ the 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. 7, where 3 groups clearly exist. It can be observed in Fig 7 that, 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. Intuitively, they have been wrongly clustered. We re-clustered them by their distributions, as shown in Fig 8.

The contrast of 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 see that the

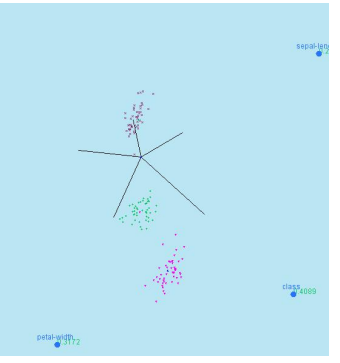

Fig. 7. Data distribution projected by HOV^3 in iVIBRATE of Iris with cluster indices maked by K-means

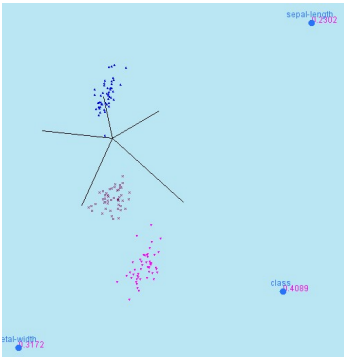

Fig. 8. Data distribution projected by HOV^3 in iVIBRATE of Iris with the new clustering indices by the user's intuition

quality of the new clustering result of Iris is better than that obtained by K-means according to their "*Variance*" comparison. 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, we have found that they are distinguished by the "*class*" attribute of Iris, i.e. *Iris-setosa*, *Iris-versicolor and Iris-virginica*. The cluster 1 generated by K-means is an outlier.

Table 1. The statistics of the clusters in Iris' produced by HOV^3 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*

With the statistical predictions in HOV^3 the user may even expose the cluster clues that are not easy to be found by random adjustments. For example, we adopted the 8th row of *auto-mpg*'s covariance matrix as a predictive measure (0.04698, -0.07657, -0.06580, 0.00187, -0.05598, 0.01343, 0.02202, 0.16102) to project *auto-mpg* by HOV^3 in MATLAB. The result is shown in Fig 9. We grouped them by their distribution as in Fig 10. Table 2 (right part) reports the statistics of the clusters generated by the projection of HOV^3, and reveals that the points in each cluster have very high similarity.

As we chose the 8th row of *auto-mpg*'s covariance matrix as the prediction, the result mainly depends on the 8th column of *auto-mpg* data, i.e., "origin" (country). Fig. 10 shows that C1, C2 and C3 are closer because they have the same "origin" value 1. The more detailed formation of clusters is given in the right part of Table 2.We believe that a domain expert could give a better and intuitive explanation about this clustering.

Then we chose number 5 to cluster *auto-mpg* by the K-means. Its clustering result is presented in the left part of Table 2. Comparing their corresponding statistics, we can see that according to the *Variance* of clusters, the quality of the clustering result by

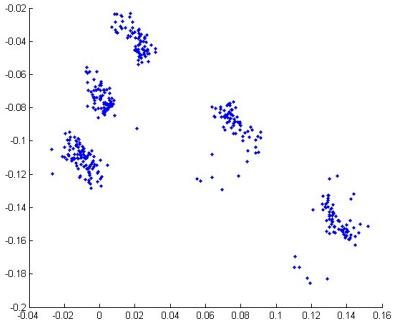

Fig. 9. Data distribution projected by HOV^3 in MATLAB of auto-mpg with 8th row of auto-map's covariance matrix as prediction

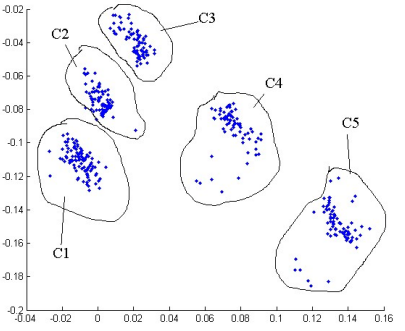

Fig. 10. Clustered distribution of data in Fig. 8 by the user's intuition

Table 2. The statistical contrast of clusters in *auto-mpg* produced by K-means and HOV^3

	Clusters produced by K-means (k=5)				Clusters generated by the user intuition on the data distribution					
C	%	Radius	Variance	MaxDis	Origin	Cylinders	%	Radius	Variance	MaxDis
1	0.503	681.231	963.406	1362.462	1	8	25.879	4129.492	0.130	4129.768
2	18.090	2649.108	0.206	2649.414	1	6	18.583	3222.493	0.098	3222.720
3	16.080	2492.388	0.139	2492.595	1	4	18.090	2441.881	0.090	2442.061
4	21.608	3048.532	0.207	3048.897	2	4	17.588	2427.449	0.142	2427.632
5	25.377	3873.052	0.220	3873.670	3	3	19.849	2225.465	0.093	2225.658
6	18.593	2417.804	0.148	2417.990						

HOV^3 with covariance prediction of *auto-mpg* is better than that one produced by K-means (k=5, cluster 1 produced by K-means is an outlier).

3.5 Predictive Cluster Validation by HOV^3

In practice, with extremely large sized datasets, it is infeasible to cluster an entire data set within an acceptable time scale. A common solution used in data mining is that, clustering algorithms are first applied to the training (a sampling) subset of data from a database to extract cluster patterns, and then the cluster scheme is assessed to see whether it is suitable for other subsets in the database. This procedure is regarded as *external cluster validation* [21]. Due to the high computational cost of statistical methods on assessing the consistency of cluster structures between large sized subsets, to achieve this goal by statistical methods is still a challenge in data mining.

Based on the assumption that if two same-sized data sets have a similar cluster structure, by applying a linear transformation to the data sets, the similarity of the newly produced distributions of the two sets would still be high, Zhang *et al* proposed a visual external validation approach by HOV^3 [23]. Technically, their approach uses a clustered subset and a same-sized unclustered subset from a database as the observation by applying the measure vectors that can separate clusters in the clustered subset by HOV^3. Thus each cluster and its geometrically covered data points (called *quasi-Cluster* in their approach) are selected. Finally, the overlapping rate of each

cluster-quasicluster pair is calculated; and if the overlapping rate approaches 1, this means that the two subsets have a similar cluster distribution.

Compared with the statistics-based validation methods, their method is not only visually intuitive, but also more effective in real applications [23]. As mentioned above, sometimes, it is time consuming to separate clusters manually in Star Coordinates or iVIBRATE. Thus, separation of clusters from lots of overlapping points is an aim of this research. As we described above, the approaches such as M-HOV^3 and HOV^3 with statistical measurement can be introduced into external cluster validation by HOV^3. In principle, any linear transformation can be employed into HOV^3 if it can separate clusters well.

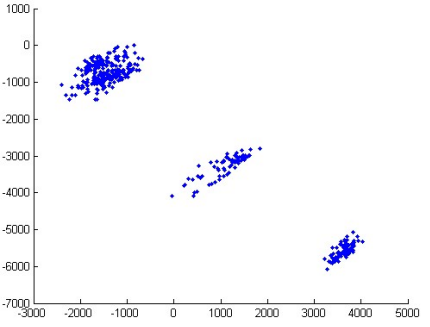

Fig. 11. The data distribution of auto-mpg projected by HOV3 with *cos(M*10i*) as the prediction

We therefore introduce the complex linear transformation to this process. We again use *auto-mpg* data set as an example. As shown in Fig. 6b, three roughly separated clusters appear there, where the vector M=[0.10, 0, 0.25, 0.2, 0.8, 0.85, 0.1, 0.95] was obtained from the axes values. Then we adopt cos($M \cdot 10i$) as a prediction, where i is the imaginary unit. The projection of HOV^3 with cos($M \cdot 10i$) is illustrated in Fig. 11, where three clusters are separated very well. In the same way, many other linear transformations can be applied to different datasets to obtain well-separated clusters. With the fully separated clusters, there will be marked improvement of the efficiency of visual cluster validation.

4 Related Work

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

For instance, Multidimensional Scaling, MDS [14] and Principal Component Analysis, PCA [10] are two commonly used 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 clusters in "Gaussian bumps", but its non-linear time complexity makes it neither suitable for dealing with very large data sets, nor for providing the contrast between clustering results. H-BLOB visualizes clusters into blob manners in a 3D hierarchical structure [17]. 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.

Kaski *el. al* [13] uses Self-organizing maps (SOM) to project high-dimensional data sets to 2D space for matching visual models [12]. 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* [7, 8] proposed the approaches based on FastMap [5] 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 [18].

5 Conclusions

In this paper, we have proposed a prediction-based visual approach to explore and verify clusters. This approach uses the HOV^3 projection technique and quantifies the previously obtained knowledge and statistical measurements about a high dimensional data set as predictions, so that users can utilize the predictions to project the data on 2D plane in order to investigate grouping clues or verify the validity of clusters based on the distribution of the data. This approach not only inherits the intuitive and easy understanding features of visualization, but also avoids the weaknesses of randomness and arbitrary exploration of the existing visual methods employed in data mining.

As a consequence, with the advantage of the quantified predictive measurement of this approach, users can identify the cluster number in the pre-processing stage of clustering efficiently, and also can intuitively verify the validity of clusters in the post-processing stage of clustering.

References

1. Ankerst, M., Breunig, M.M., Kriegel, S.H.P.J.: OPTICS: Ordering points to identify the clustering structure. In: Proc. of ACM SIGMOD Conference, pp. 49–60. ACM Press, New York (1999)
2. Ankerst, M., Keim, D.: Visual Data Mining and Exploration of Large Databases. In: 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany (September 2001)
3. Berkhin, P.: A Survey of Clustering Data Mining Techniques. In: Jacob, K., Charles, N., Marc, T. (eds.) Grouping Multidimensional Data, pp. 25–72. Springer, Heidelberg (2006)
4. Chen, K., Liu, L.: iVIBRATE: Interactive visualization-based framework for clustering large datasets. ACM Transactions on Information Systems (TOIS) 24(2), 245–294 (2006)
5. 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)
6. Fleming, W.: Functions of Several Variables. In: Gehring, F.W., Halmos, P.R. (eds.) 2nd edn. Springer, Heidelberg (1977)
7. 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)
8. Huang, Z., Lin, T.: A visual method of cluster validation with Fastmap. In: Terano, T., Chen, A.L.P. (eds.) PAKDD 2000. LNCS, vol. 1805, pp. 153–164. Springer, Heidelberg (2000)

9. Jain, A., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 31(3), 264–323 (1999)
10. Jolliffe Ian, T.: Principal Component Analysis. Springer Press, Heidelberg (2002)
11. Kandogan, E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: Proc. of ACM SIGKDD Conference, pp. 107–116. ACM Press, New York (2001)
12. Kohonen, T.: Self-Organizing Maps, 2nd extended edn. Springer, Berlin (1997)
13. Kaski, S., Sinkkonen, J., Peltonen, J.: Data Visualization and Analysis with Self-Organizing Maps in Learning Metrics. In: Kambayashi, Y., Winiwarter, W., Arikawa, M. (eds.) DaWaK 2001. LNCS, vol. 2114, pp. 162–173. Springer, Heidelberg (2001)
14. 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)
15. Oliveira, M.C., Levkowitz, H.: From Visual Data Exploration to Visual Data Mining: A Survey. IEEE Transaction on Visualization and Computer Graphs 9(3), 378–394 (2003)
16. Pampalk, E., Goebl, W., Widmer, G.: Visualizing Changes in the Structure of Data for Exploratory Feature Selection. In: SIGKDD '03, Washington, DC, USA (2003)
17. 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)
18. Seo, J., Shneiderman, B.: From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments. In: Hemmje, M., Niederée, C., Risse, T. (eds.) From Integrated Publication and Information Systems to Information and Knowledge Environments. LNCS, vol. 3379, Springer, Heidelberg (2005)
19. Shneiderman, B.: Inventing Discovery Tools: Combining Information Visualization with Data Mining. In: Jantke, K.P., Shinohara, A. (eds.) DS 2001. LNCS (LNAI), vol. 2226, pp. 17–28. Springer, Heidelberg (2001)
20. Weiss, S.M., Indurkhya, N.: Predictive Data Mining: A Practical Guide. Morgan Kaufmann Publishers, San Francisco (1998)
21. Vilalta, R., Stepinski, T., Achari, M.: An Efficient Approach to External Cluster Assessment with an Application to Martian Topography, Technical Report, No. UH-CS-05-08, Department of Computer Science, University of Houston (2005)
22. Zhang, K-B., Orgun, M.A., Zhang, K.: HOV^3, An Approach for Cluster Analysis. In: Li, X., Zaïane, O.R., Li, Z. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 317–328. Springer, Heidelberg (2006)
23. Zhang, K-B., Orgun, M.A., Zhang, K.: A Visual Approach for External Cluster Validation. In: Proc. of IEEE Symposium on Computational Intelligence and Data Mining (CIDM2007), Honolulu, Hawaii, USA, April 1-5, 2007, pp. 576–582. IEEE Press, Los Alamitos (2007)