Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

A Visual Approach for External Cluster Validation

Ke-Bing Zhang, Mehmet A. Orgun, *Senior Member, IEEE*, and Kang Zhang, *Senior Member, IEEE*

Abstract—**Visualization can be very powerful in revealing cluster structures. However, directly using visualization techniques to verify the validity of clustering results is still a challenge. This is due to the fact that visual representation lacks precision in contrasting clustering results. To remedy this problem, in this paper we propose a novel approach, which employs a visualization technique called HOV^3 (*Hypothesis Oriented Verification and Validation by Visualization*) which offers a tunable measure mechanism to project clustered subsets and non-clustered subsets from a multidimensional space to a 2D plane. By comparing the data distributions of the subsets, users not only have an intuitive visual evaluation but also have a precise evaluation on the consistency of cluster structure by calculating geometrical information of their data distributions.**

I. Introduction

THE goal of clustering is to distinguish objects into partitions/clusters based on given criteria. A large number of clustering algorithms have been developed for different application purposes [8, 14, 15]. However, due to the memory limitation of computers and the extremely large sized databases, in practice, it is infeasible to cluster entire data sets at the same time. Thus, applying clustering algorithms to sampling data to extract hidden patterns is a commonly used approach in data mining [5]. As a consequence of sampling data cluster analysis, the goal of external cluster validation is to evaluate a well-suited cluster scheme learnt from a subset of a database to see whether it is suitable for other subsets in the database. In real applications, achieving this task is still a challenge. This is not only due to the high computational cost of statistical methods for assessing the robustness of cluster structures between the subsets of a large database, but also due the non-linear time complexity of most existing clustering algorithms**.**

Visualization provides users an intuitive interpretation of cluster structures. It has been shown that visualization allows for verification of the clustering results [10]. However, the direct use of visualization techniques to evaluate the quality of clustering results has not attracted enough attention in the data mining community. This might be due to the fact that visual representation lacks precision in contrasting clustering results.

We have proposed an approach called HOV^3 to detect cluster structures [28]. In this paper, we discuss its projection mechanism to support external cluster validation. Our approach is based on the assumption that by using a measure to project the data sets in the same cluster structure, the similarity of their data distributions should be high. By comparing the distributions produced by applying the same measures to a clustered subset and other non-clustered subsets of a database by HOV^3, users can investigate the consistency of cluster structures between them both in visual form and in numerical calculation.

The rest of this paper is organized as follows. Section 2 briefly introduces ideas of cluster validation (with more details of external cluster validation) and visual cluster validation. A review of related work on cluster validation by visualization, and a more detailed account of HOV^3 are presented in Section 3. Section 4 describes our idea on verifying the consistency of cluster structure by a distribution matching based method in HOV^3. Section 5 demonstrates the application of our approach on several well-known data sets. Finally, section 6 summarizes the contributions of this paper.

II. Background

A. Cluster Validation

Cluster validation is a procedure of assessing the quality of clustering results and finding a cluster strategy fit for a specific application. It aims at finding the optimal cluster scheme and interpreting the cluster patterns. In general, cluster validation approaches are classified into the following three categories [9, 15, 27].

Internal approaches: they assess the clustering results by applying an algorithm with different parameters on a data set for finding the optimal solution [1];

Relative approaches: the idea of relative assessment is based on the evaluation of a clustering structure by comparing it to other clustering schemes [8]; and

External approaches: the external assessment of a clustering approach is based on the idea that there exists known priori clustered indices produced by a clustering algorithm, and then assessing the consistency of the clustering structures generated by applying the clustering algorithm to different data sets [12].

Manuscript received October 31, 2006.

K-B. Zhang is with the Department of Computing, Macquarie University, Sydney, NSW 2109, Australia (phone: 612-9850-9590; fax: 612-9850-9551; e-mail: kebing@ics.mq.edu.au).

M.A. Orgun is with the Department of Computing, Macquarie University, Sydney, NSW 2109, Australia (e-mail: mehmet@ics.mq.edu.au).

K. Zhang is with the Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688, USA (e-mail: kzhang@utdallas.edu).

1-4244-0705-2/07/$20.00 ©2007 IEEE

B. External Cluster Validation

As a necessary post-processing step, external cluster validation is a procedure of hypothesis test, i.e., given a set of class labels produced by a cluster scheme, compare it with the clustering results by applying the same cluster scheme to the other partitions of a database, as shown in the Fig. 1.

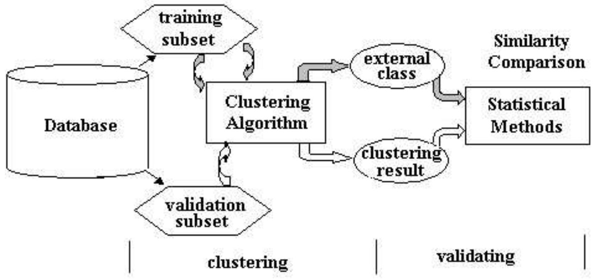


Fig. 1. External criteria based validation

The statistical methods for quality assessment are employed in external cluster validation, such as Rand statistic [24], Jaccard Coefficient [7], Folkes and Mallows index [21], Huberts Γ statistic and Normalized Γ statistic [27], and Monte Carlo method [20], to measure the similarity between the priori modeled partitions and clustering results of a dataset. However, achieving these tasks is time consuming when the database is large, due to the drawback of high computational cost of statistics-based methods for assessing the consistency of cluster structure between the sampling subsets. Recent surveys on cluster validation methods can be found in the literatures [10, 12, 27].

C. Visual Cluster Validation

In high dimensional space, traditional clustering algorithms tend to break down in terms of efficiency as well as accuracy because data do not cluster well anymore [3]. Thus, introducing 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 [23]. Visual presentations can be very powerful in revealing trends, highlighting outliers, showing clusters, and exposing gaps in data [26].

Visual cluster validation is a combination of information visualization and cluster validation techniques. In the cluster analysis process, visualization provides analysts with intuitive feedback on data distribution and supports decision-making activities.

III. Related Work

A. Previous Works

A large number of clustering algorithms have been developed, but only a small number of cluster visualization tools are available to facilitate researchers' understanding of the clustering results [25]. Several efforts have been made in cluster validation with visualization [2, 4, 11, 13, 16, 18]. While, these techniques tend to help users have intuitive comparisons and better understanding of cluster structures, but they do not focus on assessing the quality of clusters.

For example, OPTICS [2] uses a density-based technique to detect cluster structures and visualizes them in "Gaussian bumps", but its non-linear time complexity makes it neither suitable to deal with very large data sets, nor suitable to provide the contrast between clustering results. Kaski *el. al* [18] imposes the technique of Self-organizing maps (SOM) technique to project multidimensional data sets on a 2D space for matching visual models [17]. However, the SOM technique is based on a single projection strategy and not powerful enough to discover all the interesting features from the original data. Huang *et. al* [11, 13] proposed approaches based on FastMap [5] to assist users on identifying and verifying the validity of clusters in visual form. Their techniques are good on cluster identification, but are not able to evaluate the cluster quality very well.

The most prominent feature of techniques based on Star Coordinates, such as VISTA [4] and HOV^3 [28], is their linear time computational complexity. This feature makes them suitable to be used as visual interpretation and detection tools in cluster analysis. However, the characteristic of imprecise qualitative analysis of Star Coordinates and VISTA limits them as quantitative analysis tools. In addition, VISTA adopts "landmark" points as representatives from a clustered subset and re-samples them to deal with cluster validation [4]. But its experience-based "landmark" point selection does not always handle the scalability of data very well, due to the fact that well-representative landmark points selected in a subset may fail in other subsets of a database.

Visualization techniques used in data mining and cluster analysis are surveyed in the literatures [22, 25].

B. Star Coordinates

The approach HOV^3 employed in this research was inspired from the Star Coordinates [16]. For better understanding our work, we briefly describe it here.

Star Coordinates utilizes a point on a 2D surface to represent a set of points of n-dimensional data. The values of n-dimensional coordinates are mapped to the orthogonal coordinates X and Y, as shown in Fig. 2.

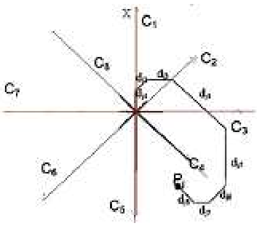


Fig. 2. Positioning a point by an 8-attribute vector in Star Coordinates [16]

The mapping from n-dimensional Star Coordinates to 2D X-Y coordinates is calculated as in Formula (1).

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

where p_j (x, y) is the normalized location of $D_j=(d_{j1}, d_{j2}, \ldots, d_{jm})$, and d_{ji} is the value of *j*th record of a data set on the *i*th coordinate C_i in Star Coordinates space; $\bar{u}_{xi}\cdot(d_{ji}-\min_i)$ and $\bar{u}_{yi}\cdot(d_{ji}-\min_i)$ are unit vectors of d_{ji} mapping to X direction and Y direction, $\min_i=\min(d_{ji}, 0\leq j<m)$ $\max_i=\max(d_{ji}, 0\leq j<m)$ are minimum and maximum values of *i*th dimension respectively; and *m* is the number of records in the data set.

C. *HOV³ Model*

The idea of HOV³ is based on hypothesis test by visualization. It treats hypotheses as measures to reveal the difference between hypotheses and real performance by projecting the test data against the measures [28]. Geometrically, the difference of a matrix D_j and a vector M can be represented by their inner product, D_j•M. Let Dj=($d_{j1}, d_{j2}, \ldots, d_{jm}$) be a data set with n attributes, and M=($m_1, m_2, \ldots, m_n$). The inner product of each vector d_{ji}, (i =1, …, n) of Dj with M can be seen as a mapping from an n-dimensional data set to one measure F: $R^n \to R^2$. It is written as:

$$< d_{ji,} M >= m_1 d_{j1} + m_2 d_{j2} + \ldots + m_n d_{jn} = \sum_{k=1}^{n} m_k d_{jk} \tag{2}$$

In order to enlarge the data analysis space, we introduce the complex number system into our study. Let $z = x + i.y$, where i is the imaginary unit. According to the Euler formula, we have: $e^{ix} = \cos x + i\sin x$. Let $z_0 = e^{2\pi i/n}$; we see that z_0^1, z_0^2, z_0^3,…, 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 the formula (1) can be simply written as:

$$P_j(z_0) = \sum_{k=1}^{n}\left[(d_{jk} - \min d_k)/(\max d_k - \min d_k)\cdot z_0^k\right] \tag{3}$$

Where, $\min_k d_{jk}$ and $\max_k d_{jk}$ represents the minimal and the maximal values of the *k*th coordinate respectively. This is the case of equally-divided circle surface. Then the more general form can be defined as:

$$P_j(z_0) = \sum_{k=1}^{n}\left[(d_{jk} - \min d_k)/(\max d_k - \min d_k)\cdot z_K\right] \tag{4}$$

where $z_k = e^{i\theta k}$; θ is the angle of neighbouring axes; and $\sum_{k=1}^{n}\theta_k = 2\pi$. In any case equation (4) can be viewed as mappings from $R^n \to C^2$.

Given a non-zero measure vector ***m*** in R^n, and a family of vectors P_j, and the projections of P_j against ***m*** according to formulas (2) and (4), the HOV³ model is given as the following equation:

$$P_j(z_0) = \sum_{k=1}^{n}\left[(d_{jk} - \min d_k)/(\max d_k - \min d_k)\cdot z_K \cdot m_k\right] \tag{5}$$

where m_k is the *k*th attribute of measure ***m***.

As shown above, a hypothesis in HOV³ is a quantified measure vector. Thus HOV³ is also able to detect the consistency of cluster structures among the subsets of a database by comparing their data distributions, because cluster validation procedure is primarily a hypothesis test process.

D. *The Axis Tuning Feature*

Overlapping and ambiguities are inevitably introduced by projecting multidimensional data into 2D space. For mitigating the problem, Star Coordinates provides several visual adjustment mechanisms, such as axis scaling, axes angles rotation; coloring data points, etc [15]. We use Iris, a well-known data set in machine learning research, as an example to demonstrate the feature of axis scaling of techniques based on Star Coordinates as follows.

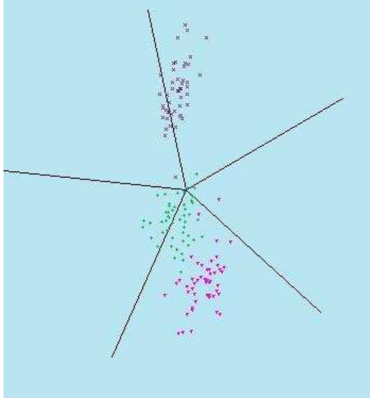


Fig. 3. The initial data distribution of clusters of Iris produced by k-means in VISTA.

Iris has 4 numeric attributes and 150 instances. We first applied K-means clustering algorithm to it and obtained 3 clusters (k=3,here), and then tuned the weight value of each axis (called *α-adjustment* in VISTA) of Iris in VISTA [4]. Fig.3 shows the original data distribution of Iris, which has overlapping among the clusters. A well-separated distribution of Iris is illustrated in Fig. 4 by a series of axis scaling. The clusters are much easier to recognize in Fig. 4 than those in the original one.

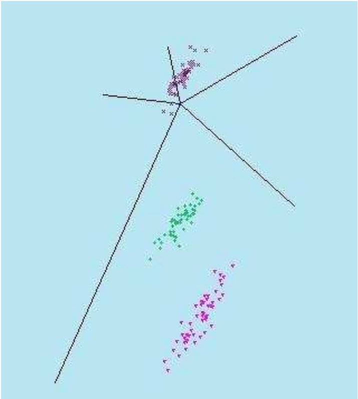


Fig. 4. The tuned version of the Iris data distribution in VISTA.

This axis-tuning feature is significant for our external cluster validation method based on distribution matching by HOV^3. We give the detailed explain next.

IV. Cluster Validation with HOV^3

The feature of tunable axis provides us a mechanism to quantitatively handle the external cluster validity by HOV^3. Our approach is based on the assumption that *by using a measure to project the data sets in the same cluster structure, the similarity of their data distributions should be high.* Based on this idea, we have implemented an approach for external cluster validation based on distribution matching by HOV^3.

A. Definitions

To explain our approach precisely, we first give a few definitions below.

Definition 1**:** A data projection from n-dimensional space to 2D plane by applying HOV^3 to a data set $\mathcal{P}$, as shown in formula (5), is denoted as $D_p=\mathcal{H}_C(\mathcal{P}, M)$, where

- $\mathcal{P}$ is an n-dimensional data set, and $\mathcal{P}=(p_1, p_2, ., p_m)$, p_j ($1\leq k \leq m$) is an instance of $\mathcal{P}$;
- $M=(w_{1t}, w_{2t}, ..., w_{nt})$, is a non-zero measure vector; w_{it} ($1\leq i \leq n$) is the weight value of *k*th coordinate at *t* moment in the Star Coordinates plane;
- D_p is the geometrical distribution of $\mathcal{P}$ in 2D space, $D_p=(p_1(x_1,y_1),\ p_2(x_2,y_2),\ ...,\ p_m(x_m,y_m))$, $p_j(x_i,y_i)$ is the location of p_j in X-Y Coordinates plane.

Definition 2: Let $\mathbb{D}$ be a database of data points. A cluster $C:=(D, L)$ is a non-empty set $D\subseteq\mathbb{D}$ on a label set L, and the *i*th cluster $C_i=\{p\in D, l\in L|\ \forall C_j.p: C_j.l=i \wedge i>0\}$ where l is the cluster label of p, $l\in\{-1, 0, 1...,k\}$, and k is the number of clusters.

As special case, an *outlier point* is an element of $\mathbb{D}$ and with cluster label –1; a *non-clustered element* of $\mathbb{D}$; has a cluster label of 0, i.e., it has not been clustered.

Definition 3: A *spy subset* $\mathcal{P}_s$ is a clustered subset of $\mathbb{D}$ produced by a clustering algorithm, where $\mathcal{P}_s=\{C_1,C_2,..., C_k, C_E\}$, C_i ($1\leq i \leq k$) is a cluster in $\mathcal{P}_s$; C_E is the outlier set of $\mathcal{P}_s$

A *spy subset* is used as a model to verify the cluster structure in the other partitions in the database $\mathbb{D}$.

Definition 4: A subset $\mathcal{P}_t\subseteq\mathbb{D}$ is a *target subset* of $\mathcal{P}_s$, $\mathcal{P}_t=\{P_t.p\in\mathbb{D}, P_t.l\in L\ |\forall P_t.p{:}P_t.l=0 \wedge |\mathcal{P}_s|=|\mathcal{P}_t|\}$.

A *target subset* $\mathcal{P}_t$ is a non-clustered subset of $\mathbb{D}$ and has the same size of a *spy subset* P_s of $\mathbb{D}$. It is used as a target to investigate the similarity of cluster structure with the *spy subset* $\mathcal{P}_s$.

Definition 5: A non-clustered point p_o is called an *overlapping point* of a cluster C_i, denoted as $C_i.\,p_o$ iff ($\exists p\in C_i \wedge p_o \notin C_i \wedge |p_o - p|\leq\delta$), where δ is the threshold distance given by the user.

Definition 6: The *overlapping point* set of cluster C_i is composed as a *quasi-cluster* of C_i, denoted as C_{qi} i.e., $\{p_o\in C_{qi}\ |\forall\ C_i.\,p_o\}$

All overlapping points of C_i are composed a *quasi-cluster* C_{qi} of $C_{i.}$

Definition 7: A cluster C_i is called a *well-separated cluster* visually, when it satisfies the condition that ($C_i\subseteq\mathcal{P}_s$, $C_j\subseteq\mathcal{P}_s|\ \forall p\in C_i$: $p\neq C_j.p_o \wedge i\neq j$).

A well-separated cluster Ci in the spy subset implies that no points in Ci are within the threshold distance δ to any other clusters in the spy subset.

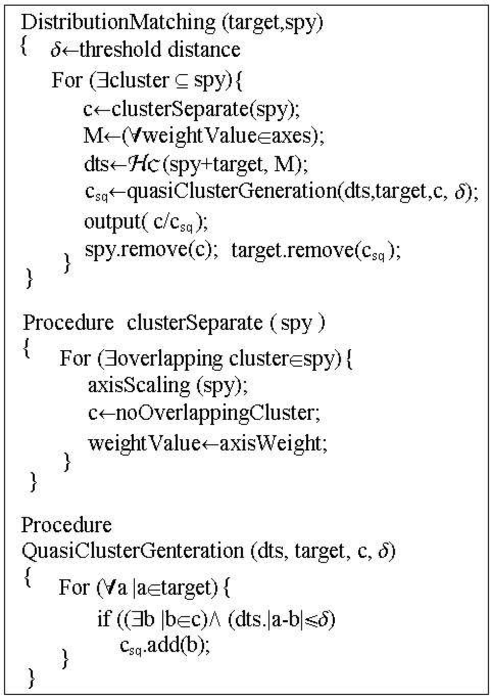
Based on above the definitions, we present the application of our approach to external cluster validation based on distribution matching by HOV^3 as follows.

B. The Stages of Our Approach

The stages of the application of our approach are summarized in the following steps:

1. ***Clustering***
   First, the user applies a clustering algorithm to a randomly selected subset $\mathcal{P}_s$ from the given dataset $\mathcal{D}$.
2. ***Cluster Separation***
   The clustering result of $\mathcal{P}_s$ is introduced and visualized in HOV^3. Then the user manually tunes the weight value of each axis to separate overlapping clusters. If one or more cluster(s) are separated from the others visually, then the weight values of each axis are recorded as a measure vector *M*.
3. ***Data Projection by HOV^3***
   The user samples another observation with the number of points as in $\mathcal{P}_s$ as a *target subset* $\mathcal{P}_t$. The clustered subset $\mathcal{P}_s$ (now as a *spy subset*) and its target subset $\mathcal{P}_t$ are projected together by HOV^3 with vector *M* to detect the distribution consistency between $\mathcal{P}_s$ and $\mathcal{P}_t$.
4. ***The Generation of Quasi-Clusters***
   The user gives a threshold δ, and then according to Definitions 5, 6 and 7, a *quasi-cluster* C_{qi} of a separated cluster C_i is computed. Then C_{qi} is removed from $\mathcal{P}_{t,}$ and C_i is removed from $\mathcal{P}_{s.}$ If $\mathcal{P}_s$ has clusters then we go back to step 2, otherwise we proceed to the next step.
5. ***The Interpretation of Result***
   The overlapping rate of each cluster-and-quasi-cluster pair is calculated as (C_{qi}, C_i)=$|C_{qi}|/|C_i|$. If the overlapping rate approaches 1, cluster C_i and its quasi-clusters C_{qi} have high similarity, since the amount ratio of the spy subset and the target subset is 1:1. Thus the overlapping analysis is simply transformed into a linear regression analysis, i.e., the points around the line $C=C_q$.

Corresponding to the procedure mentioned above, we give the algorithm of external cluster validation based on distribution matching by HOV³ below, in Fig. 5.



```
DistributionMatching (target,spy)
{   δ←threshold distance
    For (∃cluster ⊆ spy){
        c←clusterSeparate(spy);
        M←(∀weightValue∈axes);
        dts←$\mathcal{H}_C$(spy+target, M);
        $c_{sq}$←quasiClusterGeneration(dts,target,c, δ);
        output( c/$c_{sq}$);
        spy.remove(c);  target.remove($c_{sq}$);
    }
}

Procedure  clusterSeparate ( spy )
{   For (∃overlapping cluster∈spy){
        axisScaling (spy);
        c←noOverlappingCluster;
        weightValue←axisWeight;
    }
}

Procedure
QuasiClusterGenteration (dts, target, c, δ)
{   For (∀a |a∈target) {
        if ((∃b |b∈c)∧ (dts.|a-b|≤δ)
            $c_{sq}$.add(b);
    }
}
```

Fig 5. The algorithm of external cluster validation based on distribution matching in HOV³

In Fig. 5, the procedure *clusterSeparate* responds the user's axis tuning to separate the clusters in the spy subset, and to gather weight values of axes as a measure vector; the procedure *quasiClusterGeneration* produces quasi clusters in the target subset corresponding to the clusters in the spy subset.

C. Our Model

In contrast to statistics-based external cluster validation model illustrated in Fig. 2, we exhibit our model for external cluster validation by visualization in HOV³ in Fig. 6.

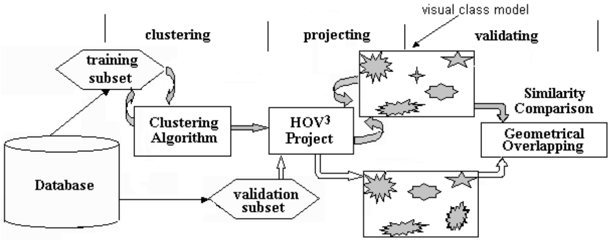


Fig. 6. External cluster validation by HOV³

Comparing these two models, we may observe that instead of using a clustering algorithm to cluster another sampling data sets, in our model, we use a clustered subset from a database as a model to verify the similarity of cluster structure between the model and the other non-clustered subsets from the database. To handle the scalability on resampling datasets, we choose the non-cluster observations with the same size as the clustered subset, and then project them together by HOV³. As a consequence, the user can easily utilize the well-separated clusters produced by scaling axes in HOV³ as a model to pick out their corresponding *quasi-clusters*, where points in a *quasi-cluster* overlap its corresponding cluster. Also, instead of using statistical methods to assess the similarity between the two subsets, we simply compute the overlapping rate between the clusters and their quasi-clusters to explore their consistency.

V. Examples and Explanation

In this section, we present several examples to demonstrate the advantages of the external cluster validation in HOV³. We have implemented our approach in MATLAB running under Windows 2000 Professional. The datasets used in the examples are obtained from the UCI machine learning website: http://www.ics.uci.edu/~mlearn/Machine-Learning. html.

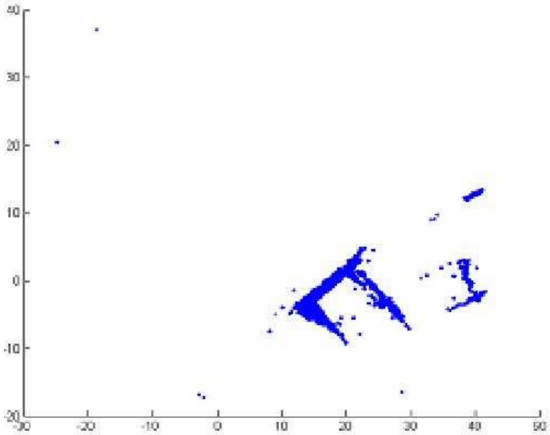


Fig. 7. The original data distribution of the first 5,000 data points of Shuttle in MATLAB by HOV³ (without cluster indices)

Shuttle data set has 9 attributes and 4,3500 instances. We choose the first 5,000 instances of Shuttle as a sampling data and apply the K-means algorithm [19] to it. Then we utilize the clustered result as a *spy subset*. We assumed that we have found the optimal cluster number k=5 for the sampling data. The original data distributions with and without cluster indices are illustrated in the diagrams of Fig.7 and Fig.8 respectively. It can been seen that there exists a cluster overlapping in Fig. 8.

To obtain well-separated clusters, we tuned the weight of each coordinate, and had a satisfied version of the data distribution as shown in Fig. 9. The weight values of axes are recorded as a measure vector [0.80, 0.55, 0.85, 0.0, 0.40, 0.95, 0.20, 0.05, 0.459], in this case. Then we chose the second 5,000 instances of Shuttle as a *target subset* and projected the target subset and the spy subset together against the measure vector by HOV³.

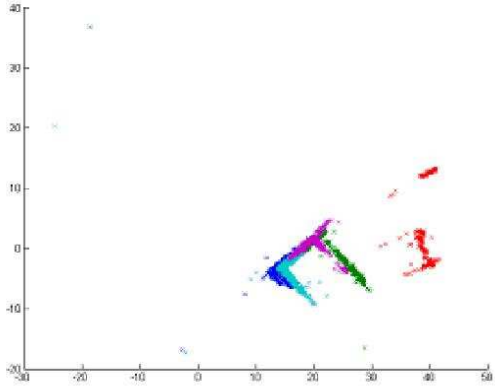


Fig. 8. The original data distribution of the first 5,000 data points of Shuttle in MATLAB by HOV^3 (with cluster indices)

Their distributions are presented in Fig. 10, where we may observe that their data distributions are matched very well. We chose the points in the enclosed area in Fig. 10 as a "cluster" then obtained a quasi-cluster in the target subset corresponding to the cluster in the enclosed area. In the same way, we can find the other quasi-clusters from the target subset.

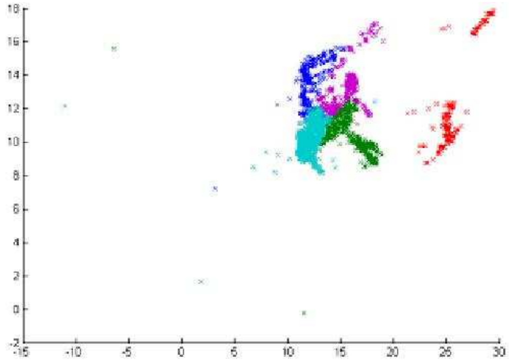


Fig. 9. A well-separated version of the spy subset distribution of Shuttle

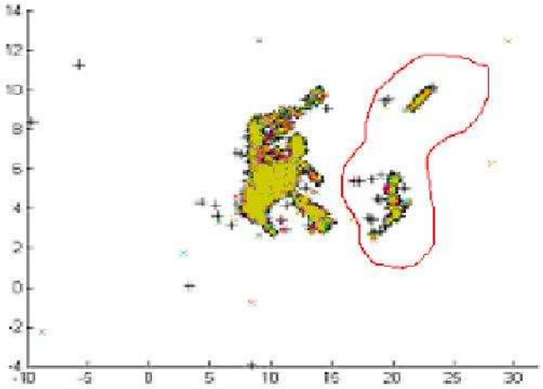


Fig. 10. The projection of the spy subset and a target subset of Shuttle by applying a measure vector.

We have done the same experiment on 4 target subsets of Shuttle. The size of each quasi-cluster and its corresponding cluster are listed in Table 1, and their curves of linear regression to the line C=C$_q$ are illustrated in Fig. 11.

TABLE I

CLUSTERS AND THEIR CORRESPONDING QUASI-CLUSTERS

Subset	C_{q1}/C_1	C_{q2}/C_2	C_{q3}/C_3	C_{q4}/C_4	C_{q5}/C_5
Spy	318	773	513	2254	1142
Target 1	278/318 =0.8742	670/773 =0.8668	503/513 =0.9805	2459/2254 =1.0909	1123/1142 =0.9834
Target 2	279/318 =0.8773	897/773 =1.1604	626/513 =1.2203	2048/2254 =0.9086	1602/1142 =1.4028
Target 3	280/318 =0.8805	875/773 =1.1320	481/513 =0.9376	2093/2254 =0.9286	1455/1142 =1.2741
Target 4	261/318 =0.8208	713/773 =0.9224	368/513 =0.7173	2416/2254 =1.0719	1169/1142 =1.0264

*At current stage, we collect the quasi-clusters manually, thus C_{qi} here may have redundancy and misloading.

It is observed that the curves are well matched to the line C=C$_q$, i.e. the overlapping rate between the clusters and their quasi-clusters are high. The standard deviation is a good way to reflect the difference between the two vectors. Thus we have calculated the standard deviation of each C_{qi}-C_i pairs among the target$_k$ (k=1,..4) and the spy subsets. They are 0.0826, 0.1975, 0.1491 and 0.1304. This means that the similarity of cluster structure in the spy and the target subsets is high.

In summary, the experiments show that the same cluster structure in the spy subset of Shuttle also exists in the target subsets of Shuttle.

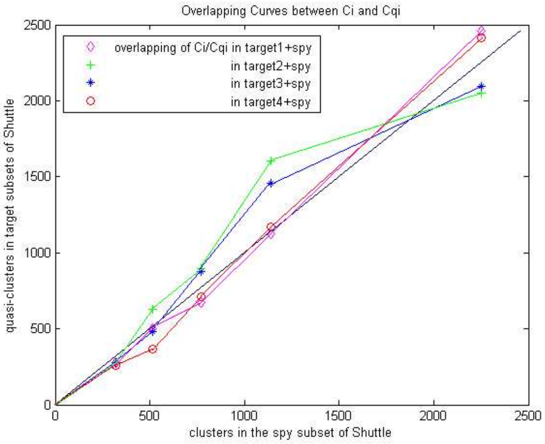


Fig. 11. The curves of linear regression to the line C=C$_q$.

In these experiments, we have also measured the timing for both clustering and projection in MATLAB. The results are listed in the Table 2.

TABLE 2

TIMING OF CLUSTERING AND PROJECTING

Clustering by K-mens (k=5)			Projecting by HOV^3		
Subset	Amount	Time (Second)	Subset	Size	Time (Second)
Target 1	5,000	.532	Syp+Target 1	10,000	.11
Target 2	5,000	.61	Syp+Target 2	10,000	.109
Target 3	5,000	.656	Syp+Target 3	10,000	.11
Target 4	5,000	.453	Syp+Target 4	10,000	.109

Based on this calculation, it has been observed that the projection by HOV^3 is much faster than the clustering process by the K-means algorithm. It is particularly effective for verifying the clustering results within extremely huge databases. Although the cluster separation in our approach may incur some time, once the well-separated clusters are found, using a measure vector to project a huge data set will be a lot more efficient than re-applying a clustering algorithm to the data set.

VI. CONCLUDING REMARKS

In this paper we have proposed a novel visual approach to assist users to verify the validity of any cluster scheme, i.e., an approach based on distribution matching for external cluster validation by visualization. The HOV^3 visualization technique has been employed in our approach, which uses measure vectors to project a data set and allows the user to iteratively adjust the measures for optimizing the result of clusters.

By comparing the data distributions of a clustered subset and non-clustered subsets projected by HOV^3 with tunable measures, users can performance intuitive visual evaluation, and also have a precise evaluation on the consistency of the cluster structure by performing geometrical computation on their data distributions as well. By comparing our approach with existing visual methods, we have observed that our method is not only efficient in performance, but also effective in real applications.

REFERENCES

[1] A. L. Abul, R. Alhajj, F. Polat and K. Barker "Cluster Validity Analysis Using Subsampling," in *proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, Washington DC, Oct. 2003 Volume 2: pp. 1435-1440.

[2] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J.Sander, "OPTICS: Ordering points to identify the clustering structure", in *proceedings of ACM SIGMOD Conference,* 1999 pp. 49-60.

[3] C. Baumgartner, C. Plant, K. Railing, H-P. Kriegel, P. Kroger, "Subspace Selection for Clustering High-Dimensional Data", Proc. of the Fourth IEEE International Conference on Data Mining (ICDM'04), 2004, pp.11-18.

[4] K. Chen and L. Liu,."VISTA: Validating and Refining Clusters via Visualization", *Journal of Information Visualization.* Volume3 (4), 2004, pp. 257-270.

[5] E. Clifford, "*Data Analysis by Resampling: Concepts and Applications"*, Duxbury Press, 2000.

[6] C. Faloutsos and K. Lin, "Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia data sets" *Proc. of ACM-SIGMOD*, 1995 pp.163- 174.

[7] Jaccard, S. (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44**,** 223–270.

[8] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*," Morgan Kaufmann Publishers, 2001.

[9] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On Clustering Validation Techniques" *Journal of Intelligent Information Systems,* Volume 17 (2/3), 2001, pp. 107–145.

[10] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "Cluster validity methods: Part I and II", *SIGMOD* Record, 31, 2002.

[11] Z. Huang, D. W. Cheung and M. K. Ng, "An Empirical Study on the Visual Cluster Validation Method with Fastmap", *Proceedings of DASFAA01*, Hong Kong, April 2001, pp.84-91.

[12] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis", *Journal of Bioinformatics* Volume 21(15), 2005, pp. 3201-3212.

[13] Z. Huang and T. Lin, "A visual method of cluster validation with Fastmap", *Proc. of PAKDD-2000*, 2000 pp. 153- 164.

[14] A. K. Jain and R. C. Dubes, "*Algorithms for Clustering Data*", Prentice Hall,1988

[15] A. Jain, M. N. Murty and P. J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, Volume 31(3), 1999, pp. 264-323.

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

[17] T. Kohonen, "*Self-Organizing Maps*" Springer, Berlin, second extended edition,1997.

[18] S. Kaski, J. Sinkkonen. and J. Peltonen, "Data Visualization and Analysis with Self-Organizing Maps in Learning Metrics", *DaWaK 2001*, LNCS 2114, 2001, pp.162-173.

[19] J. McQueen, "Some methods for classification and analysis of multivariate observations", *Proc. of 5th Berkeley Symposium on Mathematics, Statistics and Probability*, Volume 1, 1967, pp. 281-298.

[20] G. W. Milligan, "A Review Of Monte Carlo Tests Of Cluster Analysis", *Journal of* Multivariate Behavioral Research Vol. 16(3), 1981, pp. 379-407.

[21] G.W. Milligan, L.M. Sokol, & S.C. Soon "The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure", IEEE Trans PAMI, 1983 5(1):40-47.

[22] F. Oliveira, H. Levkowitz, "From Visual Data Exploration to Visual Data Mining: A Survey", *IEEE Trans.Vis.Comput. Graph*, Volume 9(3), 2003, pp.378-394.

[23] E. Pampalk, W. Goebl, and G. Widmer, "Visualizing Changes in the Structure of Data forExploratory Feature Selection", SIGKDD '03, August 24-27, 2003, Washington, DC, USA

[24] Rand, W.M., Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.*, 66:846-850, 1971.

[25] J. Seo and B. Shneiderman, "*From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*", Essays Dedicated to Erich J. Neuhold on the Occasion of His 65th Birthday. Lecture Notes in Computer Science Volume 3379, Springer, 2005.

[26] B Shneiderman, "Inventing Discovery Tools: Combining Information Visualization with Data Mining, *Proc. of Discovery Science 2001,*Lecture Notes in Computer Science Volume 2226, 2001, pp.17-28.

[27] S. Theodoridis and K. Koutroubas, "*Pattern Recognition*", Academic Press. 1999.

[28] K-B. Zhang, M. A. Orgun and K. Zhang, "HOV^3, An Approach for Cluster Analysis", *Proc. of ADMA 2006,* XiAn, China, Lecture Notes in Computer Science series, Volume. 4093, 2006, pp317-328.