An Empirical Study on the Visual Cluster Validation Method with Fastmap

Zhexue Huang, David W. Cheung
E-Business Technology Institute
The University of Hong Kong
jhuang,dcheung@eti.hku.hk

Michael K. Ng*
Department of Mathematics
The University of Hong Kong
mng@maths.hku.hk

Abstract

This paper presents an empirical study on the visual method for cluster validation based on the Fastmap projection. The visual cluster validation method attempts to tackle two clustering problems in data mining: (1) to verify partitions of data created by a clustering algorithm and (2) to identify genuine clusters from data partitions. They are achieved through projecting objects and clusters by Fastmap to the 2D space and visually examining the results by humans. A Monte Carlo evaluation of the visual method was conducted. The validation results of the visual method were compared with the results of two internal statistical cluster validation indices, which shows that the visual method is in consistence with the statistical validation methods. This indicates that the visual cluster validation method is indeed effective and applicable to data mining applications.

1 Introduction

Clustering data in real world databases is an important task in data mining applications. A typical example is customer segmentation. In database marketing, for example, a good segmentation scheme is a necessary condition for conducting effective marketing campaigns. In telecommunication service, customer segmentation is critical in identifying potential churners and deciding proper offers to retain them. However, clustering a large real world database is by no means a trivial task due to the size and complexity of data.

A notorious problem of clustering is that different clustering algorithms often impose different clustering structures on data [14][17], even the data may contain no cluster at all. A number of cluster validation methods have been developed in the past to tackle this problem [4][14][17][19]. Most methods are based on the statistical framework that one first adopts a null hypothesis of randomness (i.e., no structure in data) and then decides either rejecting or accepting it according to the distribution of a chosen statistical model for a clustering structure. A few employ graphical displays to visually verify the validity of a clustering [21]. Recent surveys on cluster validation methods can be found in [10][16]. The problem of using these cluster validation methods in data mining is that the computational cost is very high when the data sets are large and complex.

Recent work on clustering in data mining has been focused on the development of fast clustering algorithms to deal with large data sets. Interesting results include CLIQUE [1], CLARANS [20], BIRCH [24], DBSCAN [5] and the k-means extension algorithms [12]. These progresses are extremely important because without fast clustering algorithms one cannot conduct any thorough cluster analysis on large data sets. However, without effective cluster validation tools the problem of cluster analysis on large data sets is only partially solved. Unfortunately, this problem has not been well-developed in the data mining community.

In [13], we proposed a visual method for cluster validation in data mining. The visual method uses the Fastmap algorithm [7] to project objects and candidate clusters onto a two-dimensional (2D) space and allows the user to visually examine the clusters created with a clustering algorithm and determine the genuine clusters found. The visual method is based on the principle that *a cluster separate from others in the 2D space is also separate from others in the original high dimensional space* (the opposite is not true). We have used this method in a real case study to interactively cluster a mobile service marketing data set and discover a few interesting clusters of customers [13]. In that case study, we used the visual method to solve two common clustering problems, (1) to verify the separations of clusters created by a clustering algorithm and (2) to determine the number of clusters to be produced. Because the Fastmap algorithm is efficient in processing large data sets, the visual method in combination with a fast clustering algorithm provides a complete solution to the clustering problem in data mining.

*Research supported in part by Research Grant Council Grant Nos. HKU 7147/99P and 7132/00P.

0-7695-0996-7/01 $10.00 © 2001 IEEE

Projection of high dimensional data onto low dimensional spaces for clustering is a common approach in cluster analysis. Fastmap was primarily designed for this purpose [7]. Other widely used methods include principal component analysis (PCA), multidimensional scaling (MDS) [23] and dimensionality reduction techniques such as K-L transform [8]. Ganti et al. [9] has recently integrated Fastmap with the BIRCH clustering algorithm [24] to cluster data in arbitrary spaces. In their approach, Fastmap is used to project data in an arbitrary space onto a projected space in which clustering is performed. However, performing clustering in the projected space cannot guarantee the discovery of clusters in the original space. We advocate creating clusters from the original high dimensional space and using Fastmap to validate these clusters in the projected low dimensional space. When a cluster is validated, we are able to conclude that it is a cluster in the original high-dimensional space.

In this paper, we present a Monte Carlo evaluation on the visual method of cluster validation with Fastmap, proposed in [13]. We constructed a series of artificial data sets with controlled cluster structures and dimensionality and applied a clustering algorithm to cluster these data sets. For each clustering, we used both the visual method and two internal statistical indices to validate the clustering results. Then, we compared these validation results from the visual method and the statistical indices and found there exist a high degree of agreement between the visual method and the statistical methods. This indirectly proved that the visual method can produce validation results equivalent to those of statistical methods. Because the visual method is efficient in validating clusters from large data sets, it is suitable for data mining applications.

This paper is organized as follows: In Section 2, we briefly review the visual method of cluster validation with Fastmap. In Section 3, we present the two statistics used to define the internal indices for cluster validation. The synthetic data generation and validation results of the visual method and statistical methods on the synthetic data sets are discussed in Section 4. Some concluding remarks are given in Section 5.

2 Visual Cluster Validation

2.1 Cluster Validation

Cluster validation refers to the procedures that are used to evaluate clusters generated from a data set by a clustering algorithm [14]. Cluster validation is required due to the facts that no clustering algorithm can guarantee the discovery of genuine clusters from real data sets and that different clustering algorithms often impose different cluster structures on a data set even if there is no cluster structure present in it [10] [16].

Cluster validation is needed in data mining to solve the following problems:

1. To measure a partition of a real data set generated by a clustering algorithm.
2. To identify the genuine clusters from the partition.
3. To interpret the clusters.

The last problem can be solved by computing the summary statistics of each cluster and using the application domain knowledge to interpret the statistics. In this paper, we are interested in the first two problems.

In statistics, cluster validation is treated as a hypothesis test problem [10] [14] [16]. A null hypothesis of "no cluster structure in the data set" is first given on the data set in question. Then, the data set is clustered with a clustering algorithm. After that, a statistic T is calculated from the clustering result and tested against the distribution of T. The null hypothesis is rejected if the probability of the value of T calculated from the clustering result is low at certain significance level. That implies the data set indeed contains clusters. Details of the hypothesis test procedures are given in [14].

A large number of statistical indices for cluster validation were proposed [17]. Two indices will be discussed in Section 3. To use an index to validate clusters, a baseline distribution of it has to be computed from a sufficient number of randomly generated data sets in the same problem domain. Since the computational cost is extremely very high on large data sets, these statistical validation methods can hardly be used in data mining.

2.2 Fastmap Algorithm

Let $\mathcal{O}$ be a set of N objects in an n-dimensional space and d_n a dissimilarly measure between objects in $\mathcal{O}$ in the n-dimensional space. Assume d_n is a metric, having the following properties:

1. $d_n(o_i, o_i) = 0$,
2. $d_n(o_i, o_j) = d(o_j, o_i)$,
3. $d_n(o_h, o_j) \leq d(o_h, o_i) + d(o_i, o_j)$

where o_h,o_i and o_j are objects of $\mathcal{O}$.

The Fastmap projection of the N objects onto a k-dimensional space ($k < n$) uses the following formula to calculate the coordinate $x_{k,i}$ of object o_i on the kth axis of the k-dimensional space.

$$x_{k,i} = \frac{d_{n-k+1}(o_a, o_i)^2 + d_{n-k+1}(o_a, o_b)^2 - d_{n-k+1}(o_b, o_i)^2}{2d_{n-k+1}(o_a, o_b)} \quad (1)$$

where objects o_a and o_b are two pivot objects from $\mathcal{O}$ and $d_{n-k+1}(o_i, o_j)_2$ is calculated as

$$d_{n-k}(o_i, o_j)^2 = d_n(o_i, o_j)^2 - d_k(o_i, o_j)^2 \quad (2)$$

where $d_{n-k}(o_i, o_j)$ is the distance in the $(n-k)$-dimensional space, $d_n(o_i, o_j)$ the distance in the n-dimensional space and

$$d_k(o_i, o_j)^2 = \sum_k (x_{k,i} - x_{k,j})^2 \quad (3)$$

is the square distance in the k-dimensional space. Proofs of these formulas are given in [7].

Objects o_a and o_b are referred to as *pivot objects*. The line passing the pivot objects determines a projection axis. Essentially, there are k pairs of pivot objects required. The selection of the pivot objects can be arbitrary provided that the projection lines are not parallel in the original space.

For a given k, the Fastmap projection is not optimal because there is no attempt to minimizing $\sum_{o_i, o_j \in \mathcal{O}} d_{n-k}(o_i, o_j)^2$, as in multidimensional scaling. However, the Fastmap algorithm is fast ($O(kN)$) so it is suitable for data mining applications [7].

2.3 Cluster Validation with Fastmap

Faloutsos and Lin [7] has shown that the Fastmap projection has a property to reveal clusters of data existing in the original (sometimes unknown) high dimensional space. This property can be used for data clustering [7] [9] and cluster validation [13]. For clustering, we use Fastmap to project objects into a k-dimensional space and then apply a clustering algorithm to cluster data in the low dimensional space [9]. This method is useful when the dimensions of the original space are unknown. However, the clusters found in the projected space cannot be guaranteed clusters in the original space.

For cluster validation, we apply a clustering algorithm to cluster data in the original space and use the Fastmap algorithm to project clusters into a 2D space and visualize them. If a cluster is observed to be separated from other objects on the 2D plots, it is also separated from other objects in the original space. This claim is justified in (2). Let T be the minimal distance between an object o in the cluster and an object o' outside the cluster. Assume o and o' are identifiable on the 2D plot if $d_2^2(o, o') \geq T$. The two objects are also identifiable in the original space because $d_n^2(o, o') \geq CT$ where $C \geq 1$.

Based on the above principle, we can use the Fastmap display to visually validate clusters generated from high dimensional data using a fast clustering algorithm. In data mining, clustering and cluster validation can be conducted interactively. Given a real large data set, we first apply a clustering algorithm to partition it into k clusters. Then, we use the Fastmap algorithm to project the clusters into a 2D space and visualize objects in different clusters in different colors and/or symbols. In this way, we can visually identify some clusters which are separate from other objects. According to the above principle, these clusters are also separate from other objects in the original high dimensional space. In analysis, we can extract these clusters from the data set.

Figure 1 shows a real example of the Fastmap projection of two clusters generated from a telecommunication data set. The symbols "plus" and "triangle" represent two clusters respectively. We see that the objects shown in the symbol "plus" form a compact cluster, which represents a group of customers who churned from the service provider shortly after joining it [13].

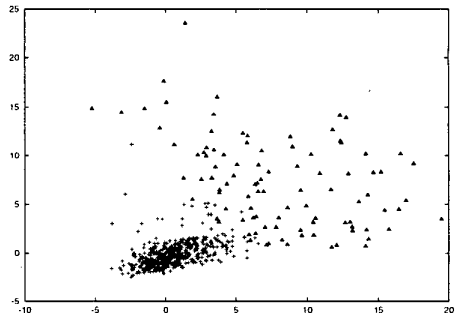

Figure 1. Fastmap projection of two clusters in a real data set.

In [13], we described an interactive approach to clustering and cluster validation with Fastmap for data mining. We use a top-down approach to interactively building cluster trees from data. Starting with the whole data set that is considered as a cluster on its own right, we stepwise decompose the data and grow a tree of clusters. In the tree, a node containing children is a composite cluster while all leaves are atomic clusters.

In this interactive approach, we have to make two decisions at each node to proceed the process. That is, to decide whether a node being considered is a composite or atomic cluster and to determine the number of clusters to be created from it if the node is a composite cluster. With synthetic data, since we know the details of clusters, we will have no difficulty to make these decisions. However, when we deal with real data, we usually have no knowledge about the structure of clusters. The Fastmap algorithm and visualization help us to obtain such knowledge and make decisions. Through Fastmap and visualization, we can make decisions based on what we see. In clustering real world data, this kind of human involvement has a great advantage because

we can use our domain knowledge to accept or reject the clusters generated by the clustering algorithm.

3 Statistical Cluster Validation

3.1 Poisson Model

A null model specifies the type of random data sets that are generated to calculate the baseline distribution of the statistic T used in cluster validation. The Poisson model is one of null models widely used [10]. The Poisson model assumes that the objects are random points which are uniformly distributed in some region A of the n-dimensional space. The Poisson model is used to validate the clustering results of the data sets which can be represented in an N-by-n matrix. A uniform distribution random number generator can be used to generate the random data sets.

3.2 Hypothesis Test

Let T be a statistic and H_0 a null hypothesis stating that no cluster structure exists in data set $\mathcal{X}$. Let Prob($\mathbf{B}|H_0$) be the baseline distribution $\mathbf{B}$ under H_0. $\mathbf{B}$ could be either $T \geq t_\alpha$ or $T \leq t_\alpha$, where t_α is a fixed number called a threshold at significance level α and

$$\text{Prob}(T \geq t_\alpha) = \alpha.$$

Suppose that t_* is the value of T calculated from a clustering result of data set $\mathcal{X}$. If $t_* \geq t_\alpha$, then we reject the hypothesis that states $\mathcal{X}$ contains no cluster structure. This is because the probability that H_0 is true is low (α).

There are many criterion measures to reflect the goodness-of-fit between a data set and its clustering. Several reviews of these criterion measures can be found in [2][17]. In our study, we are interested in determining good partitions of data sets produced by a clustering algorithm and identifying genuine clusters from the partitions. Therefore, we focus on two criterion measures for validating the partition of a data set and a cluster.

3.3 C-index for Validating Partitions

In [17], it has been shown that C-index is an effective criterion measure for validating the partition of a data set. Let D be the sum of all within-cluster dissimilarities, $D_{\min}$ and $D_{\max}$ be the minimum and maximum sums of the within-cluster dissimilarities in the baseline distribution. The C-index is defined by

$$C = \frac{D - D_{\min}}{D_{\max} - D_{\min}}. \tag{4}$$

The smaller C, the better the clustering. Therefore we can use the C-index to validate the partition of a data set.

3.4 U-statistic for Validating Clusters

The statistic U was used to identify genuine clusters. The idea is to compare the dissimilarities between the objects in a cluster $\mathcal{W}$ and the objects outside the cluster $\mathcal{X} - \mathcal{W}$. More precisely, the statistic U is defined as follows:

$$U = \sum_{(i,j)\in\mathcal{W}} \sum_{(k,l)\in\mathcal{X}-\mathcal{W}} U_{ijkl} \tag{5}$$

where

$$U_{ijkl} = \begin{cases} 0, & \text{if } d_n(x_i, x_j) < d_n(x_k, x_l), \\ \frac{1}{2}, & \text{if } d_n(x_i, x_j) = d_n(x_k, x_l), \\ 1, & \text{if } d_n(x_i, x_j) > d_n(x_k, x_l). \end{cases}$$

If a cluster is well-separated from other objects or clusters, the value of the statistic U should be small. Therefore, we can use the U-statistic to validate a cluster.

4 A Monte Carlo Evaluation

It is a common approach to using synthetic data sets with known cluster structures to evaluate clustering algorithms and cluster validation methods [17][18] [19]. We adopted this approach to analytically evaluate the visual cluster validation method. Our purpose was to investigate whether the visual method can produce a result which would be in consistence with the result of a statistical validation method. If the two methods are consistent, then we can conformably use the visual method to validate clusters in data mining without the need to calculate the statistic baseline. This section presents our evaluation results.

4.1 Data Generation

Two types of synthetic data sets were generated. The data sets of the first type contained well-formed clusters distributed in the specified region in an n-dimensional space. The clusters were generated by a multidimensional normal distribution random number generator.

Table 1 lists the control parameters used to generate these data sets. The first three parameters were randomly generated for each data set. However, we restricted the number of objects in each cluster between $0.8 \times N/K \leq N_k \leq 1.2 \times N/K$. We specified $N = 100$ for all data sets. We tested the number of dimensions n on the range between 3 and 5, and the number of clusters K on the range between 3 and 5. We generated 5 different configurations of data sets. For each configuration, 20 data sets were generated. Totally, we generated 100 synthetic data sets containing randomly distributed clusters within a region of unit hyper-boxes.

Table 1. Control parameters

No.	Parameter	Definition
1	μ_k	Mean vector of a cluster
2	Σ_k	Covariance of a cluster
3	N_k	Number of objects in the kth cluster
4	N	Number of objects in a data set
5	n	Number of dimensions
6	K	Number of clusters in a data set

In each data set, we randomly generated K cluster centers. The distances between cluster centers were set between 1.5 and 3. Then we computed the minimum distances among cluster centers. Using these minimum distances, we generated the covariance Σ_k for the kth cluster. N_k objects were generated by using the multi-normal distribution generator with the cluster center as the mean and Σ_k as the covariance. Finally, we re-scaled all the data points to the range between [-1,1] in each dimension.

The data sets of the second type contained randomly generated objects which were uniformly distributed in the same region of unit hyper-boxes. For each specific n, we generated 100 data sets to calculate the baseline distribution. Totally, 500 random data sets were generated using a uniform distribution random generator. Each data set contained 100 objects.

4.2 Experiment

The visual cluster validation method was designed to validate partitions generated from a data set by a clustering algorithm and to identify genuine clusters from a partition. In this experiment, we used the k-means algorithm to generate a partition from a synthetic data set because all the data sets were numeric. In dealing with real data sets, we use the k-prototypes algorithm that can process both numeric and categorical data [12].

To visually validate a partition, we used the Fastmap algorithm to project the clusters into a 2D space and displayed the clusters in different colors and symbols. If we saw well-separated clusters on the display, we considered the partition was valid. If we saw any overlapping between clusters, we considered the partition was not valid. Figure 2 shows two partitions with five clusters. We considered Figure (a) was valid but (b) was not. There were two problems in (b). The clusters shown in boxes and crosses were overlapping and the k-means algorithm failed to separate two clusters shown in stars.

For each partition, we calculated its C-index value using (4). Since we randomly generated 20 synthetic data sets for each configuration, we conducted this validation process twenty times on the data sets in the same configuration.

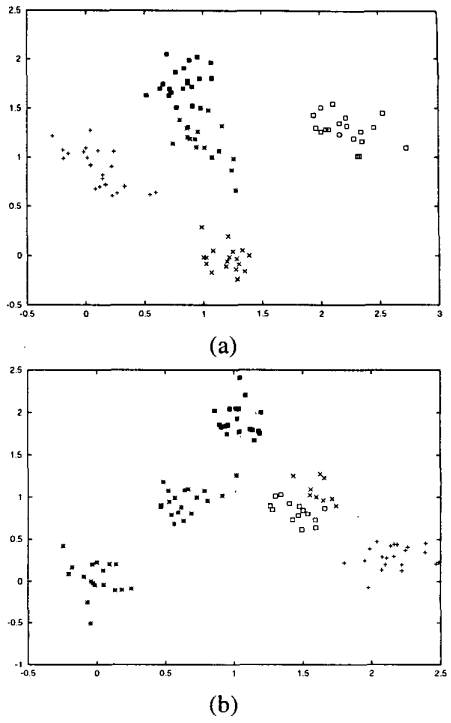

Figure 2. (a) Valid partition. (b) Invalid partition.

We counted the numbers of valid and invalid partitions generated from the 20 data sets and calculated the minimum, mean and maximum C-index values for the valid and invalid partitions. The result is summarized in Table 2, which presents the analytical results of five different configurations of data sets. One observation from the table was that the visual validity of a partition is in correspondence to a smaller C-index value, which implies the partition was significant. This result indicated that the two validation methods were in consistence to each other.

Another observation from Table 2 was that the high percentage of valid partitions indicated that it was easier to generate and validate the valid partitions by the k-means algorithm and the visual validation method from the data sets in low dimensions with less clusters than from those in high dimensions with more clusters. The reason could be that given a fixed number of objects in data sets, the higher the dimensions and the more the clusters, the sparser the distribution of the objects in space.

We noticed that most C-index values are negative, which means the cluster structures of the partitions are significant, in comparison with the baseline distribution because the synthetic data sets were generated with some well formed clusters. This implies that all these invalid partitions con-

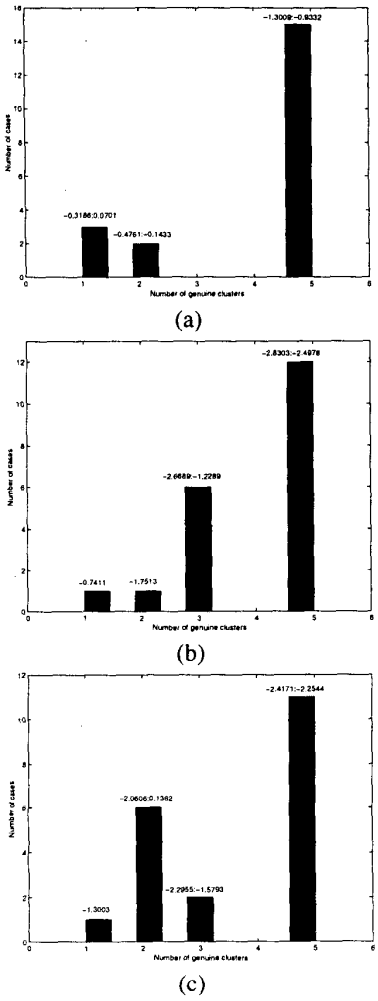

Figure 3. Relations between the number of identifiable genuine clusters and the average of C-index values. (a) 3D, (b) 4D, (c) 5D.

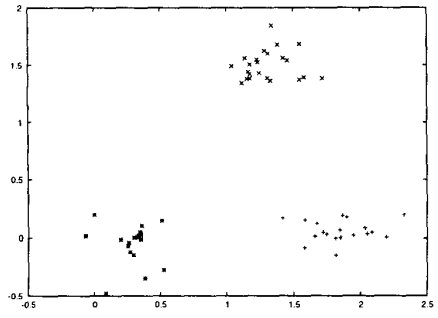

Figure 4. The partition of the data set in Figure 2(b) after removing two identified genuine clusters and re-clustering.

tain some genuine clusters. From the displays, We could identify the genuine clusters from the invalid partitions. For example, from Figure 2(b) we could identify two genuine clusters, one in solid boxes and one in pluses. To summarize the validation results of the invalid partitions, we could reveal the relations, as shown in Figure 3, between the number of genuine clusters identifiable from invalid partitions and the average C-index values calculated from the partitions. One can observe that the average C-index value decreases as the number of identifiable genuine clusters increases. This indicates that the more genuine clusters are recognized in a partition, the more significant the partition is.

To statistically validate an individual cluster shown in a partition, we calculated the U-statistic value of the cluster using (5) and compared the value with the baseline distribution. We found that the more identifiable a cluster is on the display, the more significant the U-statistic is. For example, the five clusters in Figure 2(a) are very identifiable. Table 3 shows their U-statistic values compared with the baseline distribution. These values are very low. However, the clusters in Figure 2(b) are not well-separated, some of their U-statistic values (clusters in crosses, boxes and stars) are quite large compared with the baseline distribution (cf. Table 4). Therefore we could not statistically accept these clusters as genuine clusters. On the other hand, we see from Figure 2(b) that clusters in pluses and solid boxes are quite well-separated from the other clusters. Correspondingly, their U-statistic values (33 and 9) are also small, so we could accept them as genuine clusters. This indicates that genuine clusters can be identified from the 2D display.

After genuine clusters were identified from a data set, we removed them from the data set and used the k-means algorithm to further cluster the rest of the data set. For example, after removing clusters "1" (pluses) and "5" (solid boxes) from the data set in Figure 2(b), we used k-means algorithm to cluster data set (the remaining objects in clusters "2" (crosses), "3" (stars) and "4" (boxes)) and projected them into a 2D space again by Fastmap. The display is shown in Figure 4. From the figure we can see that this partition is visually valid. The C-index value for this partition is -2.5078 and also significant. The U-statistic values for these clusters are given in Table 5, which are also very small and significant. This shows that, by combining the k-means algorithm and visual cluster validation method, genuine clusters can be identified from multiple levels of clustering started from the same data set.

5 Concluding Remarks

In this paper, we have presented a preliminary evaluation on the visual method of cluster validation using Fastmap which we proposed in [13]. We have already shown in [13]

that the visual method was effective in identifying interesting clusters of customers from a telecommunication marketing data set. The contribution of the Monte Carlo evaluation in this paper was to show that the results of visual validation were in consistence with the results of statistic validation.

The visual cluster validation method has a few advantages over statistical methods. First of all, it is fast in processing large data sets, thus suitable for data mining applications. The requirement for baseline construction prohibits the statistical methods from being used in data mining. Secondly, it is easy to comprehend and to be used by non-experts. Thirdly, it tackles three validation problems in one method, namely, (1) the number of clusters in a data set, (2) evaluation of good data partitions and (3) identification of individual clusters. This method is especially useful in solving the first and third problems. Finally, in combination with a fast clustering algorithm, it offers a flexible way to identify genuine clusters in multiple clustering steps. In this approach, the user is involved in making decisions, which is critical in data mining tasks. A real case study on using this multiple step clustering method and visual validation of clusters to identify interesting groups of telecommunication customers was described in [13].

Fastmap algorithm [7] is efficient in projecting high dimensional data into low dimensional spaces. However, unlike the multidimensional scaling methods [23], the projection is not optimized. An interesting problem remains on how to select the projection lines (the pairs of pivot objects) which can best reveal the cluster structures in data. Our next objective is to investigate this problem. We believe the solution will significantly improve the visual cluster validation method for data mining. We will also evaluate the visual method on some benchmark data sets available on the Internet such as http://www.ncc.up.pt/liacc/ML/statlog/.

References

[1] Agrawal, R., Gehrke, J, Gunopulos, D. and Raghavan, P. (1998) Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of SIGMOD Conference.

[2] Cormack, R. (1971) A review of classification. Journal of Royal Statistical Society, Series A, Vol. 134, pp. 321-367.

[3] Dubes, R. C. (1987) How many clusters are best? - an experiment. Pattern Recognition, Vol. 20, No. 6, pp.645-663.

[4] Dubes, R. and Jain, A. K. (1979) Validity studies in clustering methodologies. Pattern Recognition, Vol. 11, pp. 235-254.

[5] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, USA.

[6] Everitt, B. (1974) Cluster Analysis. Heinemann Educational Books Ltd.

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

[8] Fukunaga, K. (1990) Introduction to Statistical Pattern Recognition. Academic Press.

[9] Ganti, V., Ramakrishnan, R., Gehrke, J, Powell, A. L. and French, J. C. (1999) Clustering large datasets in arbitrary metric spaces. ICDE 1999, pp. 502-511.

[10] Gordon, A. D. (1998) Cluster validation, In Data Science, Classification, and Related Methods, ed. C Hayashi, N Ohsumi, K Yajima, Y Tanaka, H-H Bock and Y Baba, Springer, Tokyo, pp 22-39.

[11] Gordon, A. D. (1994) Identifying genuine clusters in a classification. Computational Statistics and Data Analysis 18, pp.516-581.

[12] Huang, Z. (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, Vol. 2, No. 3, pp. 283-304.

[13] Huang, Z. and Lin, T. (2000) A visual method of cluster validation with Fastmap. In Proceedings of PAKDD-2000, pp. 153-164.

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

[15] Kruskal, J. B. and Carroll, J. D. (1969) Geometrical models and badness-of-fit functions. in Multivariate Analysis II, ed. P. R. Krishnaiah, Academic Press, pp.639-670.

[16] Milligan, G. W. (1996) Clustering validation: results and implications for applied analysis. in Clustering and Classification, ed. P. Arabie, L. J. Hubert and G. De Soete, World Scientific, pp.341-375.

[17] Milligan, G. W. (1981) A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, Vol. 46, No. 2, pp.187-199.

Table 2. Summary of visual validation results and C-index values

Dimensions	Clusters	Visual Result		C-index values					
		Valid	Invalid	Valid			Invalid		
				min.	max.	mean	min.	max.	mean
3	5	75 %	25 %	-1.3009	-0.9332	-1.1748	-0.4761	0.0761	-0.2084
4	5	60 %	40 %	-2.8303	-2.4978	-2.6872	-2.6689	-0.7411	-1.8500
5	5	55 %	45 %	-2.4171	-2.2544	-2.3417	-2.2955	-0.1362	-1.4705
5	4	65 %	35 %	-1.8469	-1.5690	-1.6817	-1.4604	0.2366	-0.5306
5	3	90 %	10 %	-2.1288	-1.6248	-1.8910	-0.7243	0.7230	-0.0001

[18] Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika, Vol. 50, No. 2, pp.159-179.

[19] Milligan, G. W. and Isaac, P. D. (1980) The validation of four ultrametric clustering algorithms. Pattern Recognition, Vol. 12, pp.41-50.

[20] Ng, R. and Han, J. (1994) Efficient and effective clustering methods for spatial data mining. In Proceedings of VLDB, 1994.

[21] Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, Vol. 20, pp.53-65.

[22] Theodoridis, S. and Koutroumbas, K. (1999) Pattern Recognition. Academic Press.

[23] Young, F. W. (1987) Multidimensional scaling: history, theory and applications. Lawrence Erlbaum Associates.

[24] Zhang, T. and Ramakrishnan, R. (1997) BIRCH: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, Vol. 1, No. 2, pp. 141-182.

Table 3. U-statistic values (U) for a valid partition shown as Figure 2(a), where M is the number of objects.

Clusters	M	U	Baseline Distribution		
			minimum	maximum	mean
1 (+)	23	83	204300	310920	254300
2 (×)	19	47	144430	210920	171740
3 (*)	17	1015	110250	160830	137240
4 (□)	19	0	144430	210920	171740
5 (■)	22	22	202490	294870	250530

Table 4. U-statistic values (U) for a valid partition shown as Figure 2(b), where M is the number of objects.

Clusters	M	U	Baseline Distribution		
			minimum	maximum	mean
1 (+)	21	33	173890	252540	211150
2 (×)	9	1823	25273	42379	54265
3 (*)	37	205283	725700	889710	807870
4 (□)	14	2337	74819	118600	95471
5 (■)	19	9	144430	210920	171740

Table 5. U-statistic values (U) for the three remaining clusters shown in Figure 4, where M is the number of objects.

Clusters	M	U	Baseline Distribution		
			minimum	maximum	mean
1 (+)	19	4	144430	210920	171740
2 (×)	23	0	204300	310920	254300
3 (*)	18	6	117910	186730	152760