A Visual Method for High-Dimensional Data Cluster Exploration

Ke-Bing Zhang[1], Mao Lin Huang[2], Mehmet A. Orgun[1], and Quang Vinh Nguyen[3]

[1] Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
{kebing,mehmet}@science.mq.edu.au
[2] Faculty of Engineering and Information Technology, University of Technology, Sydney NSW 2007, Australia
maolin@it.uts.edu.au
[3] School of Computing and Mathematics, University of Western Sydney, NSW 1797, Australia
vinh@scm.uws.edu.au

Abstract. Visualization is helpful for clustering high dimensional data. The goals of visualization in data mining are exploration, confirmation and presentation of the clustering results. However, the most of visual techniques developed for cluster analysis are primarily focused on cluster presentation rather than cluster exploration. Several techniques have been proposed to explore cluster information by visualization, but most of them depend heavily on the individual user's experience. Inevitably, this incurs subjectivity and randomness in the clustering process. In this paper, we employ the statistical features of datasets as predictions to estimate the number of clusters by a visual technique called HOV^3. This approach mitigates the problem of the randomness and subjectivity of the user during the process of cluster exploration by other visual techniques. As a result, our approach provides an effective visual method for cluster exploration.

Keywords: Cluster Exploration, Visualization, Statistics.

1 Introduction

Cluster analysis is an important technique of knowledge acquisition in data mining. To address the requirements of different applications, a large number of clustering algorithms have been developed [9, 3]. However, those algorithms are not very effective in coping with arbitrarily shaped clusters. In addition, cluster analysis is a highly iterative process. However most of existing clustering methods are too automated to exploit the domain experts' knowledge in the intermediate process of clustering. As a consequence, they are not always effective to cluster datasets with a large number of variables and/or huge-sized datasets in real world applications. In a high dimensional space, traditional clustering algorithms tend to break down in terms of efficiency as well as accuracy because data does not cluster well anymore [1].

In order to solve those problems, Shneiderman [19] proposed to present data as a visual plot, so that the user could see the interesting features easily. He pointed out that, visualization can be very powerful and effective in revealing trends, highlighting

C.S. Leung, M. Lee, and J.H. Chan (Eds.): ICONIP 2009, Part II, LNCS 5864, pp. 699–709, 2009.

outliers, showing clusters, and exposing gaps in high-dimensional data analysis. Therefore, the use of visualization 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].

Clustering is an exploratory activity [9]. It is an iterative process under the guidance of user domain knowledge. In most cases of the preprocessing stage of clustering, it is hard for the user to estimate the proper cluster number [3]. Visualization is very helpful for the user to do that. However, cluster exploration by visualization mostly depends on the individual user's experience. Thus, subjectivity, randomness and impreciseness may be introduced into the cluster exploration process. As a result, cluster analysis based on imprecise results may be inefficient and ineffective. On the other hand, cluster exploration based on the user's random interaction is arbitrary and it may not be easy to interpret from where the grouped results come.

In this paper, based on the projection of a technique called Hypothesis Oriented Verification and Validation by Visualization (HOV^3) [22], we introduce the statistical features of datasets as the predictions of HOV^3 to guide the user on cluster exploration, because the statistical summaries objectively reflect the features of datasets. As a result, it provides the user an effective method on determining cluster numbers in the preprocessing stage of cluster analysis.

The rest of this paper is organized as follows. Section 2 briefly introduces cluster visualization techniques and gives a short introduction to the HOV^3 technique. Section 3 presents the algorithm of statistics-guided visual approach for cluster detection by HOV^3. Section 4 demonstrates the effectiveness of our approach by an experimental analysis on several datasets[1]. Finally, Section 5 summarizes the contributions of this paper.

2 Background

2.1 Visual Cluster Analysis

Visual cluster analysis is a combination of visualization and cluster analysis. It is believed that the combined strength of visualization and data mining would enrich both approaches and enable more successful solutions [20]. However, the data to be processed by clustering is usually high dimensional. It is not easy to visualize multidimensional data on 2D or 3D space and still give a "genuine" visual interpretation. This is because mapping higher dimensional data onto lower dimensional space inevitably introduces ambiguities, overlapping and even bias. Thus, choosing a technique to fit visualizing clusters of high dimensional data is the first and most crucial task of visual cluster analysis.

In practice, instead of providing a quantitative guidance on cluster exploration, the most of the cluster visualization techniques are typically used as an observational mechanism to assist the user in having intuitive comparisons and understanding of clustering results better. Several approaches have been proposed to help the user on cluster exploration.

---

[1] The datasets used in this paper are available from http://archive.ics.uci.edu/ml/

For example, Multidimensional scaling (MDS) maps multidimensional data as points into 2D Euclidean space, where the distances between data points reflect the similarity/dissimilarity of them [14]. However, the relative high computational cost of MDS, with polynomial time complexity $O(N^2)$, limits its usability on very large datasets. PCA is a commonly used multivariate analysis technique [10], mainly used for reducing the dimensionality of high dimensional data by extracting the representative variables. However, PCA is sensitive to deal with the non-linear data structure. It is not suitable for the exploration of unknown data. A Grand Tour based visual technique is proposed to visualize cluster structures [5], but this technique visualizes 3 clusters only. To deal with more than 3 clusters with a more sophisticated Grand Tour technique, more assistance is required.

OPTICS uses a density-based technique to detect cluster structures and visualizes them in "Gaussian bumps" [2]. It is an intuitive method to assist the user to observe cluster structures, 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.

Huang *et. al* [7, 8] proposed several approaches 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.

CVAP [21] is a recently proposed prototype with several integrated clustering algorithms and cluster validation methods. It is a convenient toolkit to assist the user on the selection of clustering scheme for the application of small-sized datasets. However, CVAP is only for displaying the clustering and cluster validation results, rather than the purpose for directly evolving the user into the cluster exploration process.

2.2 Star Coordinates

The projection of Star Coordinates [11] has only linear time complexity, which is significant for interactive cluster visualization of very large datasets. VISTA [4] and HOV^3 [22] extend Star Coordinates by additional features to mitigate the problem of overlapping and ambiguities caused by projecting high dimensional data onto 2D space. The visual approach reported in this paper has been developed based on the projection of HOV^3. For the sake of completeness, we briefly introduce the Star Coordinates technique here.

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 surface on the 2D space [11]. Star Coordinates first normalizes data in each dimension into a unit interval [0, 1]. Then the values of all axes are mapped to an orthogonal X-Y coordinate which shares the centre point with Star Coordinates on the 2D space. Thus, an n-dimensional data item is represented as a point in the X-Y 2D plane by Star Coordinates. Based on this projection, several interaction mechanisms, such as axis scaling, axis rotation, data point filtering are provided in Star Coordinates to change the data distribution of a dataset in order to detect cluster characteristics and render clustering results.

However, it is not easy to give an explanation of the grouping results produced by the user's random interactions in Star Coordinates and VISTA, also the grouping results are usually not repeatable. On the other hand, in the Star Coordinates space, the user's interactions cannot change the data distribution too much when the dimensionality of the dataset is very high (a hundred or more dimensions, which is very common in data mining). This is because the alteration of the data distribution by applying interactions to an axis is much less than that of lower dimensional data in the Star Coordinates space. As a result, in very high dimensional space, it is not effective anymore to separate clusters or explore grouping clues by the interactions of Star Coordinates and VISTA.

As discussed above, the issues of arbitrary exploration and/or complicated visual representation of cluster structures make those techniques inefficient and time consuming on cluster exploration of large and high dimensional data. As Seo and Shneiderman [18] mentioned that "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". Thus developing an effective visualization technique to assist the user during cluster exploration and detection is the main aim of this research.

2.3 HOV^3

To remedy the randomness and arbitrariness of visualization on cluster analysis, Zhang *et al.* mathematically generalized the Star Coordinates model by the Euler formula and proposed their visual approach HOV^3 to detect clusters [22]. 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. Then Star Coordinates can be simply written as:

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

where $\min d_k$ and $\max d_k$ represent the minimal and maximal values of the kth coordinate respectively. Equation (1) can be viewed as a mapping from $R^n \rightarrow C^2$.

Conversely, instead of using a random exploration of cluster information by axis scaling or axis rotation in Star Coordinates/VISTA, HOV^3 quantifies the user's apriori knowledge/estimation of a studied dataset as a measure vector to precisely guide the user on the exploration of group information. A measure vector M in HOV^3 represents the corresponding axes' weight values. Then given a non-zero measure vector M in 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:

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

where m_k is the kth variable of measure M .

It can be observed that, equation (2) is a standard form of linear transformation of n variables, where m_k is the coefficient of the kth variable of P_j.

3 Cluster Exploration by HOV^3

We propose a statistics-guided cluster exploration approach by HOV^3 based on the following idea. In analytic geometry, the difference of two vectors *A* and *B* can be expressed by their inner product *A.B*, with its geometrical meaning that the data distribution is plotted by vector *A* against vector *B* (and vice versa). The inner product between a dataset and a measure vector in HOV^3 can be geometrically viewed as a data distribution plotted by a set of vectors against the measure vector in the HOV^3 space, as shown in equation (2).

Predictive knowledge discovery is an important knowledge acquisition method, which utilizes the existing knowledge to deduce, infer, reason and establish predictions, and verify the validity of the predictions. As mentioned above, the user can quantify his/her priori knowledge of a studied dataset as the guidance on the exploration of group information. Thus the statistical summaries of a dataset can be directly employed as the statistical predictions (measure vectors) of the dataset in HOV^3, since the statistical summaries reflect the nature comparisons of data objectively [19]. Also, it is easy to interpret the grouping results of a dataset plotted by statistical predictions in HOV^3. The detailed description of our approach is presented below.

3.1 The Algorithm

We formalized our idea of using statistical predictions to explore clusters by HOV^3 into the algorithm in table 1. The detailed explanation of our algorithm is given next.

Table 1. The Algorithm of Statistics-guided Cluster Exploration by HOV^3

Algorithm: Statistics-guided Cluster Exploration by HOV^3

Input: *D*: a dataset; *M*: statistical measures of *D*;

Output: G**:** data distribution of *D* or subsets of *D*;

```
cluster exploration ← true;
p ← D;
mi ← a statistical measure of p;  (mi ∈ M)
m ← mi;
while (cluster exploration)
    G ← Hc(D, mi);
    if (G well grouped?)
        if (stop exploration?)
            cluster exploration ← false;
            break;
        endif
    endif
    if (new statistical measure of p?)
        mi ← a statistical measure of p;
        m ← mi;
    endif
    mi ←m·mi ;
endwhile
```

3.2 Supported Features

There are two significant features of the use of statistical predictions to explore clusters by HOV^3: *Enhanced separation of data groups* and *quantitatively guided exploration*. The projection of HOV^3 is simply written as G ← *Hc(D, m)* [23], where *D* is the processing dataset, *m* is a measure vector, and G is the distribution of D projected by HOV^3.

- **Enhanced Group Separation**

It is proved that if there are several data point groups that can be roughly separated by applying a measure vector *m* in HOV^3 to a dataset, then multiple applications of the projection in HOV^3 with the same measure vector to the dataset would lead to the groups being more condensed, i.e., have a good separation of the groups [24].

This feature is achieved by step 6 and step17 in the **while** loop (steps 5-18) of the algorithm, as shown in Table 1. The enhanced group feature is significant for cluster exploration by HOV^3 with statistical predictions, since clearly separated groups cannot be usually observed by applying a measure vector to a dataset in HOV^3 once.

- **Quantitatively Guided Exploration**

The HOV^3 technique provides a quantitative mechanism to visually detect cluster clue by measure vectors. In fact, the statistical summaries of a dataset are quantitative depictions of the dataset. They objectively reflect the natural comparisons of the dataset. Thus introducing them as the predications in HOV^3 avoids the randomness and subjectivity which may be introduced by the user during the cluster exploration process by visualization.

To highlight these two features and demonstrate the effectiveness of our approach, we provide several experiments in the next section.

4 The Experiments

4.1 Parkinson's Disease Dataset

Parkinson's disease dataset has 23 attributes and 195 instances. The original data distribution of Parkinson's disease dataset is shown in Fig. 1, where we cannot recognize any groups in the dataset. Then we choose the standard deviation of the dataset *pstd*=[0.24096, 0.18676, 0.25056, 0.15401, 0.13764, 0.14296, 0.14786, 0.14293, 0.17215, 0.16013, 0.19555, 0.16314, 0.12977, 0.19553, 0.12865, 0.17987, 0.43188, 0.24253, 0.22046, 0.13688, 0.18776, 0.17029, 0.18665] as a statistical prediction to explore the clusters of the dataset. Its projected data distribution is illustrated in Fig. 2, where data points are roughly separated, but we still cannot distinguish groups clearly (3 or 4 groups?).

According to the enhanced separation feature of HOV^3 [24], we adopt two times inner product of *pstd* as a statistical prediction and try again. The newly projected result is shown in Fig.3, where the data points are separated into two mains groups, based on the user's observation. We have also used three times mean value of Parkinson's dataset as the statistical prediction to plot the dataset. Its data distribution is shown in Fig.4. It can be observed that, clearly, there are two groups in both Fig.3 and Fig.4.

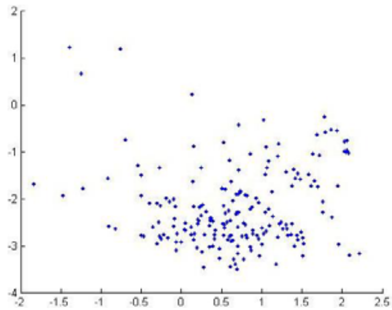

Fig. 1. Projecting data distribution by HOV^3 of Parkinson's disease dataset without any measurement

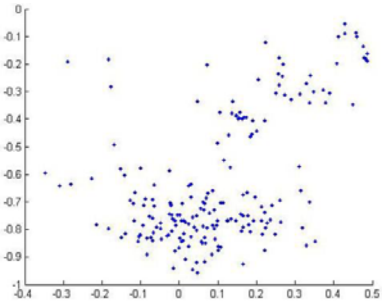

Fig. 2. The data distribution projected by HOV^3 of Parkinson's disease dataset with its standard deviation, *pstd* as a statistical prediction

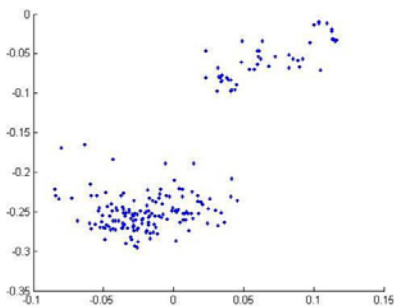

Fig. 3. The data distribution projected by HOV^3 of Parkinson's disease dataset in Fig.1 with two times of *pstd* as the prediction

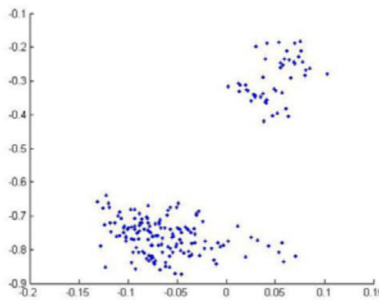

Fig. 4. The data distribution projected by HOV^3 of Parkinson's disease dataset with three times of mean values of the dataset as a prediction

Based on the above experiments, there are two well-separated clusters in Parkinson's disease dataset. The cluster exploration process can be done iteratively until the user is satisfied by the grouping result by HOV^3. He/she can terminate the cluster exploration process by his/her decision (steps 7-12) in table 1.

To verify the validation of the above experiments produced by HOV^3, we employed the CVAP system [21] to check the quality of clustering results of Parkinson's disease dataset by K-means [15] and PAM [12] clustering algorithms with a cluster number of 2 to 10. Then we checked the quality of those clustering results by the cluster validation methods of Silhouette index [17] and Dunn index [6]. The higher Silhouette and Dunn indices indicate the better quality of clustering results. The quality tests of those clustering results are illustrated in Fig.5 and Fig.6. It is clear that number 2 is the optimal cluster number of Parkinson's disease dataset for K-means and PAM clustering. This example shows that statistics-guided cluster exploration by HOV^3 provides an effective visual method to assist the user on the acquisition of the cluster number in the preprocessing stage of clustering.

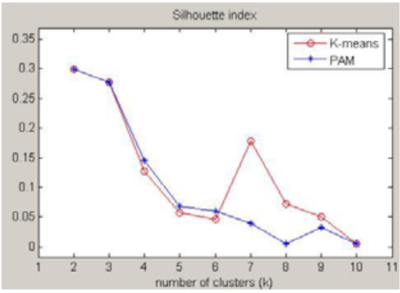

Fig. 5. The quality indicated by Silhouette index of clustering results of Parkinson's disease dataset produced by K-means and PAM clustering algorithms with the cluster number ranging from 2 to 10

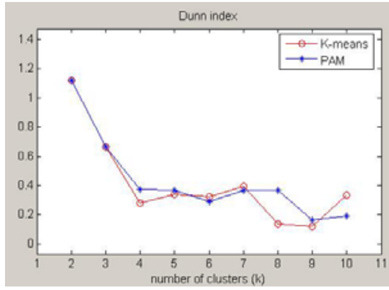

Fig. 6. The quality indicated by Dunn index of clustering results of Parkinson's disease dataset produced by K-means and PAM clustering algorithms with the cluster number ranging from 2 to 10

4.2 Wine Dataset

We have also applied our approach to the *wine* dataset, which has 13 attributes and 178 instances. Fig.7 and Fig.8 present the original data distribution of the wine dataset and the data distribution projected by HOV^3 with three times standard deviation of the dataset respectively. Clearly, there are three well-separated groups in Fig.8. Then we cluster these three groups (C_H).

SOM (*Self-organizing Map*) is a neural network based clustering algorithm [13], which has been widely applied in machine learning and data mining. We applied the SOM to the *wine* dataset with cluster number 2-10, and employed the Silhouette index validation algorithm to verify the clustering results in CVAP. Fig.9 illustrates the curve of validation results produced by Silhouette index in CVAP, where we can observe that number 3 is the optimal cluster number of the wine dataset.

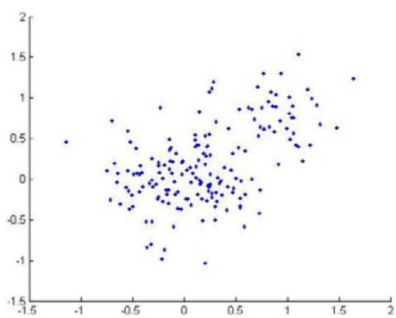

Fig. 7. The original data distribution of the *wine* dataset by HOV^3

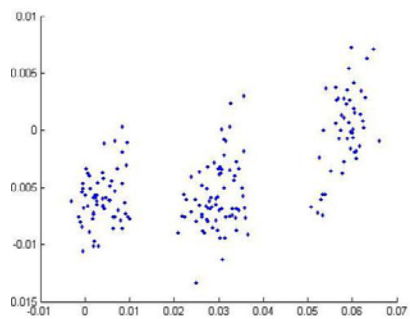

Fig. 8. The data distribution projected by HOV^3 of the *wine* dataset with its three times of stand deviation values

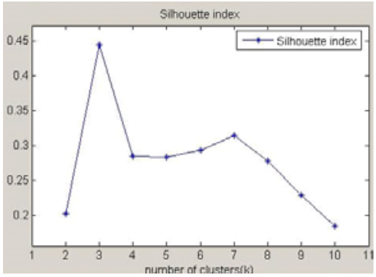

Fig. 9. The quality indicated by Silhouette index of clustering results of the wine dataset produced by SOM clustering algorithms with a cluster number of 2 to 10

Table 2. The statistical contrast between the clusters (k=3) produced by HOV^3 with three times standard deviation of the wine dataset and the clusters produced by SOM clustering algorithms

C_H	*%*	*Radius*	*Variance*	*Weighted Variance*	C_S	*%*	Radius	Variance	*Weighted Variance*
1	26.966	102.286	0.125	3.37075	1	33.708	107.980	0.124	4.179792
2	39.888	97.221	0.182	7.259616	2	38.764	97.449	0.185	7.17134
3	33.146	108.289	0.124	4.110104	3	27.528	102.008	0.126	3.468528
				14.74047					**14.81966**

The contrast of the clusters (C_H) projected by HOV^3 and the clustering result (C_S) produced by the SOM clustering algorithm is summarized in Table 2. The weighted variance of the two clustering results is listed in the last row of the table. We can see that the quality of C_H is even slightly better than the quality of C_S based on the variance contrast. We believe that a domain expert could give a better and intuitive explanation about this clustering result. This experiment also supports the effectiveness of our approach.

As the examples have demonstrated, visual projection based on the statistical prediction by HOV^3 is a more purposeful and effective method for cluster exploration, and also it is easier to obtain a geometrical interpretation of the clustering results.

5 Conclusions

We have proposed a statistics-guided visual approach to assist the user during cluster exploration, and demonstrated its effectiveness by experiments on several datasets. This approach adopts the statistical summaries of a high dimensional dataset as predictions to project the data so that the user can have an intuitive observation of clusters during cluster exploration. The use of statistical features of data mitigates the weaknesses of randomness and arbitrary exploration of the existing visual methods employed in data mining. As a consequence, with the features of enhanced group separation and quantitatively guided exploration of our approach, the user can effectively identify the cluster number in the preprocessing stage of clustering.

References

1. Abul, A.L., Alhajj, R., Polat, F., Barker, K.: Cluster Validity Analysis Using Subsampling. In: Proc. of IEEE International Conference on Systems, Man, and Cybernetics, vol. (2), pp. 1435–1440 (2003)
2. Ankerst, M., Breunig, M., Kriegel, H.-P., Sander, J.: OPTICS: Ordering Points To Identify the Clustering Structure. In: Proceedings of ACM SIGMOD 1999, pp. 49–60 (1999)
3. Berkhin, P.: A Survey of Clustering Data Mining Techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.) Grouping Multidimensional Data, pp. 25–72. Springer, Heidelberg (2006)
4. Chen, K., Liu, L.: VISTA: Validating and Refining Clusters via Visualization. Journal of Information Visualization l3(4), 257–270 (2004)
5. Dhillon, I.S., Modha, D.S., Spangler, W.S.: Visualizing class structure of multidimensional data. In: The 30th Symposium on the Interface: Computing Science and Statistics, vol. (30), pp. 488–493 (1998)
6. Dunn, J.C.: Well Separated Clusters and Optimal Fuzzy Partitions. Journal of Cybern. (4), 95–104 (1974)
7. Huang, Z., Cheung, D.W., Ng, M.K.: An Empirical Study on the Visual Cluster Validation Method with Fastmap. In: The proceedings of DASFAA 2001, pp. 84–91. Springer, Heidelberg (2001)
8. Huang, Z., Lin, T.: A visual method of cluster validation with Fastmap. In: The proceedings of PAKDD 2000, 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, I.T.: Principal Component Analysis. Springer, Heidelberg (2002)
11. Kandogan, E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. In: The proceedings of ACM SIGKDD 2001, pp. 107–116 (2001)
12. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data, An Introduction to Cluster Analysis. John Wiley and Sons, Brussels (1990)
13. Kohonen, T.: Self-Organizing Maps, vol. 30. Springer, Heidelberg (1995)
14. Kruskal, J.B., Wish, M.: Multidimensional Scaling, SAGE university paper series on quantitive applications in the social sciences, pp. 07–011. Sage Publications, CA (1978)
15. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: The proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
16. Pampalk, E., Goebl, W., Widmer, G.: Visualizing Changes in the Structure of Data for Exploratory Feature Selection. In: The proceedings of SIGKDD 2003, Washington, DC, USA (2003)
17. Rousseeuw, P.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computing 20, 53–65 (1987)
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, pp. 232–240. 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. Westphal, C., Blaxton, T.: Data Mining Solutions: Methods and Tools for Solving Real-World Problems. John Wiley and Sons, Chichester (1999)

21. Wang, K., Wang, B., Peng, L.: CVAP: Validation for Cluster Analyses. Data Science Journal 8(20), 88–93 (2009)
22. Zhang, K.-B., Orgun, M.A., Zhang, K.: HOV^3: An Approach to Visual Cluster Analysis. In: Li, X., Zaïane, O.R., Li, Z.-h. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 316–327. 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 (CIDM 2007), Honolulu, Hawaii, USA, April 1-5, pp. 576–582. IEEE Press, Los Alamitos (2007)
24. Zhang, K.-B., Orgun, M.A., Zhang, K.: A Prediction-based Visual Approach for Cluster Exploration and Cluster Validation by HOV^3. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) PKDD 2007. LNCS (LNAI), vol. 4702, pp. 336–349. Springer, Heidelberg (2007)