Validating and Refining Clusters via Visual Rendering

Keke Chen Ling Liu
College of Computing, Georgia Institute of Technology
{kekechen, lingliu}@cc.gatech.edu

Abstract

The automatic clustering algorithms are known to work well in dealing with clusters of regular shapes, e.g. compact spherical/elongated shapes, but may incur higher error rates when dealing with arbitrarily shaped clusters. Although some efforts have been devoted to addressing the problem of skewed datasets, the problem of handling clusters with irregular shapes is still in its infancy, especially in terms of dimensionality of the datasets and the precision of the clustering results considered. Not surprisingly, the statistical indices works ineffective in validating clusters of irregular shapes, too. In this paper, we address the problem of clustering and validating arbitrarily shaped clusters with a visual framework (VISTA). The main idea of the VISTA approach is to capitalize on the power of visualization and interactive feedbacks to encourage domain experts to participate in the clustering revision and clustering validation process.

1. Introduction

Over the past decades most of the clustering research has been focused on automatic clustering algorithms and statistical validity indices. The automatic methods are known to work well in dealing with clusters of regular shapes, e.g. compact spherical or elongated shapes, but incur high error when dealing with arbitrarily shaped clusters. Some new algorithms like CURE [2], WaveCluster [12], DBSCAN [9], and OPTICS [15] have addressed this problem and try to solve it in restricted situations (low dimensional datasets or the cluster shapes are elongated/enlarged). Yet it is still considered as an unsolved hard problem due to the complexity in multi-dimensional space and the unpredictable skewed cluster distributions.

Since clustering is an unsupervised process, cluster validity indices are used to evaluate the quality of clusters, (the compactness or density of clusters, and the dissimilarity between clusters, etc.[11]) and particularly, cluster validity indices are used to decide the optimal number of clusters. The arbitrarily shaped clusters also make the traditional statistical cluster validity indices ineffective [11], which leaves it difficult to determine the optimal cluster structure.

It is possible to invent some complicated automatic clustering algorithms or statistical methods to adapt various specific irregular situations. However, the irregularity cannot be anticipated in applications. Some irregularly shaped clusters may be formed by combining two regular clusters or by splitting one large cluster with the incorporation of domain knowledge. There are no general rules to describe the irregularity. Therefore, the automatic algorithms or statistical methods are not flexible enough to adapt all application-specific requirements.

One feature of the automatic clustering algorithms is that it almost excludes human from the clustering process. What the user can do is usually setting the parameters before the clustering algorithm running, waiting for the algorithm producing the results, validating the results and repeating the entire process if the results are not satisfactory. Once the clustering algorithm starts running, the user cannot monitor or steer the cluster process, which also makes it hard to incorporate domain knowledge into the clustering process and especially inconvenient for large-scale clustering since the iterative cycle is long.

Since the geometry and density features of clusters derived from the distance (similarity) relationship, determines the validity of clustering results, no wonder that visualization is the most intuitive method for validating clusters, especially the clusters in irregular shape. However, cluster visualization is also highly challenging because of the difficulty in visualizing multi-dimensional (>3D) datasets.

Generally speaking, clustering algorithms and validity indices have to answer the two questions: “how to recognize the special structure of each particular dataset?” and “how to refine a given imprecise cluster definition?” In this paper, we propose a visual framework that allows the user to be involved into the clustering/validating process via interactive visualization. The core of the visual framework is the visual cluster rendering system VISTA. VISTA can work with any algorithmic results – at the beginning, VISTA imports the algorithmic clustering result into the visual cluster rendering system, and then lets the user participate in the following “clustering-evaluation” iterations interactively. With the reliable mapping mechanism employed by VISTA system, the user can visually validate the defined clusters via interactive operations. The interactive operations also allow the user to refine the clusters or incorporate domain knowledge to define better cluster structure.

We organize the paper as following. The visual framework and VISTA system are introduced in section 2; in section 3, two empirical examples are demonstrated in details to show the power of VISTA in validating and refining clusters for real datasets. The related work is discussed in section 4. Finally, we conclude our work.

2. VISTA visual framework

Most frequently, the clustering is not finished when the computer/algorithm finishes unless the user has evaluated, understood and accepted the patterns or results, therefore, the user has to be involved in the "clustering – analysis/evaluation" iteration. Concrete discussion about the framework can be found in [8]. We observed that with automatic approaches clustering phase and validating phase should only be done in sequence. In order to interweave these two phases to improve the efficiency, we develop an interactive cluster visual rendering system to get human involved in. With this framework, the user can participate in the clustering process, validating the clusters, monitoring and steering the clustering process.

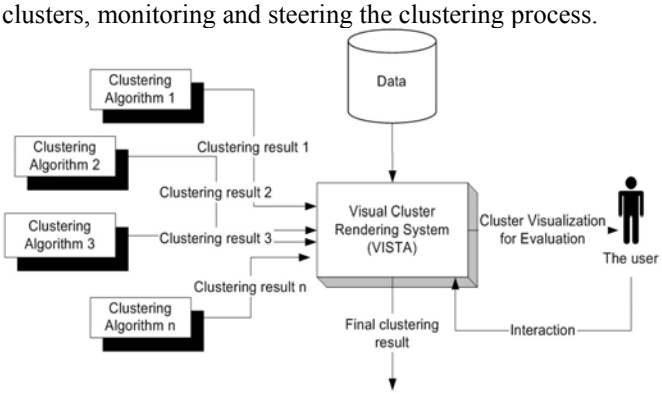

Figure 1. VISTA validating and refining clusters

There are some challenges for interactive cluster visualization techniques, among which the most challenging one is cluster preserving – the clusters appearing in the 2D/3D visualization should be the real clusters in k-D (k>=3) space. Since a k-D to 2D/3D mapping inevitably introduces visual bias, such as broken clusters, overlapping clusters or fake clusters formed by outliers, additional interactive rendering techniques are needed to improve the visual quality.

In VISTA cluster rendering system, we use a linear (or affine) mapping [13] – α-mapping to avoid the breaking of clusters after mapping, but the overlapping and fake clusters may still exist. The compensative techniques are interactive operations to produce dynamic visualization. The interactive operations are used to change the projection plane, which allows the user to observe the datasets from different perspectives. While the visual cluster rendering system is combined with the algorithmic result, the two can improve each other.

To illustrate how the VISTA works, we will briefly introduce the α-mapping and some interactive operations. The initial version of VISTA is used to render Euclidean datasets, where the similarity is defined by Euclidean distance, since the Euclidean distance is widely used in applications.

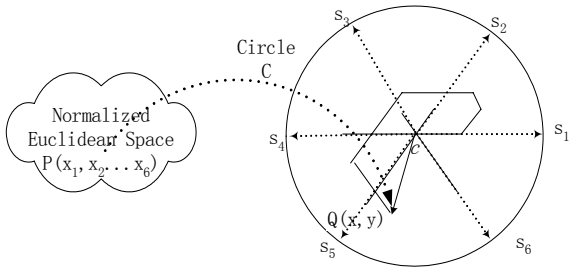

Figure 2. Illustration of α-mapping with *k*=6

We invent a linear mapping α-mapping that partially preserves *k*-dimensional (*k*-D) information in 2D space and is used to build a *k*-parameter-adjustable interactive visualization system. A *k*-axis 2D star coordinates is defined by an origin $\vec{o}\,(x_0, y_0)$ and *k* coordinates $S_1, S_2, \ldots, S_k$, which represent the *k* dimensions in 2D spaces. The *k* coordinates are equidistantly distributed on the circumference of the circle *C*, as in Figure 2, where the unit vectors are $\vec{S}_i = (\hat{u}_{xi}, \hat{u}_{yi})$, $i = 1..k$, $\hat{u}_{xi} = \cos(2\pi/i), \hat{u}_{yi} = \sin(2\pi/i)$. The radius *c* of the circle *C* is the scaling factor, which determines the size and the detail level of the visualization. Let a 2D point *Q* (*x*, *y*) represent the mapping of a *k*-dimensional max-min normalized (with normalization bounds [-1, 1]) data point $P(x_0, x_1, \ldots x_i \ldots, x_k)$ on the 2D star coordinates.

α-mapping:

$$A(x_1,\ldots,\ x_k,\alpha_1,\ldots,\ \alpha_k) = (c/k)\sum_{i=1}^{k} \alpha_i x_i \vec{s}_i - \vec{o} \quad (1)$$

i.e. the position of *Q* (*x*, *y*)is determined by,

$$\left\{(c/k)\sum_{i=1}^{k}\alpha_i x_i \cos(2\pi/i) - x_0,\quad (c/k)\sum_{i=1}^{k}\alpha_i x_i \sin(2\pi/i) - y_0\right\}$$

The α_i (i = 1,2,…*k*, $-1 \le \alpha_i \le 1$) in the definition are dimensional adjustment parameters, one for each of the *k* dimensions. α_i is set to 0.5 initially.

The α-mapping has two important properties: (1) the mapping is linear, and thus it does not break clusters and the gaps in visualization are the real gaps in the original space. (2) The mapping is dimension-by-dimension adjustable by α_i, which enables the dynamic rendering operations to find the cluster overlapping.

Since α-parameter adjustment is the most frequently used one, some operations, such as random rendering and automatic rendering, are used to increase the efficiency of α-parameter adjustment [1]. Another set of set-oriented operations are used to refine visual cluster definition after we get initial cluster visualization with α-parameter

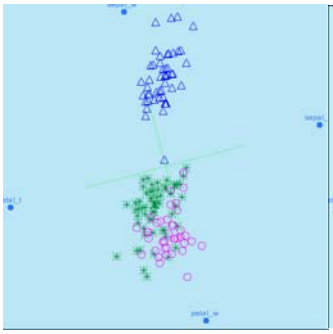

Figure 3-1 load in

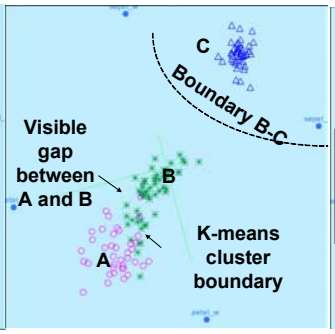

Figure 3-2 Visual clusters

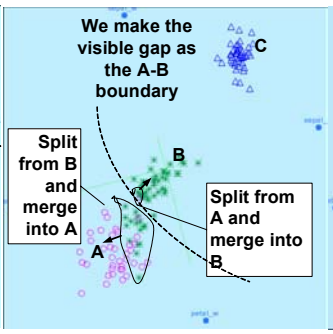

Figure 3-3 editing

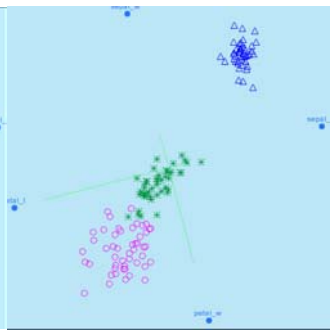

Figure 3-4: real clusters

* k-means result, RMSSTD =0.4108, RS = 0.8002, S_Dbw = 1.4158. After editing, EMSSTD = 0.4396, RS=0.7712, S_Dbw =1.5115

adjustment. These operations include subset selection, cluster marking, cluster splitting, cluster merging, and hierarchical structure defining. Domain knowledge in form of labelled items can be incorporated into visualization conveniently. Due to the space limitation, we will not introduce them concretely.

3. Empirical study

In this section, we will introduce two examples of visual rendering. The first one demonstrates the ability of VISTA visual validating and interactive refining. The second one shows how to incorporate domain knowledge into VISTA visual cluster rendering. The datasets used in the examples can be found at UCI Machine Learning database (http://www.ics.uci.edu/~mlearn/).

3.1 Analyzing the "Iris" dataset

In this example, we will use the most popular clustering algorithm – k-means [6] to produce the clustering result on the dataset "iris", and then import the result into VISTA system. With VISTA system, we will validate the k-means result visually and then try to refine the clusters to improve the quality of the k-means clusters. The quality of clusters is also evaluated by statistical indices RMSSTD, RS, and S_Dbw [11] to see if the statistical indices are consistent with the visual improvement.

"Iris" dataset is a famous dataset widely used in pattern recognition and clustering. It is a 4-D dataset containing 150 instances, and there are three clusters, each has 50 instances. One cluster is linearly separable from the other two; the latter two are not exactly linearly separable from each other according to the literature.

In initial visualization Figure 3-1, we can find one cluster has been separated from the other two. After interactive cluster rendering, mainly the α-parameter adjustment, the visual boundaries become clearer (Figure 3-2). The boundary B-C clearly separates cluster C from the other two clusters. The gap between cluster A and B can be visually perceived. The α-mapping model confirms that this gap does exist in the 4-D space. We make this gap as the visual boundary A-B. This visually perceived boundary A-B is not consistent with the k-means boundary, but we have more confidence with it since it has been intuitively confirmed. As the literature of the "iris" dataset mentioned, the two clusters are not linearly separable. To further refine the cluster definition, we can also informally define a small "ambiguous area" around the gap between A and B, the points in which have equal probability of belonging to A or B.

With this visual boundary, we can edit the k-means result as Figure 3-3 shows. After editing, the points are shown more homogeneously distributed in the clusters. The visual partition is also highly consistent with the real cluster distribution (comparing Figure 3-3 and 3-4). However, the statistical validity indices do not agree with the visual improvement. All of the three indices show the visual re-partitioning reduces the cluster quality. (Smaller RMSSTD, larger RS and the smaller S_Dbw imply the better quality. [11]).

Extended experiments with trained users show all users can find the visualization like Figure 3-2 in less than 2 minutes, which means visual validity could be very practical in exploring datasets. Experimental results on various datasets, showing the effectiveness and efficiency of visual validating and rendering, are not listed here, due to the space limitation.

3.2 Incorporating domain knowledge

Domain knowledge plays a critical role in the clustering process [8]. It is the semantic explanation to the data, which is different from the structural clustering criteria, such as distance between points and usually leads to a high-level cluster definition, for example, splitting or combining the parts of the basic clusters.

Domain knowledge can be represented in various forms [8]. In VISTA system, we define the domain knowledge as additional labeled items to the original dataset. The labels indicate the domain criterion about the clustering. We name the labeled items "landmarks". The number of landmarks is usually so small that they cannot work effectively as a training dataset to classify the entire datasets with classification algorithms.

When visualizing a dataset, the landmark points are loaded and visualized in different colors according to their labels. This guiding information can direct the user

to define the high-level cluster structure, or repartition the algorithmic clustering results. The alternative method is to visualize the dataset first and then sample some points from the "critical areas" on the visualization such as the connection/boundary area. The sample points then work as the "landmarks". It is very inefficient or clumsy to incorporate such functionality into automatic algorithms.

We use the "shuttle" dataset and the alternative method to demonstrate how the VISTA system incorporates the domain knowledge into the clustering process. "Shuttle" dataset is a 9-D dataset. It has three large clusters and some tiny clusters in irregular shapes. We use the testing dataset, which has 14500 items for visualization.

Several points are interactively picked from the critical areas in the initial visualization (Figure 4-1) working as the landmarks. Using the labels from the original datasets to mimic the domain expert, the "landmarks" show we could partition the dataset in the way of Figure 4-2.

4. Related work

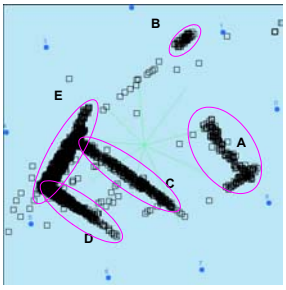

Figure 4-1: Initial

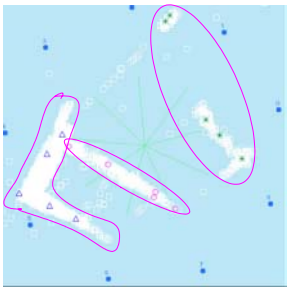
Figure 4-2: landmarks

The common framework of cluster analysis is described in the clustering review paper [14]. Recently, some algorithms [2][7][9][10][12][15]have been developed aiming at the arbitrarily shaped clusters. Some typical statistical validity indices are introduced in [11].

The early research on general plot-based data visualization is Grand Tour and Projection Pursuit [3]. L.Yang [4] utilizes the Grand Tour technique to show projections of datasets in an animation. Star Coordinates [5] is a visualization system designed to visualize and analyze the clusters interactively. We utilize the form of Star Coordinates and build the normalized α-mapping model in our system. HD-Eye [14] is another interactive visual clustering system based on density-plots of any two interesting dimensions. The 1D visualization based OPTICS [15] works well in finding the basic arbitrarily shaped clusters but lacks the ability in helping understand the inter-cluster relation. In the KDD 2002 tutorial [7], more visualization methods were also discussed.

5. Conclusion

Most of researchers have focused on automatic clustering algorithms, but very few have addressed the human factor in the clustering process, especially in dealing with arbitrarily shaped clusters. The VISTA system demonstrates some possible ways to introduce the users into the clustering process, and helps them validating and refining the clustering results visually.

Reference

[1] Keke Chen and Ling Liu: "Cluster Rendering of Skewed Datasets via Visualization". ACM Symposium on Applied Computing 2003, Melborne, FL.

[2] G.Guha, R.Rastogi, and K.Shim. "CURE: An efficient clustering algorithm for large databases", in Proc. of the 1998 ACM SIGMOD

[3] Cook, D.R, Buja, A., Cabrea, J., and Hurley, H. "Grand tour and projection pursuit", Journal of Computational and Graphical Statistics, V23, pp. 225-250

[4] Li Yang. "Interactive Exploration of Very Large Relational Datasets through 3D Dynamic Projections", in Proc. of SIGKDD2000

[5] E. Kandogan. "Visualizing Multi-dimensional Clusters, Trends, and Outliers using Star Coordinates", in Proc. of SIGKDD2001.

[6] A. Jain and R.Dubes. "Algorithms for Clustering Data", Prentice hall, Englewood Cliffs, NJ, 1988

[7] Grinstein G., Ankerst M., Keim D.A.: Visual Data Mining: Background, Applications, and Drug Discovery Applications", Tutorial at ACM SIGKDD2002.

[8] Jain, A.K., Murty, M.N. and Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys, 31(3), P264-323

[9] Ester, M., Kriegel, H., Sander, J. and Xu, X. "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"

[10] Hinneburg, A. and Keim, D. " An Efficient Approach to Clustering in Large Multimedia Databases with Noise", in Proc. of KDD-98, pp. 58-65

[11] Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis: "Cluster Validity Methods: Part I&II", SIGMOD Record, Vol31, No.2&3, 2002

[12]G.Sheikholeslami, S.Chatterjee, and A.Zhang. "Wavecluster: A multi-resolution clustering approach for very large spatial databases", In Proc. VLDB98', 1998

[13] Jean Gallier: "Geometric methods and applications: for computer science and engineering", Springer-Verlag, NY, c2001

[14] A. Hinneburg, D. Keim, and M. Wawryniuk: "Visual Mining of High-dimensional data", IEEE Computer Graphics and Applications. V19, No 5, 1999

[15] M. Ankerst, M. Breunig, H. Kriegel and J. Sander: "OPTICS: Ordering Points To Identify the Clustering Structure", in proc. of SIGMOD1999