Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging

Cheng-Ru Lin and Ming-Syan Chen, *Fellow*, *IEEE*

Abstract—Data clustering has attracted a lot of research attention in the field of computational statistics and data mining. In most related studies, the dissimilarity between two clusters is defined as the distance between their centroids or the distance between two closest (or farthest) data points. However, all of these measures are vulnerable to outliers and removing the outliers precisely is yet another difficult task. In view of this, we propose a new similarity measure, referred to as cohesion, to measure the intercluster distances. By using this new measure of cohesion, we have designed a two-phase clustering algorithm, called cohesion-based self-merging (abbreviated as CSM), which runs in time linear to the size of input data set. Combining the features of partitional and hierarchical clustering methods, algorithm CSM partitions the input data set into several small subclusters in the first phase and then continuously merges the subclusters based on cohesion in a hierarchical manner in the second phase. The time and the space complexities of algorithm CSM are analyzed. As shown by our performance studies, the cohesion-based clustering is very robust and possesses excellent tolerance to outliers in various workloads. More importantly, algorithm CSM is shown to be able to cluster the data sets of arbitrary shapes very efficiently and provide better clustering results than those by prior methods.

Index Terms—Data mining, data clustering, hierarchical clustering, partitional clustering.

1 Introduction

Data mining has attracted a significant amount of research attention due to its usefulness in many applications, including selective marketing, decision support, business management, and user profile analysis, to name a few [6], [7], [11], [14]. Among others, data clustering is one of the most active research areas [16], [18]. The data clustering techniques can be used to perform similarity search, pattern recognition, trend analysis, grouping, classification, and so forth [7], [14], [26].

In general, there are two types of attributes associated with input data in clustering algorithms, i.e., numerical attributes [1], [4], [25], [29], [31] and categorical attributes [13], [30]. Numerical attributes are those with a finite or infinite number of ordered values, such as the height of a person or the x-coordinate of a point on a 2D domain. On the other hand, categorical attributes are those with finite unordered values, such as the occupation or the blood type of a person. In this paper, we focus only on the clustering of numerical data.

Many data clustering algorithms have been proposed in the literature. These algorithms can be categorized into nearest-neighbor clustering [22], fuzzy clustering [3], partitional clustering [9], [20], [23], hierarchical clustering [19], [27], artificial neural networks for clustering [15], statistical clustering algorithms [8], [28], density-based clustering algorithm [5], [10], and so on. In theses methods, hierarchical and partitional clustering algorithms are two primary approaches in research communities. Hierarchical clustering algorithms can usually find satisfiable clustering results. A hierarchical clustering algorithm is able to obtain different clustering results for different similarity requirements. However, most of those hierarchical algorithms are very computationally intensive and require much memory space. Both of the two well-known hierarchical algorithms, i.e., *single-link* [27] and *complete-link* [19] algorithms, require time complexity of $O(n^2 \log n)$, where n is the number of input points. Algorithm CURE [12], which is an improvement of the single-link algorithm, though being able to lead to good clustering results, is, in essence, very costly in its computational overhead.

On the other hand, most partitional clustering algorithms run in linear time. With better efficiency, the clustering quality of a partitional algorithm is, however, not as good as that of hierarchical algorithms. The *k-means* clustering algorithm is one of the most famous partitional clustering algorithms [23]. Although widely used, the k-means algorithm suffers from some drawbacks, including: 1) dependency on the input order, 2) the tendency to result in local minimum, and 3) limited applicability to only the data set consisting of isotropic clusters (i.e., a circle in the 2D domain or a sphericity in the 3D domain).

Several clustering methods have been proposed to combine the features of hierarchical and partitional clustering algorithms. In general, these algorithms first partition the input data set into m subclusters. Then, these algorithms construct a hierarchical structure based on these m subclusters. As shown in Fig. 1, the data set is first partitioned into 15 subclusters and these subclusters are next grouped into two clusters.

• *The authors are with the Electrical Engineering Department and Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, ROC. E-mail: {owenlin, mschen}@arbor.ee.ntu.edu.tw.*

Manuscript received 16 Aug. 2002; revised 21 Aug. 2003; accepted 9 Mar. 2004; published online 17 Dec. 2004.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 117145.

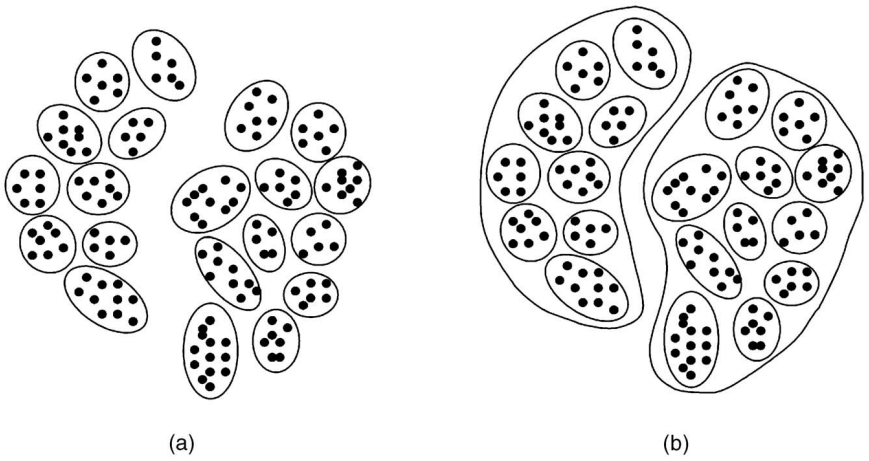

Fig. 1. An illustration of a hybrid clustering algorithm. (a) Step 1: Form several small clusters. (b) Step 2: Merge small subclusters into clusters.

This hybrid idea of clustering is first proposed in [24] where a multilevel algorithm is developed. At the first level, the multilevel algorithm partitions the data set into several partitions and then performs the k-means algorithm on each partition to obtain several subclusters. In subsequent levels, this algorithm uses the centroids of the subclusters identified in the previous level as the new input data points and performs the hierarchical clustering algorithm on those points. This process continues until exactly k clusters are determined. Finally, the algorithm performs a top-down process to reassign all points of each subcluster to the cluster of their centroids. However, representing a subcluster by only one point makes this multilevel algorithm not applicable to some cases, especially when the sizes of those subclusters are similar to that of the merged cluster.

Algorithm BIRCH is one of the most efficient clustering algorithms [31]. In BIRCH, only one scan is needed to obtain good clustering results. One or more additional passes can be used to further improve the clustering qualities. Algorithm BIRCH first partitions the input data set into many small subclusters and then applies a global clustering algorithm on those subclusters to achieve the final results. The main contribution of BIRCH is as an efficient data preprocessor for large input data sets. It does not focus on the design of global clustering algorithm, which can be effectively applied on subclusters. In view of this, a new definition of distances between two data bubbles (subclusters) are proposed in [5]. The algorithm proposed first executes BIRCH to obtain many small data bubbles and then applies a modified algorithm of OPTICS with the new definition of distance between data bubbles.

In most related studies, the dissimilarity between two clusters is defined as the distance between their centroids or the distance between two closest (or farthest) data points. However, these measures are vulnerable to outliers and removing outliers precisely is yet another difficult task. This is the very reason that most prior clustering methods do not perform stably for various types of inputs. In view of this, we propose a new similarity measure, referred to as cohesion, to measure the intercluster distances. Using cohesion, we have designed a two-phase clustering algorithm, called cohesion-based self-merging (abbreviated as CSM), which runs in time linear to the size of input data set. Combining the features of partitional and hierarchical clustering methods, algorithm CSM partitions the input data set into several small subclusters in the first phase and then continuously merges the subclusters based on cohesion in a hierarchical manner in the second phase. The time and the space complexities of algorithm CSM are analyzed. Our performance studies show that the proposed similarity measure, cohesion, is more effective than others. With cohesion, algorithm CSM is not only very robust to the existence of outliers, but also able to lead to better clustering results than most prior methods while incurring much shorter execution times.

The rest of the paper is organized as follows: Section 2 presents the preliminaries. Algorithm CSM is presented in Section 3. Performance studies are conducted in Section 4. This paper concludes with Section 5.

2 Preliminaries

Given a desired number of clusters k, data clustering is the process of partitioning the input data points into k clusters so that the points in each cluster are similar to one another and are different from the points in other clusters. For example, for the data set shown in Fig. 2a, a data clustering algorithm with $k = 2$ may partition these points into two clusters, $\{A, B, C, D\}$ and $\{E, F, G\}$. We review several prior algorithms related to this study in the following.

2.1 Hierarchical Clustering Algorithms

As its name implies, a hierarchical clustering algorithm establishes a hierarchical structure as the clustering result. Consider the example in Fig. 2a. One possible hierarchical structure is shown in Fig. 2b. With the hierarchical structure, we can obtain different clustering results for different similarity requirements. As shown in Fig. 2b, if the similarity requirement is set at *level 1*, the input data set is partitioned into two clusters, i.e., $\{A, B, C, D\}$ and $\{E, F, G\}$. However, if the similarity requirement is set at *level 2*, then the input data set is partitioned into six clusters, i.e., $\{A, B\}$, $\{C\}$, $\{D\}$, $\{E\}$, $\{F\}$, and $\{G\}$.

Most existing hierarchical clustering algorithms are variations of the single-link and complete-link algorithms. Both algorithms require time complexity of $O(n^2 \log n)$,

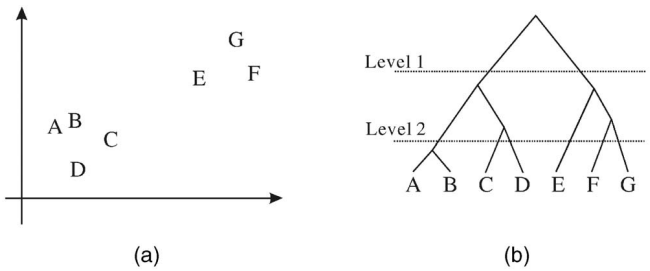

Fig. 2. Illustrative hierarchical clustering results for a data set of seven points. (a) The input data set. (b) A possible hierarchical tree.

where n is the size of the input data set. Owing to their good quality of clustering results, hierarchical algorithms are widely used, especially in document clustering and classification. The outline of a general hierarchical clustering algorithm is given below:

Hierarchical Clustering Algorithm

1. Initially, each data point forms a cluster by itself.
2. The algorithm repetitively merges the two closest clusters.
3. Output the hierarchical structure constructed.

A single-link clustering algorithm differs from a complete-link clustering algorithm in the intercluster distance measure, i.e., Step 2. The single-link algorithm uses the distance between the two closest points of the two clusters as the intercluster distance, i.e.,

$$dist(C_i, C_j) = \min\{dist(o_i, o_j) | o_i \in C_i, o_j \in C_j\},$$

while the complete-link algorithm uses the distance of two farthest points as the intercluster distance, i.e.,

$$dist(C_i, C_j) = \max\{dist(o_i, o_j) | o_i \in C_i, o_j \in C_j\}.$$

As shown in Fig. 3a, the single-link algorithm suffers from the so-called chaining effect and the complete-link clustering algorithm has problems in dealing with particular shapes, such as the circles shown in Fig. 3b.

In the single-link algorithm, we can maintain a pointer to the closest neighboring cluster for each cluster. After merging Cluster C_j into Cluster C_i, the single-link algorithm needs only to update those clusters whose closest neighbor is C_j and change it to C_i. With this implementation, the space complexity required by the single-link clustering algorithm is only $O(n)$. However, the same mechanism cannot be applied to the complete-link algorithm. Instead, the complete-link algorithm needs to keep all the distances of any two points and, thus, is of space complexity $O(n^2)$.

Algorithm CURE [12] is an improvement over the single-link clustering algorithm. CURE selects several scattered data points carefully as the representatives for each cluster and shrinks these representatives toward the centroid in order to eliminate the effects of outliers and avoid the chaining effect. The distance between two clusters in CURE is defined as the minimal distance between the two representatives of each cluster. In each iteration, it merges the two closest clusters.

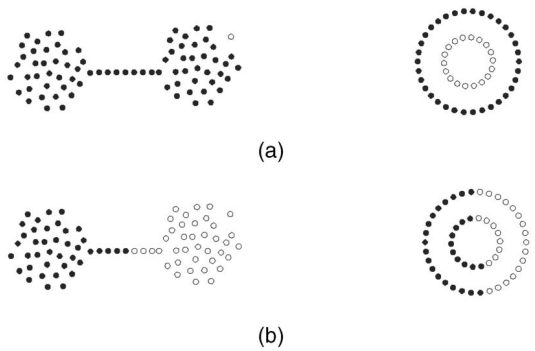

Fig. 3. Examples of the clustering capability of the single-link and complete-link clustering algorithms. (a) Clustering results by the single-link algorithm. (b) Clustering results of the complete-link algorithm.

It is known that CURE is one of the most successful clustering methods for clustering the data of any shape. However, it is very computationally intensive. The time complexity of algorithm CURE is essentially $O(n^2 \log n)$. In addition, algorithm CURE utilizes a spatial searching data structure, kd-tree, to search the nearest representatives. As pointed out in [2], this kind of searching structure does not work well in a high-dimensional data set. This fact limits the applicability of CURE. The sampling and partitioning technique adopted by algorithm CURE may accelerate the computation. However, most sampling techniques are orthogonal to clustering algorithms, i.e., a clustering algorithm can usually be easily integrated with most sampling techniques by no or slight modification. Please also note that, while being widely used, random sampling suffers from the problem of missing small clusters. In view of this, a density biased sampling algorithm is proposed in [10]. In this paper, we will focus on the clustering problem.

2.2 Partitional Clustering Algorithms

The k-means algorithm is one popular partitional algorithm for data clustering. However, the k-means algorithm suffers from the tendency of resulting in local minimum and is likely to obtain different clustering results with different initial states. The clustering results are also dependent on the sequence of the input data. The adoption of Euclidean distance as the similarity measure causes the application of the k-means algorithm to be limited to data sets consisting of only isotropic clusters, thus refraining it from being used in many real applications. Nevertheless, most partitional algorithms have the advantages on the execution time and the space required. The outline of the k-means algorithm is given as follows:

Algorithm K-Means

1. Initially, select k centroids arbitrarily for each cluster C_i, $i \in [1, k]$.
2. Assign each data point to the cluster whose centroid is closest to the data point.
3. Calculate the centroid c_i of cluster C_i, $i \in [1, k]$.
4. Repeat Step 2 and Step 3 until no points change between clusters.

2.3 Expectation Maximization Algorithms

In addition to using distance as the similarity measure, the Expectation Maximization (EM) algorithm uses probabilities to measure the similarities. It is assumed that the points of a cluster follow a certain distribution [8]. By assuming the parameters of the distribution of each cluster, EM utilizes probability to judge which cluster a data point should be assigned to. Algorithm EM then adjusts the parameters of each cluster's distribution according to the data points in that cluster. Next, it reassigns these points according to the new distributions. These iterations continue until the clustering results converge. For example, if the distribution of Cluster C_i follows a given probability density function (abbreviated as *pdf*) $f_{C_i}(v)$, then the probability for a point with position v to belong to this cluster is:

$$P(C_i|v) = \frac{P(v|C_i) \times P(C_i)}{P(v)} = \frac{P(C_i)}{P(v)} f_{C_i}(v).$$

If a point at location v is more likely to belong to Cluster C_i than to Cluster C_j, i.e., $P(C_i|v) > P(C_j|v)$, then this point will be assigned to Cluster C_i.

2.4 Hybrid Clustering Algorithms

Since the complexity of hierarchical data clustering is relatively high, several improved algorithms have been proposed, such as the hybrid clustering algorithm proposed in [24] and algorithm BIRCH [31]. The hybrid algorithm in [24] is a multilevel algorithm. In the first level, the input data set is divided into P_1 partitions equally. Then, the hybrid algorithm performs a k-means algorithm to get C_1 clusters in each partition. In level i, where $i > 1$, the centroids of the clusters obtained in level $i-1$ are taken to this level for merging. These centroids are partitioned into P_i partitions. In each partition, the hybrid algorithm performs a hierarchical clustering algorithm to get C_i clusters. This process continues until exactly k clusters are identified. Finally, the algorithm performs a top-down process to reassign all points of each subcluster to the cluster of their centroid. This algorithm is shown to be very efficient in both aspects of computation and memory space. However, using only one point to represent the whole cluster may easily lose some information about the distributions of clusters, which are important to the similarity of two clusters.

Another hybrid clustering algorithm, BIRCH, is designed to deal with large input data sets. BIRCH uses *cluster features* (CF) to represent a subcluster. Given the CF of a subcluster, one can obtain the *centroid*, *radius*, and *diameter* of that subcluster easily (in constant time). Furthermore, the CF vector of a new cluster formed by merging two subclusters can be directly derived from the CF vectors of the two subclusters by algebra operations. Algorithm BIRCH consists of four phases. In Phase 1, BIRCH partitions the input data set into many subclusters by a CF tree. In Phase 2, it reduces the size of the CF tree (i.e., the number of subclusters) in order to apply a global clustering algorithm in Phase 3 on those generated subclusters. In Phase 4, each point in the data set is redistributed to the closest centroids of the clusters produced in Phase 3. Among these phases, Phase 2 and Phase 4 are used to further improve the clustering quality and are thus optional. Algorithm BIRCH does not specify the global clustering algorithm. However, if a clustering algorithm, which can find clusters in any shape, is adopted in Phase 3, then Phase 4 will need some slight modifications to keep the shape information. More specifically, each cluster produced in Phase 3 should be represented by several points instead of only the centroid. However, the main contribution of BIRCH is as an efficient data preprocessor for a large input data set so that the global clustering algorithm can be executed efficiently.

Algorithm CHAMELEON is also a hybrid clustering algorithm [17]. The outline of the algorithm is as follows:

Algorithm CHAMELEON

1. Construct a k-nearest neighbor graph.
2. Partition the k-nearest neighbor graph into many small subclusters.
3. Merge those subclusters into final clustering results.

Note that the k-nearest-neighbor graph is built by connecting each node with its k nearest-neighboring nodes. Different from most of clustering algorithms, it considers not only the interconnectivity (the number of links between two clusters) but also the closeness (the length of those links) as the similarity measure of two clusters. Their experiments also show the excellent clustering quality. However, the time complexity of building a k-nearest-neighbor graph of a high-dimensional data set is as high as $O(n^2)$, which makes CHAMELEON infeasible for large data sets.

3 Cohesion-Based Self-Merging Algorithm

In this section, we describe the details of the cohesion-based self-merging algorithm (abbreviated as CSM). In Section 3.1, we propose a new measure for the similarity of two subclusters, *cohesion*, which is intrinsically different from other prior measures. Cohesion is more appropriate for an intercluster similarity measure because it does not judge the similarity of two subclusters by only some data points. Rather, the cohesion measure takes the distributions of the two clusters into account. In Section 3.2, based on cohesion, we devise a clustering algorithm, CSM, which fully utilizes the features of cohesion. We also describe a general outlier removal mechanism in Section 3.3 to enable algorithm CSM to resist outliers. (Note that the colored versions of some figures in this paper can be found at http://arbor.ee.ntu.tw/~owenlin/tkde_csm/.)

3.1 Similarity Measure between Subclusters

As shown in Fig. 4, the distance between the centroids of the two clusters in Fig. 4a and that of the two clusters in Fig. 4b are the same. The two clusters shown in Fig. 4b, however, are more inclined to be merged together. In addition to the distance between centroids, the average complete distance (i.e., $\frac{\sum d(p_i - p_j)}{|C_i| \times |C_j|}$ for each $p_i \in C_i$ and $p_j \in C_j$) is another method to measure the intercluster similarity of two clusters. However, with much more computation, the average complete distance cannot even distinguish between the two cases shown in Fig. 4. The average complete distance is 20.40 in Fig. 4a and 21.57 in Fig. 4b. (In contrast, as can be verified later, the corresponding cohesions are 2.21×10^{-11} in Fig. 4a and 2.17×10^{-4} in

(a) (b)

Fig. 4. Illustration for subclusters that have different cohesion values. (a) Two clusters with smaller cohesion. (b) Two clusters with greater cohesion.

Fig. 4b.) Several alternatives, such as the distance between the closest (or farthest) points of the two clusters, could be employed to redeem this deficiency. However, those measures are very vulnerable to random noises (outliers).

Consequently, we propose a new similarity measure, namely, *cohesion*, based on the *joinability* of two clusters, referring to the existence of a data point. Conceptually, *joinability* is the merging inclination of two clusters according to the existence of a shared data point. Thus, *joinability* is expected to have the following properties: 1) Data points located closer to the boundary of the two clusters are more important and 2) the merging inclination should not be determined by only a few points, i.e., the value of *joinability* should not vary dramatically. Formally, we have the following definition for *joinability*:

Definition 1. *The* joinability *of the two clusters (C_i and C_j) referring to the existence of the point p with location v is defined as:*

$$join(p, C_i, C_j) = \min\big(f_i(v), f_j(v)\big),$$

where f_i and f_j are the probability (density) function of the distributions in Cluster C_i and C_j, respectively.

An illustration of *joinability* is shown in Fig. 5. In Fig. 5, the joinabilities of the data points at v_1 and v_2 are j_1 and j_2, respectively. With the notion of *joinability*, the definition of *cohesion* of two clusters is given below:

Definition 2. *The* cohesion *of two clusters (C_i and C_j) is defined as*

$$chs\big(C_i, C_j\big) = \frac{\sum_{p \in C_i, C_j} join\big(p, C_i, C_j\big)}{|C_i| + |C_j|},$$

where $|C_i|$ is the size of Cluster C_i.

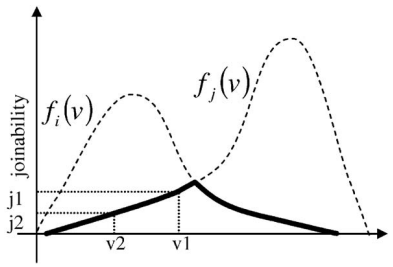

Fig. 5. An illustration for the meaning of *joinability*.

In general, the *pdf* $f(v)$ can be evaluated in constant time. Thus, the time complexity of the computation of cohesion of two clusters (C_i and C_j) is linear to the size of the two clusters, i.e., $O\big(|C_i| + |C_j|\big)$.

In this paper, we assume the location of a point in each cluster follows a multivariate normal distribution, i.e., $V \sim N_d(\mu, \psi)$, where d is the dimension of the space, μ is the mean vector, and ψ is the covariance matrix. The probability density function is

$$f(v) = (2\pi)^{-\frac{d}{2}} (\det \psi)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\Delta^2(v)\right],$$

where

$$\Delta^2(v) = (v - \mu)^T \psi^{-1} (v - \mu).$$

Please note that, in this formula, the location of a point (i.e., v_i) and the mean vector (i.e., μ) are both d-variate vectors and the covariance matrix (i.e., ψ) is a positive, definite $d \times d$ matrix.

Since the values of the mean vector and covariance matrix are unknown in advance, we use the maximum likelihood estimator of (μ, ψ) in practice. Given a cluster of n points with locations $(v_1, v_2, \ldots, v_n)$, the values of (μ, ψ) of the cluster are estimated by the following formulae:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} v_i$$

and

$$\psi = \frac{1}{N}\sum_{i=1}^{n} (v_i - \widehat{\boldsymbol{\mu}})(v_i - \widehat{\boldsymbol{\mu}})^T.$$

This similarity measure of cohesion is robust to the existence of outliers due to the following two reasons: 1) Using the cohesion measure, instead of only a few of points, all points of the two subclusters are considered to evaluate this intercluster similarity and 2) this cohesion measure makes the effect of outliers much smaller than that of other points since outliers are much farther from the centroid of the two clusters. For example, for the four pairs of clusters shown in Fig. 6, cohesions between these pairs of clusters are 2.99×10^{-9} in Fig. 6a, 1.63×10^{-6} in Fig. 6b, 5.23×10^{-8} in Fig. 6c, and 8.52×10^{-9} in Fig. 6d. It can be noted that, comparing with the cohesion value in Fig. 6b, i.e., 1.63×10^{-6}, the cohesions of Fig. 6b and Fig. 6c are similar to the original one, i.e., cohesion is robust to the effects of outliers.

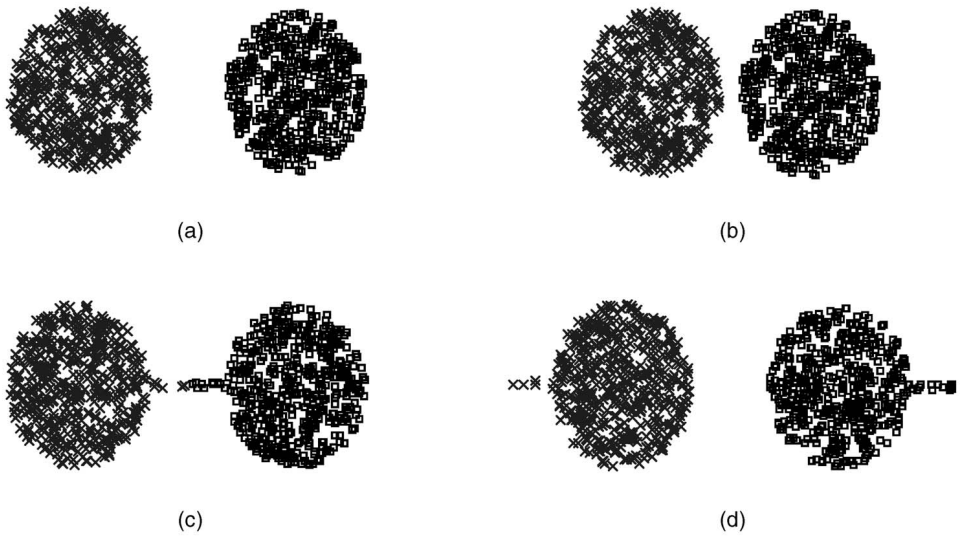

Fig. 6. An illustration of the robustness of the cohesion measure against outliers. (a) Two clusters. (b) Two closer clusters. (c) Two clusters with a link between them. (d) Two clusters with tails.

TABLE 1
All the Cohesions in Example 3.2

	Cl_1	Cl_2	Cl_3	Cl_4	Cl_5	Cl_6
Cl_1	–	2.32×10^{-25}	2.05×10^{-60}	4.88×10^{-21}	5.34×10^{-5}	3.74×10^{-5}
Cl_2	2.32×10^{-25}	–	6.31×10^{-5}	3.32×10^{-5}	6.36×10^{-12}	3.91×10^{-11}
Cl_3	2.05×10^{-60}	6.31×10^{-5}	–	4.11×10^{-5}	3.33×10^{-23}	2.51×10^{-19}
Cl_4	4.88×10^{-21}	3.32×10^{-5}	4.11×10^{-5}	–	5.82×10^{-12}	2.92×10^{-13}
Cl_5	5.34×10^{-5}	6.36×10^{-12}	3.33×10^{-23}	5.82×10^{-12}	–	4.08×10^{-5}
Cl_6	3.74×10^{-5}	3.91×10^{-11}	2.51×10^{-19}	2.92×10^{-13}	4.08×10^{-5}	–

3.2 Algorithm CSM

Now, we describe the proposed algorithm based on cohesion self-merging as follows:

Algorithm CSM

//Input: The input data set, the size of the data set, n, the number of subclusters, m, and the desired number of clusters, k.
//Output: The hierarchical structure of the k clusters.

1. Apply k-means on the input data set to obtain m subclusters.
2. Apply the single-link clustering algorithm on the m subclusters produced in Step 1 with cohesion as the similarity measure and stop when k clusters are obtained.

Example 3.2. Consider the data set shown in Fig. 7. We apply algorithm CSM with $k = 2$ and $m = 6$ to it. In the first phase, as shown in Fig. 7, algorithm CSM partitions the data set into six subclusters. In the second phase, algorithm CSM tries to merge the six subclusters into two clusters by their cohesions. The cohesions between these subclusters are shown in Table 1.

Since the largest cohesion is $chs(Cl_2, Cl_3) = 6.31 \times 10^{-5}$, we first merge cluster Cl_2 and cluster Cl_3. Then, we find the next large cohesion, i.e., $chs(Cl_1, Cl_5) = 5.34 \times 10^{-5}$, and merge cluster Cl_1 and cluster Cl_5. Similarly, we merge Cl_3 and Cl_4 and then Cl_5 and Cl_6. Finally, the two clusters reached by CSM are $\{Cl_1, Cl_5, Cl_6\}$ and $\{Cl_2, Cl_3, Cl_4\}$.

Fig. 7. An illustrative example of algorithm CSM.

Algorithm CSM is a two-phase clustering algorithm. In the first phase, it adopts the k-means algorithm to divide the input data set into m subclusters. At the beginning of Phase 2, it obtains the cohesions of these m subclusters produced in the first phase. Then, algorithm CSM performs a single-link clustering algorithm based on cohesion to obtain the final clusters.

In algorithm CSM, the parameter m, i.e., the number of subclusters, is the only additional parameter. We can obtain desired clustering results by adjusting the value of m. It is clear that the value of parameter m falls in the range of $[k, n)$. When $m = k$, algorithm CSM is degenerated to the k-means clustering algorithm. When m approaches n, this algorithm is reduced to the single-link algorithm. It is known that the k-means algorithm is good for obtaining clusters of isotropic shape, while the single-link algorithm is able to find clusters of any shape. With prior knowledge, we can make the clustering algorithm adapt to various inputs by adjusting the parameter m. As will be validated

by our performance studies later, algorithm CSM, in general, leads to better clustering results than those by most prior methods, showing not only the generality but also the advantage of algorithm CSM over the k-means and single-link methods.

3.3 Complexity Analysis

Theorem 1. *The time complexity of algorithm CSM is $O(mnl + m^2 \log m)$, where l is the number of iterations in the k-means algorithm.*

Proof. In Phase 1, algorithm CSM takes time of complexity $O(nml)$ to apply the k-means algorithm on the input data set to obtain m subclusters. Then, at the beginning of Phase 2, the values of cohesion between any two subclusters are evaluated. Recall that the time complexity of each evaluation is linear to the size of the clusters. Thus, the time complexity is

$$O\left(\sum_i \sum_{j>i} \left(|C_i| + |C_j|\right)\right) = O\left(\frac{1}{2}\sum_i \sum_{j\neq i} \left(|C_i| + |C_j|\right)\right)$$
$$= O\left(\frac{1}{2}\sum_i \left((m-2)|C_i| + \sum_j |C_j|\right)\right)$$
$$= O\left(\frac{1}{2}n(m-1)\right) = O(nm).$$

Next, algorithm CSM applies the single-link clustering algorithm to those subclusters and has time in order of $O(m^2 \log m)$. The time complexity of algorithm CSM is therefore $O(mnl + nm + m^2 \log m) = O(mnl + m^2 \log m)$. □

Note that the number of main iterations (i.e., l) in algorithm k-means is data dependent and determined empirically. However, experiments reveal that the k-means algorithm takes only a few iterations to converge.

Theorem 2. *The space complexity of the algorithm is $O(n)$.*

Proof. In the first phase, we only need the memory space to store the input data set and the group relationship of subclusters, which requires the memory space of complexity $O(n)$. In the second phase, memory space is required for the single-link clustering algorithm. Thus, the space complexity is $O(m)$. Since m is always smaller than n, the total space complexity of algorithm CSM is $O(n)$. □

3.4 Resilience to Outliers

Outliers are those random points that are very different from others and do not belong to any clusters. In algorithm CSM, we adopt cohesion as our similarity measure to resist the effects of outliers within a subcluster. However, it is possible that some subclusters consist only of outliers. We will get those noise clusters especially when we partition the input data set into too many subclusters. Those noise clusters will affect the correctness of the subsequent merging. Consider an example shown in Fig. 8. (One may note that the data sets in these figures are not identical to one another. This is because the data sets have been properly sampled in order to be clearly displayed.) In this example, algorithm CSM first partitions the input data set into 128 subclusters and then start to merge closest clusters pair by pair. The clustering process works well until those subclusters are merged into 12 clusters (as shown in Fig. 8a). However, in the next step, the upper two clusters are merged together undesirably (as shown in Fig. 8b) due to the existence of those noise clusters. Finally, those subclusters are merged into five clusters, as shown in Fig. 8c. However, after applying the noise removal mechanism introduced in this section, we can correctly identify the five clusters, as shown in Fig. 8d.

As shown in Fig. 9, there are two kinds of noise clusters, i.e., sparse noises and dense noises. Note that those sparse noise clusters are likely to be merged with others and, thus, may become bridges between subclusters which should be separated. For example, the sparse noise cluster shown in Fig. 9 may be merged with Cluster A and Cluster B. It also should be noted that the density of clusters could vary in different regions in a large data set. The noise clusters in one region are possibly even denser than some normal clusters in another region. Thus, it is hard to identify those sparse noise clusters. Instead of trying to identify those noises and remove them immediately, algorithm CSM first makes those sparse noises harder to join into normal clusters and then they can be removed with those dense noises together. (It is assumed that the number of points in a noise cluster is much less than that of a normal cluster.) Algorithm CSM adopts *density impedance* to achieve this. More specifically, given two clusters C_i and C_j, the *density impedance* of the two clusters is defined as

$$impedance(C_i, C_j) = \frac{C_i.density() + C_j.density()}{2\sqrt{C_i.density() \times C_j.density()}}.$$

Algorithm CSM then defines the similarity of two clusters as

$$\frac{cohesion(C_i, C_j)}{impedance(C_i, C_j)^{\alpha}},$$

where α is named the *impedance factor* and is specified by users to control the effect of impedance. Then, when those m subclusters are merged into m' subclusters, where m' is a predefined number such that $k < m' < m$, algorithm CSM will try to remove those sparse and dense noises clusters. The value of density impedance is only related to the density ratio of the two clusters. The relationship between them is shown in Fig. 10.

In order to quantify the meaning of sparseness, we first define the volume and the density of a cluster. Recall that the *pdf* of a multivariate normal distribution ($N_p(\mu, \psi)$) is

$$f(v) = (2\pi)^{-\frac{d}{2}} (\det \psi)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\Delta^2(v)\right],$$

where

$$\Delta^2(v) = (v-\mu)^T \psi^{-1} (v-\mu).$$

Thus, the contours of the *pdf* will satisfy the condition

$$(v-\mu)^T \psi^{-1} (v-\mu) = c.$$

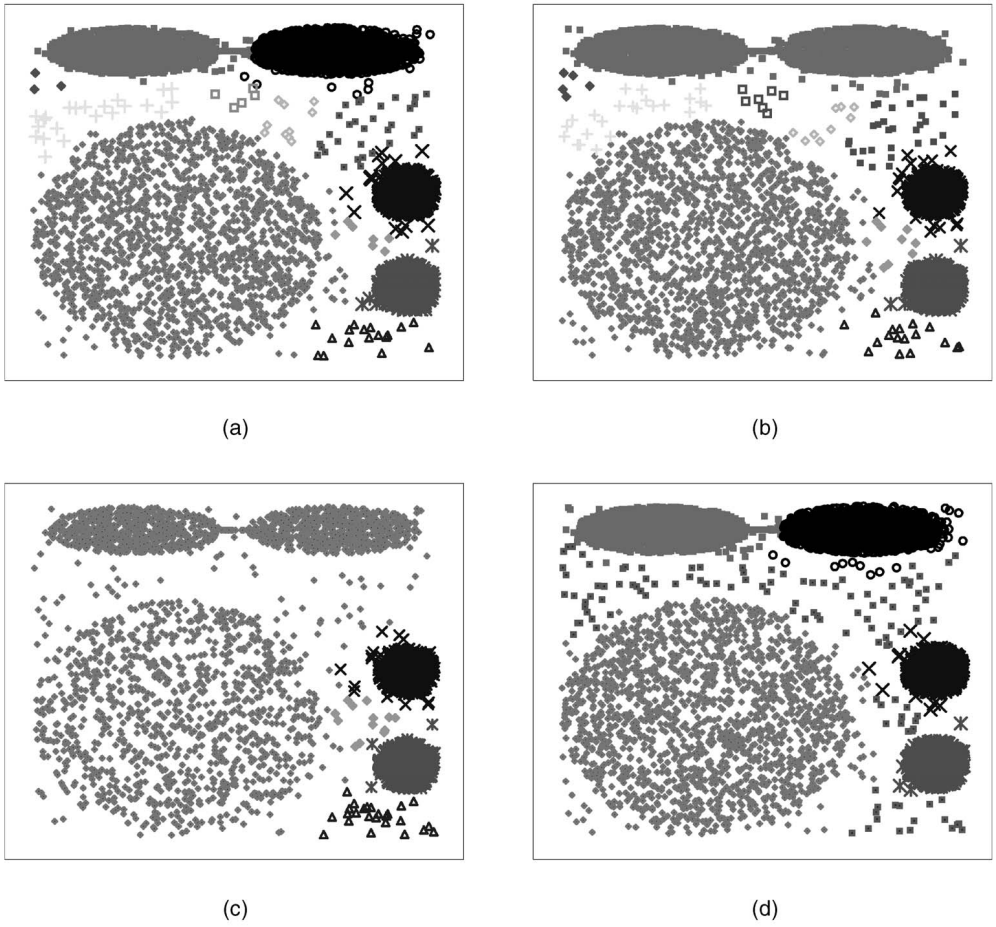

Fig. 8. An illustrative example of the effects of noise clusters. (a) Intermediate clustering state of 12 subclusters (five normal clusters and seven noise clusters). (b) Intermediate clustering state of 11 subclusters. (c) The final clustering result (three clusters are mixed together). (d) The clustering result after removing those noise clusters.

For simplicity, we assume $\mu = \mathbf{0}$ in the following discussion. Since ψ is a $d \times d$ symmetric matrix, there exist d real eigenvalues, i.e., $\lambda_1, \lambda_2, ..., \lambda_d$, and associated eigenvectors, i.e., $\beta_1, \beta_2, ..., \beta_d$, of unit length that can be chosen to be mutually orthogonal to one another. Next, making a $d \times d$ matrix $\mathbf{B}$ as $(\beta_1, \beta_2, ..., \beta_d)$, we have $\mathbf{B}^T\mathbf{B} = \mathbf{I}_d$, i.e., $\mathbf{B}^T = \mathbf{B}^{-1}$. Then, we change the coordinate in such a way that the new coordinate v' of vector v will satisfy the equation: $v = \mathbf{B}v'$. Note that this is a rigid transformation, i.e., all the angles and lengths will be preserved after the transformation. According to this transformation, we can present the function of the contour as $v'^T(\mathbf{B}^T\psi^{-1}\mathbf{B})v' = c$. Since matrix $\mathbf{B}$ consists of eigenvectors, we have

$$(\mathbf{B}^T\psi^{-1}\mathbf{B}) = \begin{bmatrix} \lambda_1^{-1} & 0 & 0 & 0 \\ 0 & \lambda_2^{-1} & 0 & 0 \\ & & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d^{-1} \end{bmatrix}.$$

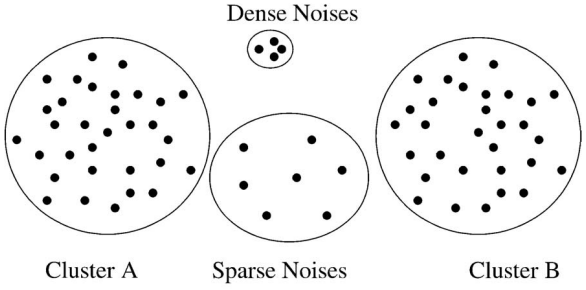

Fig. 9. The two kinds of noise clusters generated in Phase 1 of algorithm CSM.

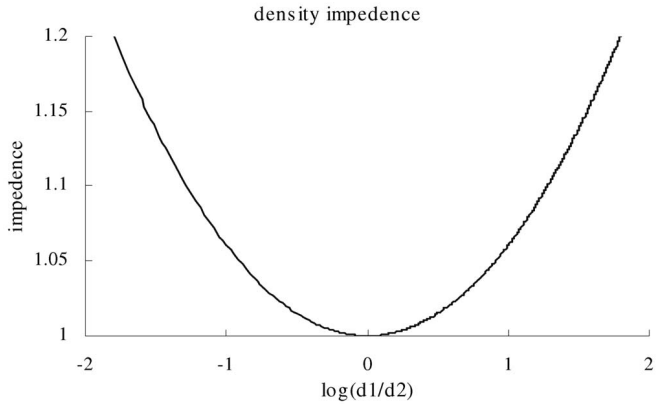

Fig 10. The density impedence of two clusters with density as d_1 and d_2.

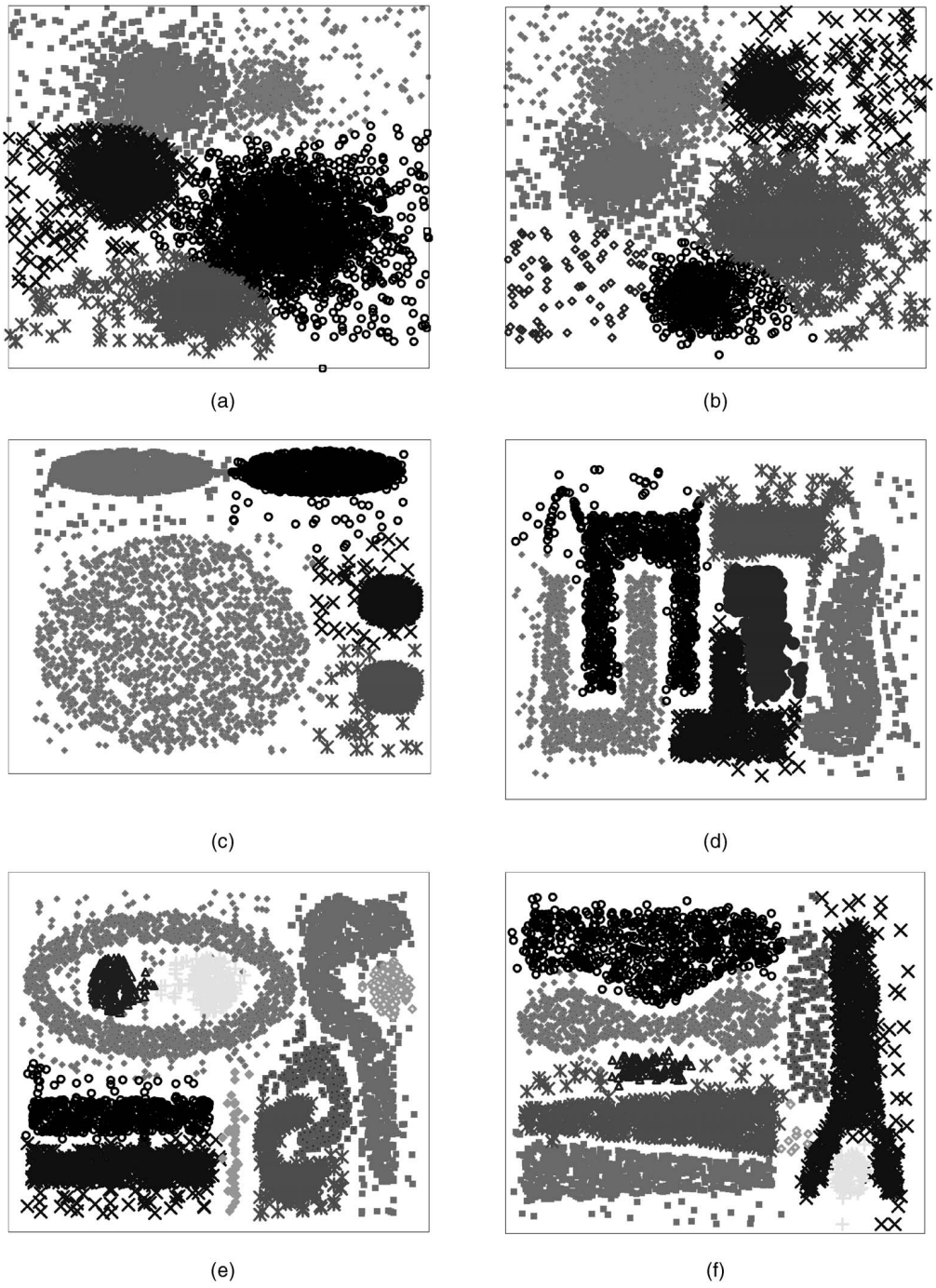

Fig. 11. The clustering results of algorithm CSM on the four input data sets. (a) CSM on Data Set 1. (b) CSM on Data Set 2. (c) CSM on Data Set 3. (d) CSM on Data Set 4. (e) CSM on Data Set 5. (f) CSM on Data Set 6.

As a result, the contour is a hyperellipse with $\sqrt{\lambda_1}$, $\sqrt{\lambda_2}, ...\sqrt{\lambda_d}$ as the length of each axis. Therefore, the volume and density of a cluster can be defined as follows:

Definition 3. *The volume of Cluster C_i is defined as*

$$C_i.volume = \sqrt{\lambda_1\lambda_2\ldots\lambda_d} = \sqrt{\det\left(\mathbf{B}^T\psi_i\mathbf{B}\right)} = \sqrt{\det\mathbf{B}^T\det(\psi_i)\det\mathbf{B}} = \sqrt{\det(\psi_i)},$$

where ψ_i is the covariance matrix of cluster C_i.

Definition 4. *The density of Cluster C_i is defined as*

$$C_i.density() = \frac{|C_i|}{C_i.volume},$$

where $C_i.volume$ is the volume of C_i and $|C_i|$ is the number of points in Cluster C_i.

The outlier removal mechanism is formally stated as follows. Note that the impedance factor, α, and $size\_ratio$ are two parameters specified by users.

TABLE 2
Execution Details on Each Data Set by CSM

	n	m	k	Average execution time
Data Set 1	10000	10	5	0.75 sec.
Data Set 2	10000	10	5	0.82 sec.
Data Set 3	10000	16	5	1.82 sec.
Data Set 4	8000	48	6	3.18 sec.
Data Set 5	10000	64	9	5.51 sec.
Data Set 6	8000	96	8	6.13 sec.

Outlier Removal Procedure

1. At Phase 2, algorithm CSM merges those subclusters by the similarity defined as
$$\frac{cohesion(C_i, C_j)}{impedance(C_i, C_j)^{\alpha}}.$$
2. When those m subclusters are merged into m' subclusters, algorithm CSM removes all the subclusters whose sizes are less than the threshold defined as
$$size\_ratio \times \frac{1}{m'} \sum |C_i|.$$

With proper parameter settings, algorithm CSM can precisely remove those noise clusters and result in clustering of better quality.

4 Performance Studies

To assess the performance of algorithm CSM, we have conducted a series of experiments. These experiments are performed on a computer with an 800 Mhz Intel CPU and 1.7G of memory. Several similarity measures are evaluated in Section 4.1. In Section 4.2, we compare the clustering quality of CSM with several prior clustering methods.

4.1 Experiment I: Comparing with Other Measures

We perform our experiments on the data sets shown in Fig. 11. Data Set 1 is generated from normal distribution and the sizes of clusters follow the Zipf distribution. Data Set 2 is in the same layout of Data Set 1. However, the density of clusters in Data Set 2 varies, while the density of clusters in Data Set 1 is uniform. Data Set 3 is the same one used in CURE [12]. As stated in [12], both BIRCH and single-link cannot partition this data set correctly. The other data sets are obtained from [17]. As stated in [17], both DBScan and CURE cannot successfully partition some of the data sets. The clustering results of algorithm CSM are shown in Fig. 11, while the details are shown in Table 2. Note that the values of parameter m are chosen so that algorithm CSM can obtain the similar clustering results each time. Recall that the algorithm is randomized. Thus, we execute the algorithm 20 times to obtain the average execution time. As shown in these figures, algorithm CSM is able to successfully partition these data sets. Note that the clustering results shown in this paper have been properly sampled in order to be clearly displayed. However, all algorithms are performed on the whole data sets.

First, cohesion is compared with other similarity measures, including those measures ($D0$, $D1$, $D2$, $D3$, $D4$) defined in [31] and the distance of data bubbles defined in [5]. We also define a new similarity measure based on the density decrement and obtain an improvement over the distance of data bubbles. These similarity measures are briefly defined below:

Definition 5. *The similarity measures $D0$, $D1$, $D2$, $D3$, and $D4$ between clusters C_i and C_j are defined in Table 3.*

In the definitions presented in Table 3, the notation c_i is the centroid of Cluster C_i, $c_i[k]$ is the value of c_i in the kth coordinate, $|C_i|$ means the number of points in Cluster C_i, and $var(C_i)$ is the variance of Cluster C_i, which is defined as $\sum_{p \in C_i} |p - c_i|^2$.

Definition 6. *The distance between two data bubbles in Euclidean vector space is defined as*

$$dist(C_i, C_j) = \begin{cases} |c_i - c_j| - (e_i + e_j) + nnDist(1, C_i) + nnDist(1, C_j) \\ \qquad \text{if } |c_i - c_j| - (e_i + e_j) \geq 0, \\ \max(nnDist(1, C_i),\ nnDist(1, C_j)) \text{ otherwise,} \end{cases}$$

where c_i is the centroid of Cluster C_i, e_i is extent of Cluster C_i, and $nnDist(k, C_i)$ is a function of the estimated average k-nearest-neighbor distance within the Cluster C_i. The definitions of e_i and $nnDist(k, C_i)$ are formulated below:

TABLE 3
The Definitions of Similarity Measures

Name	Definition
Centroid Euclidian Distance	$D0 = \lvert\overrightarrow{c_i} - \overrightarrow{c_j}\rvert$
Centroid Manhattan Distance	$D1 = \sum_{k=1}^{d} \lvert c_i[k] - c_j[k]\rvert$
Average Inter-cluster Distance	$D2 = \sqrt{\frac{\sum_{p_i \in C_i} \sum_{p_j \in C_j} \lvert p_i - p_j\rvert^2}{\lvert C_i\rvert \lvert C_j\rvert}}$
Average Intra-cluster Distance	$D3 = \sqrt{\frac{\sum_{p_i, p_j \in C_i \cup C_j} \lvert p_i - p_j\rvert^2}{(\lvert C_i\rvert + \lvert C_j\rvert)(\lvert C_i\rvert + \lvert C_j\rvert - 1)}}$
Variance Increase Distance	$D4 = var(C_i \cup C_j) - var(C_i) - var(C_j)$

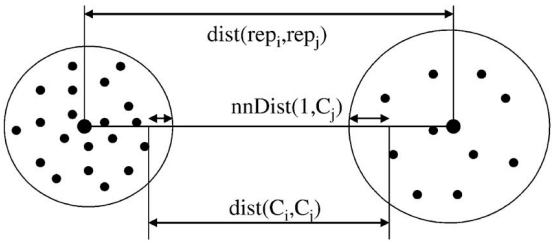

Fig. 12. Distance between two data bubbles.

$$e_i = \sqrt{\frac{\sum_{p_i,p_j \in C_i} |p_i - p_j|^2}{|C_i|(|C_i|-1)}} \text{ and}$$

$$nnDist(k, C_i) = \left(\frac{k}{|C_i|}\right)^{1/d} \times e_i,$$

where d is the dimension of the vector space.

As shown in Fig. 12, the distance between two data bubbles is defined under the assumption that all the data bubbles are spherical. However, this assumption does not always hold. With the covariance matrix of each data bubble (subcluster), we can extend the formula to elliptic data bubbles. This improvement can be made by changing the definition of *extent* as below.

Definition 7. *The extent e_i of an elliptic data bubble C_i along the direction of a unit length vector a is defined as*

$$e_i = 2\sqrt{\frac{1}{a^T \psi_i^{-1} a}},$$

where ψ_i, a $d \times d$ matrix, is the covariance matrix of the data bubble C_i.

As shown in Fig. 13, this formula is derived from the fact that the contours of the *pdf*, in a multivariate normal distribution, satisfies the condition $(v-\mu)^T \psi^{-1} (v-\mu) = c$. Thus, we have

$$(ta)^T \psi_i^{-1} (ta) = c \Rightarrow t^2 \times a^T \psi_i^{-1} a = c$$

$$e_i = t = \left(\frac{c}{a^T \psi_i^{-1} a}\right)^{\frac{1}{2}}.$$

Here, the value of parameter c is chosen as four so that the derived formula is consistent with the original one when the data bubbles are spherical.

Intuitively, if two neighboring clusters merge together, the density of the merged cluster will be expressed as

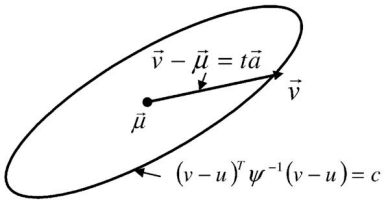

Fig. 13. The definition of extent in an oblique ellipse.

TABLE 4
Summary of Clustering Results of Different Measures

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6
Cohesion	O	O	O	O	O	O
Cohesion$^-$	O	O	O	X	X	X
D0	O	O	X	X	X	X
D1	O	O	X	X	X	X
D2	O	O	X	X	X	X
D3	O	O	X	X	X	X
D4	O	O	X	X	X	X
Bubble	O	O	O	X	X	X
E-Bubble	O	O	O	O	X	X
Density	O	O	O	O	X	X

O is successful and X is failed.

$$\frac{|C_i| + |C_j|}{C_i.volume + C_j.volume.}$$

However, the density drops if the two clusters are far away from each other. Thus, we define a new similarity measure as the ratio of expected density over the estimated density.

Definition 8. *The density ratio of merging two clusters, C_i and C_j, is*

$$dd(C_i, C_j) = \frac{\frac{|C_i|+|C_j|}{C_i.volume+C_j.volume}}{(C_i \cup C_j).density} = \frac{(C_i \cup C_j).volume}{C_i.volume + C_j.volume}.$$

We also compare with the similarity measurement defined in our prior work [21], where the *joinability* of two clusters, C_i and C_j, according the existence of a point p_i in Cluster C_i, is defined as

$$join(p_i, C_i, C_j) = \exp\left(-\frac{|p_i - c_i| - |p_i - c_j|}{r_i}\right),$$

where c_i and c_j are the centroids of the two clusters and r_i is the radius of cluster C_i. We improve the definition of *joinability* because it is observed that, once the distance from the point to the two centroids is the same, the value of *joinability* becomes one. Therefore, the value of cohesion becomes large for two clusters that are far away from each other with some noise points located at the middle of the two clusters. Although very rare, the occurrence of this scenario is still deemed a defect and is thus remedied in this paper. In addition, by using multivariate normal distribution, the new definition of *joinability* is more powerful for handling elliptic clusters. We denote the distance measure defined in [21] as Cohesion$^-$.

We replace cohesion with these similarity measures in algorithm CSM and then apply it to the six data sets. The clustering results are summarized in Table 4, where the measure E-Bubble refers to the distance of elliptical data bubbles. As shown, only cohesion can successfully partition all the data sets. To fairly compare these similarity measures, we perform this experiment 20 times. Each time, we perform algorithm CSM with each similarity measure and a unique random number so that the subclusters generated in Phase 1 are identical for each measure. We use the probability of successful clustering to judge the quality of each similarity measure. The probabilities of some better similarity measures are shown in Fig. 15. Note that the success rates of these similarity measures on Data Set 1 and

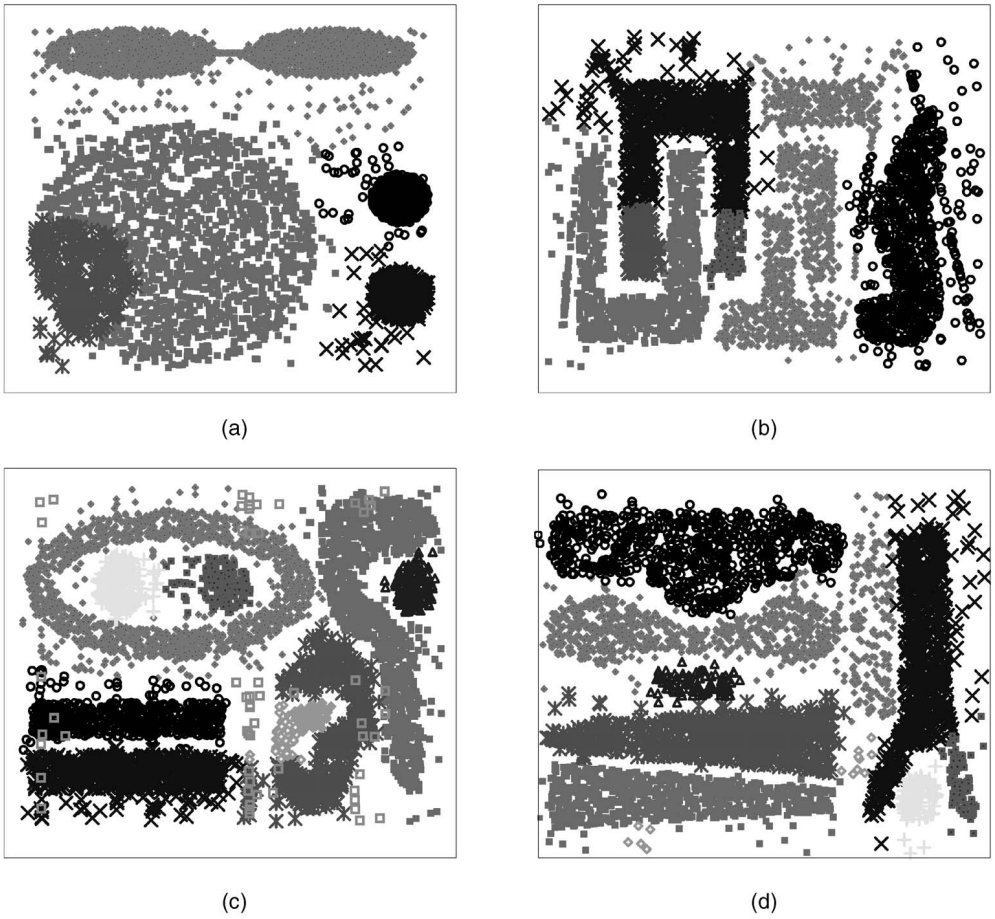

Fig. 14. Some clustering results of different similarity measures. (a) D0 on Data Set 3. (b) Bubble on Data Set 4. (c) Bubble on Data Set 5. (d) E-Bubble on Data Set 6.

Data Set 2 are all 100 percent and, thus, are omitted in Fig. 15. Some of those unsuccessful clustering results are shown in Fig. 14.

4.2 Experiment II: Comparing with Other Algorithms

We also apply other clustering algorithms on those data sets. The programs of algorithm BIRCH and DBScan are obtained from public domain. The k-means, single-link, and complete-link clustering algorithms are implemented as described in [19], [23], [27] using our best efforts. We also implement the algorithm CURE with the outlier elimination enhancement as described in [12]. For comparison reasons, we have carefully chosen the parameters for each algorithm. The clustering results are summarized in Table 5.

Some of unsuccessful clustering results are shown in Fig. 16. Note that the single-link algorithm is equipped with the outlier elimination enhancement as described in CURE. Thus, it can successfully partition Data Set 4. Otherwise, it will fail on all the data sets. In our observation, algorithm CURE fails on Data Set 5 and Data Set 6 because some clusters are in the shape of long stripes, while some noise links exist between neighboring clusters. Recall that algorithm CURE shrinks representatives toward the center of the cluster to avoid the link-effects. However, the shrink

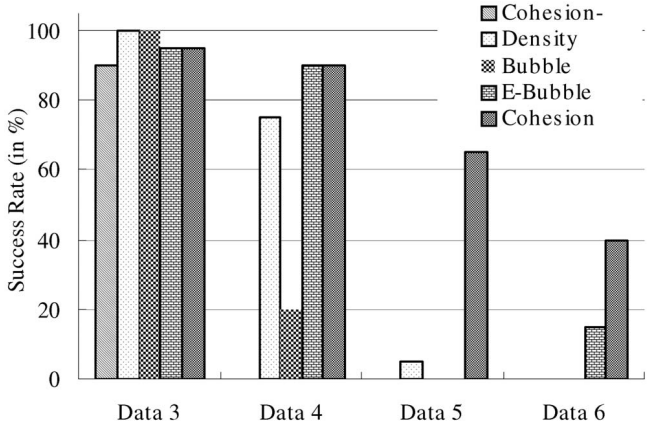

Fig. 15. The probabilities of successfully partitioned input data sets for some similarity measures.

TABLE 5
Summary of Clustering Results of Different Algorithms

	Data 1	Data 2	Data 3	Data 4	Data 5	Data 6
CSM	O	O	O	O	O	O
BIRCH	X	X	X	X	X	X
DBScan	O	X	X	O	O	X
CURE	O	O	O	O	X	X
K-means	O	O	X	X	X	X
Single-Link	X	X	X	O	X	X
Complete-Link	O	X	X	X	X	X

O is successful and X is failed.

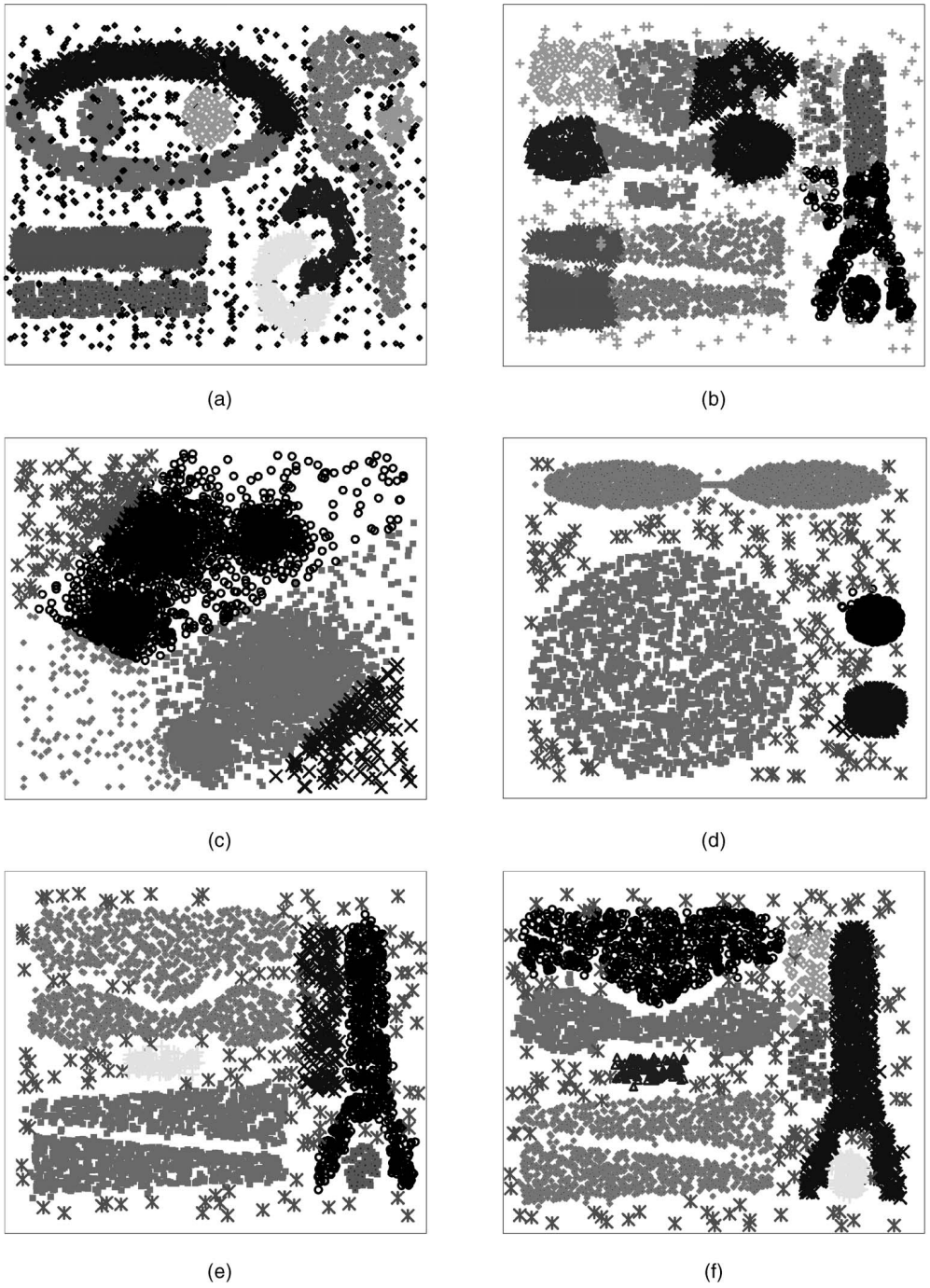

Fig. 16. Some clustering results of different algorithms. (a) Single Link on Data Set 5. (b) CURE on Data Set 6. (c) BIRCH on Data Set 1. (d) DBScan on Data Set 3. (e) DBScan on Data Set 6. (f) DBScan on Data Set 6.

mechanism will cause those slender clusters to be split. On the other hand, if we weaken the shrink mechanism, some clusters will be merged by noise links. In Data Set 4, algorithm CURE successfully finds a balanced shrink factor. However, it fails to find one in Data Set 5 and 6. Algorithm DBScan fails to find the five clusters in Data Set 2 because of the variety of density of those clusters. It also fails in Data Set 3 because there is a strong link between the upper two clusters. Note that the largest cluster in Data Set 1 is much sparser than the others. Thus, if we try to separate the upper two clusters by increasing the values of ε and/or $MinPts$, the other clusters will be merged with one another and/or the largest cluster will be regarded as noise. In Data Set 6, algorithm DBScan also faces a dilemma. As shown in Fig. 16e, it fails to separate two neighboring clusters linked by noise. If we try to separate them by increasing the values ε and/or $MinPts$, it will fail to identify the sparsest cluster, as shown in Fig. 16e. Among these algorithms, only algorithm CSM can obtain the correct clustering results.

Next, we conduct a scale-up experiment by applying these clustering algorithms on Data Set 3 of various sizes. Since the time complexities of these algorithms are much different from one another, we show the experimental results in Fig. 17a and Fig. 17b. For comparison reasons, we

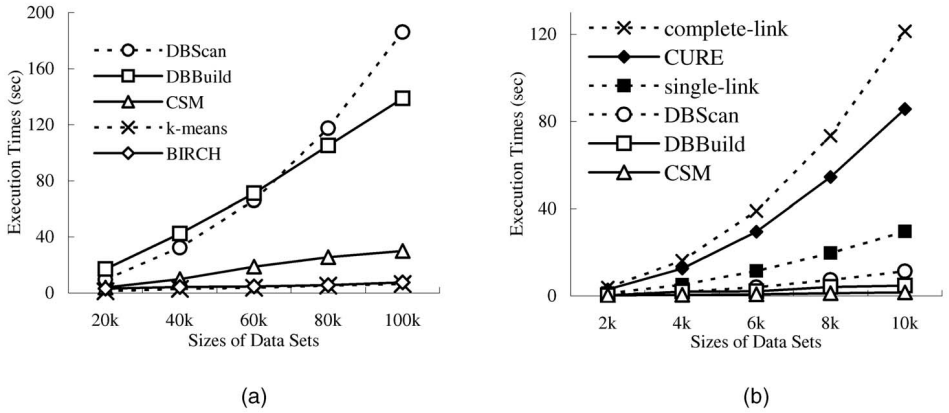

Fig. 17. Scale up experiment on the sizes of input data sets. (a) Comparison of faster algorithms. (b) Comparison of slower algorithms.

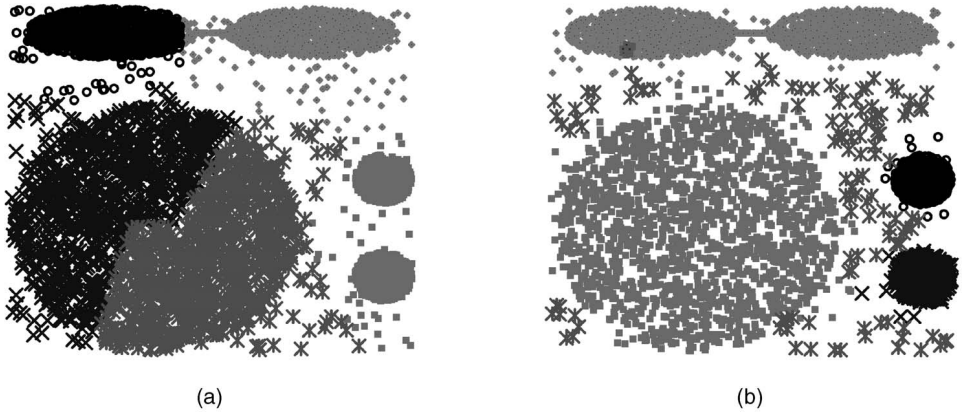

Fig. 18. The clustering results when m is too small or too large. (a) Small value of parameter m (10). (b) Large value of parameter m (256).

have tuned all the parameters so that the algorithm can be executed as fast as possible while accruing acceptable clustering qualities. For example, the value of parameter ε is 0.5 (the range of the Data Set 3 is 30×30), which is the smallest value that can successfully identify the four clusters (recall that algorithm DBScan fails to separate the upper two clusters). The value of parameter m in algorithm CSM is chosen as 16 and the number of representatives in algorithm CURE is chosen as 10. Note that the time required by algorithm DBScan to build its search structure, R^*-Tree, is also shown in the figure and denoted as DBBuild. Note that algorithm k-means and algorithm CSM are both randomized algorithms, thus we show the average execution times in these figures. As shown, algorithm CSM is faster than most algorithms (except for BIRCH and k-means) and scales linear to the size of input data set.

4.3 Experiment III: On the Number of Intermediate Clusters

Finally, we conduct a sensitivity analysis on the value of parameter m. It is observed that, when the value of m is too small, the subclusters produced in phase one may not properly partition the input data set, as shown in Fig. 18a. Thus, algorithm CSM results in an incorrect partition. On the other hand, if we partition the input data set into toomany subclusters, algorithm CSM may also fail to partition the input data set due to two problems. The first problem is the existence of many noise clusters. They will merge with other clusters and affect the clustering results. The second problem is those small clusters may form a link and connect two neighboring clusters. As shown in Fig. 18b, the upper two clusters are connected prior to the merging of the small subcluster and the left upper cluster. The first problem may be cured by our outlier resilience mechanism, while the second problem is hard to prevent when m is too large.

Next, we conduct a scale-up experiment on the value of parameter m. Specifically, we apply algorithm CSM on a $20k$-point subset of Data Set 3 with different values of parameter m. As shown in Fig. 19, algorithm CSM scales linear to the value of parameter m.

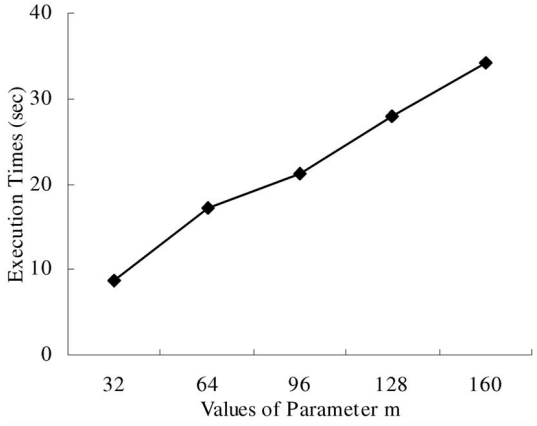

Fig. 19. The scale-up experiment on the value of parameter m.

5 Conclusion

In this paper, we propose a new similarity measure, cohesion, to measure the intercluster distances. By using cohesion, we proposed a two-phase clustering algorithm, CSM, whose time complexity is linear to the size of the input data set. Combining the features of partitional and hierarchical algorithms, algorithm CSM is able to not only resist outliers, but also lead to good clustering results while incurring a much shorter execution time than other algorithms. The time and the space complexities of CSM are also analyzed. As shown by our performance studies, the cohesion-based clustering is very robust and possesses excellent tolerance to outliers in various workloads. More importantly, algorithm CSM is shown to be able to cluster the data sets of arbitrary shapes very efficiently and to provide better clustering results than those by prior methods.

Acknowledgments

The authors would like to thank the anonymous referees for their helpful comments to improve this paper. This work was supported in part by the National Science Council of Taiwan, R.O.C., under Contracts NSC93-2752-E-002-006-PAE.

References

[1] C.C. Aggarwal, C. Procopiuc, J.L. Wolf, P.S. Yu, and J.-S. Park, "Fast Algorithms for Projected Clustering," *Proc. ACM SIGMOD '99,* 1999.

[2] K.S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When Is 'Nearest Neighbor' Meaningful?" *Proc. Int'l Conf. Database Theory (ICDT '99),* pp. 217-235, 1999.

[3] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York: Plenum Press, 1981.

[4] P.S. Bradley, K.P. Bennett, and A. Demiriz, "Constrained K-Means Clustering," Technical Report MSR-TR-2000-65, Microsoft Research, May 2000.

[5] M.M. Breunig, H.-P. Kriegel, P. Kröger, and J. Sander, "Data Bubbles: Quality Preserving Performance Boosting for Hierarchical Clustering," *Proc. ACM SIGMOD '01,* vol. 30, no. 2, pp. 79-90, 2001.

[6] A.G. Buchner and M. Mulvenna, "Discovery Internet Marketing Intelligence through Online Analytical Web Usage Mining," *Proc. ACM SIGMOD '98,* vol. 27, no. 4, pp. 54-61, Dec. 1998.

[7] M.-S. Chen, J. Han, and P.S. Yu, "Data Mining: An Overview from Database Perspective," *IEEE Trans. Knowledge and Data Eng.,* vol. 5, no. 1, pp. 866-883, Dec. 1996.

[8] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. B,* vol. 39, no. 1, pp. 1-38, 1977.

[9] R.C. Dubes, "How Many Clusters Are Best?—An Experiment," *Pattern Recognition,* vol. 20, no. 6, pp. 645-663, 1987.

[10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining,* pp. 226-231, 1996.

[11] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurasamy, *Advances in Knowledge Discovery and Data Mining.* Cambridge, Mass.: MIT Press, 1996.

[12] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," *Proc. Conf. Management of Data (ACM SIGMOD '98),* pp. 73-84, 1998.

[13] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," *Proc. 15th Int'l Conf. Data Eng.,* 1999.

[14] J. Han and M. Kamber, *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2000.

[15] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation.* Reading, Mass.: Addison-Wesley, 1991.

[16] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," *ACM Computer Surverys,* vol. 31, no. 3, Sept. 1999.

[17] G. Karypis, E.-H. Han, and V. Kumar, "Chameleon: Hierarchical Clustering Using Dynamic Modeling," *Computer,* vol. 32, no. 8, pp. 68-75, Aug. 1999.

[18] D.A. Keim and A. Hinneburg, "Clustering Techniques for the Large Data Sets—from the Past to the Future," *Tutorial Notes for ACM SIGKDD '99,* pp. 141-181, Aug. 1999.

[19] B. King, "Step-Wise Clustering Procedures," *J. Am. Statistical Assoc.,* vol. 69, pp. 86-101, 1967.

[20] C.-R. Lin and M.-S. Chen, "On the Optimal Clustering of Sequential Data," *Proc. Second Int'l Conf. Data Mining,* April 2002.

[21] C.-R. Lin and M.-S. Chen, "Robust and Efficient Clustering Algorithm Based on Cohesion Self-Merging," *Proc. Eighth Int'l Conf. Knowledge Discovery and Data Mining (ACM SIGKDD '00),* Aug. 2002.

[22] S.Y. Lu and K.S. Fu, "A Sentence-to-Sentence Clustering Procedure for Pattern Analysis," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 8, pp. 381-389, 1978.

[23] J. McQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. Fifth Berkeley Symp. Math. Statistics and Probability,* 1967.

[24] N.M. Murty and G. Krishna, "A Hybrid Clustering Procedure for Concentric and Chain-Like Clusters," *Int'l J. Computer and Information Sciences,* vol. 10, no. 6, pp. 397-412, 1981.

[25] R.T. Ng and J. Han, "Efficient and Effective Clustering Methods for Spatial Data Mining," *Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94),* 1994.

[26] J.O. Pedersen, D.R. Cutting, D.R. Karger, and J.W. Tukey, "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections," *Proc. 15th Int'l Conf. Research and Development in Information Retrieval (ACM),* pp. 318–329, 1992.

[27] P.H.A. Sneath and R.R. Sokal, *Numerical Taxonomy.* London: Freeman, 1973.

[28] J. Tantrum, A. Murua, and W. Stuetzle, "Hierarchical Model-Based Clustering of Large Datasets through Fractionation and Refractionation." *Proc. Eighth Int'l Conf. Knowledge Discovery and Data Mining (ACM SIGKDD '02),* Aug. 2002.

[29] A.K.H. Tung, J. Han, L.V.S. Lakshmanan, and R.T. Ng, "Constraint-Based Clustering in Large Databases," *Proc. Int'l Conf. Database Theory (ICDT '01),* Jan. 2001.

[30] C.-H. Yun, K.-T. Chuang, and M.-S Chen, "An Efficient Clustering Algorithm for Market Basket Data Based on Small-Large Ratios," *Proc. 25th Int'l Conf. Computer Software and Applications,* Oct. 2001.

[31] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *Proc. Conf. Management of Data (ACM SIGMOD '96),* pp. 103-114, 1996.

Cheng-Ru Lin received the BS and PhD degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2003, respectively. He is currently a researcher at Arcadyan Advanced Technology Center, Taipei, Taiwan. His research interests include data mining, distributed system, network, and multimedia applications.

Ming-Syan Chen received the MS and PhD degrees in computer, information, and control engineering from the University of Michigan, Ann Arbor, in 1985 and 1988, respectively. He received the BS degree in electrical engineering from National Taiwan University, Taipei, Taiwan. He is currently the chairman of the Graduate Institute of Communication Engineering and a professor at National Taiwan University, Taipei, Taiwan. His research interests include database systems, data mining, mobile computing systems, and multimedia networking. He is a fellow of the IEEE and a member of the ACM.