Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 641927, 10 pages
http://dx.doi.org/10.1155/2013/641927

Research Article

Active Semisupervised Clustering Algorithm with Label Propagation for Imbalanced and Multidensity Datasets

Mingwei Leng,[1] Jianjun Cheng,[1] Jinjin Wang,[1] Zhengquan Zhang,[2] Hanhai Zhou,[1] and Xiaoyun Chen[1]

[1] *School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China*
[2] *Gansu Computing Center, Lanzhou 730000, China*

Correspondence should be addressed to Xiaoyun Chen; chenxy@lzu.edu.cn

Received 27 August 2013; Revised 11 October 2013; Accepted 13 October 2013

Academic Editor: Gelan Yang

The accuracy of most of the existing semisupervised clustering algorithms based on small size of labeled dataset is low when dealing with multidensity and imbalanced datasets, and labeling data is quite expensive and time consuming in many real-world applications. This paper focuses on active data selection and semisupervised clustering algorithm in multidensity and imbalanced datasets and proposes an active semisupervised clustering algorithm. The proposed algorithm uses an active mechanism for data selection to minimize the amount of labeled data, and it utilizes multithreshold to expand labeled datasets on multidensity and imbalanced datasets. Three standard datasets and one synthetic dataset are used to demonstrate the proposed algorithm, and the experimental results show that the proposed semisupervised clustering algorithm has a higher accuracy and a more stable performance in comparison to other clustering and semisupervised clustering algorithms, especially when the datasets are multidensity and imbalanced.

1. Introduction

Semisupervised clustering algorithm has been studied recently as a method for improving the performance of clustering algorithm, and it allows the human expert to incorporate domain knowledge into the process of clustering and thus guides it to get better results. The use of domain knowledge in clustering task is motivated by the fact that the priori knowledge for some data objects can be obtained in many applications, the priori knowledge can be the labels of the data objects or the relationships between data objects. The "must-link" and "cannot-link" constraints capture relationships among data objects. Labeled objects could be used in clustering algorithms to help determine the groups and get more meaningful results. Most of the existing semisupervised clustering algorithms can be divided into three categories: method based on labeled data [1–9], pairwise constraints method [10–16], and fuzzy semisupervised method [17–22].

Semisupervised clustering algorithms based on labeled data utilized the label information to improve the performance of clustering. Semisupervised k-means clustering algorithm is a popular semisupervised clustering method [1–4]. Basu et al. exploited labeled data to generate initial seed clusters [1]. Bilenko et al. proposed a principled probabilistic framework based on hidden markov random fields for semisupervised clustering and presented HMRF-KMEANS based on EM and hidden markov random fields framework [2]. Leng et al. used labeled data to initialize the process of k-means clustering and obtained the similarity threshold of clusters based on the label information; they also utilized similarity threshold to guide k-means clustering algorithm [3]. Dang et al. presented a novel initialization method by propagating the labels of labeled data to more unlabeled data [4]. Zhong used deterministic annealing to expand three semisupervised clustering methods seeded clustering, constrained clustering, and feedback clustering, and their performances were compared with real text datasets [5]. Semisupervised density-based clustering is another kind of popular semisupervised clustering method [6, 7]. Lelis and Sander exploited labeled data to find values for ε. They gave a fixed value of MinPts and used the minimum spanning tree (MST) to partition dataset [6]. Böhm and Plant expanded

the clusters starting at all labeled objects simultaneously and proposed a semisupervised hierarchical clustering algorithm [7]. Guan et al. proposed an asymmetric similarity measure for two different documents and a new semisupervised clustering algorithm by expanding affinity propagation [8]. Shiga and Mamitsuka combined soft spectral clustering and label propagation and proposed a semisupervised clustering algorithm by learning locally informative data from multiple graphs [9].

The concepts of two basic pairwise constraints were defined by Wagstaff et al. [10]; they made the insertion of domain knowledge into the clustering (k-means in this case) process, and the pairwise constraints were given as the must-link and cannot-link. Reference [11] divided the pairwise constraints method into instance-level semisupervised clustering [10, 12, 13] and space-level semisupervised clustering [11, 14–16]. Wagstaff et al. viewed the pairwise constraints as instance-level constraints in the process of clustering and proposed the semisupervised clustering algorithm COP-KMeans [10]. Ruiz et al. proposed a semisupervised clustering algorithm called C-DBSCAN [12], which built a set of initial local clusters by partitioning data space into denser subspaces and cannot-link constraints, then merged density-connected local clusters and enforced the must-link constraints, finally, C-DBSCAN merged adjacent neighborhoods in a bottom-up fashion and enforced the remaining cannot-link constraints. Wang and Davidson combined spectral clustering and pairwise constraints in a principled and flexible manner [13]. They used a user-specified threshold to lower-bound how well the given constraints were satisfied, instead of trying to satisfy every given constraint, and they proposed a flexible and generalized framework for constrained spectral clustering. Instance-level semisupervised clustering method introduces pairwise constraints into clustering only and does not utilize the priori knowledge with the highest degree. Space-level semisupervised clustering not only makes use of constraints but also employs the space information provided by the constraints to adjust the process of clustering.

Fuzzy clustering model adopts membership to show the results of clustering, and membership grades are used as probabilities that each data object belongs to every class. In order to improve the performance of fuzzy clustering, the priori knowledge has been applied into it, and most of them used the priori knowledge to modify the objective function. Labeled data [17–19] and pairwise constraints [20–22] are two principal forms of priori knowledge in the fuzzy semisupervised clustering. Pedrycz and Waletzky improved the performance of clustering algorithm by using the information provided by labeled patterns to aid the process of clustering [17]. Bouchachia and Pedrycz utilized the information provided by labeled data to modify the objective function of fuzzy c-means [18]. Gao et al. proposed a fuzzy semisupervised clustering algorithm based on distance, which guided the process of clustering by using background information provided by labeled data and optimized the objective function by adding the label information into it [19]. Grira et al. added the information of pairwise constraints into the process of updating memberships and proposed a fuzzy semisupervised clustering algorithm based on pairwise constraints, which guided the process of solving membership matrix [20]. Pedrycz et al. used pairwise constraints information to optimize fuzzy c-means by adding an optimization step into the iteration process [21]. Yan et al. proposed fuzzy semisupervised coclustering algorithm for document by using the pairwise constraints to guide the process of constructing it [22].

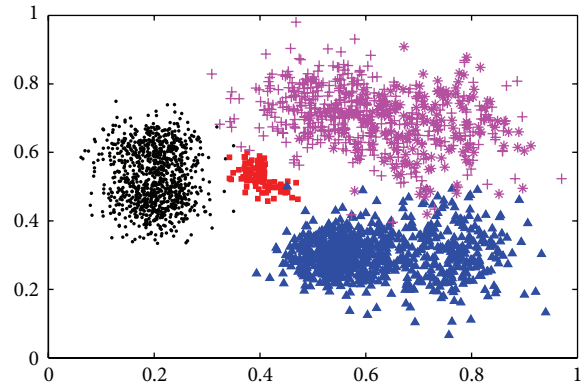

Figure 1: An imbalanced and multidensity dataset which contains 4 clusters.

Most of the semisupervised clustering algorithms assume that the labeled dataset or pairwise constraints are given. In practice, getting the priori knowledge is very expensive and time consuming. In addition, if the size of labeled dataset is too small in the process of constructing semisupervised clustering based on labeled data, some clusters may have no labeled data in imbalanced dataset, and then the data in those clusters will be assigned to other clusters forcibly. For example, the dataset shown in Figure 1 contains four clusters (these clusters are labeled with shapes "·", "■", "▲", and "∗", resp.). The size of cluster "■" is much less than that of the rest of the clusters. If the labeled data are randomly selected from the whole dataset, the data objects in cluster "■" are very difficult to be selected. If cluster "■" has no labeled data, then most of the semisupervised clustering algorithms will miss the cluster "■". How to select data from imbalanced dataset to guarantee that each cluster has more than one data that can be selected is one work of this paper. One of the solutions to this problem is to adopt active learning method to guide the process of selecting data points, which aims to cover as many clusters as possible.

The active learning method, which aims to achieve high accuracy using labeled data as few as possible, selects informative data actively and labels them by oracle. The active learning method can minimize the cost of obtaining labeled data points greatly without compromising the performance of clustering algorithm, and this is very attractive and valuable in real-world applications.

Perhaps the simplest and most commonly used active learning technique is uncertainty sampling [23], and least confident strategy, margin sampling, and entropy are the most popular uncertainty sampling strategies. Since the most likely label sequence can be efficiently computed using dynamic programming, least confident strategy has been

popular with statistical sequence models in information extraction tasks [24, 25]. However, the least confident strategy only considers information about the most probable label, and it discards information about the remaining label distribution, whereas margin sampling was proposed to correct for a shortcoming in least confident strategy by incorporating the second most likely label [26]. Entropy may be the most popular uncertainty sampling strategy, and it is easily applied to more complex structured instances, such as sequences [25] and trees [27]. Scheffer and Wrobel presented an active learning algorithm to reduce the required data labeling effort and increase the quality of the learning model by selecting "difficult" unlabeled samples [28].

Although most of the active learning strategies are applied into classification tasks, in the recent years, active learning is also introduced into clustering [29–35]. Mallapragada et al. selected constraints through using a min-max criterion to improve the performance of semisupervised clustering algorithms by selecting most uncertain data [29]. The uncertainty sampling technique selects the data objects which lie in the boundaries of clusters, and they are not "representative" of other data in the same cluster. Since knowing their labels is unlikely to improve the performance of the clustering algorithm as a whole, the "representative" method was proposed to solve this problem [30, 31]. Nguyen and Smeulders selected the most representative samples to avoid repeatedly labeling samples in the same cluster [30]. Vu et al. selected useful examples according to a min-max approach to determine the set of labeled data [31]. Active learning technique was also introduced into semisupervised clustering based on pairwise constraints [32–35]. Zhao et al. selected informative document pairs for obtaining user feedback by using active learning approach and incorporated instance-level constraints to guide the clustering process in DBSCAN [32]. Grira et al. defined an active mechanism for the selection of candidate constraints to minimize the amount of constraints required [33]. Wang and Davidson presented an active query strategy based on maximum expected error reduction and a constrained spectral clustering algorithm that can handle both hard and soft constraints [34]. Huang et al. conducted a preliminary clustering process to estimate the true clustering assignments and chose informative document pairs by means of learning the intermediate cluster structure [35].

Most of the existing active learning algorithms are pool-based or stream-based, and they are mainly applied in supervised learning. Although active learning is introduced into semisupervised clustering, the performances of these clustering algorithms are unsatisfiying when dealing with the imbalanced and multidensity datasets. The most uncertain data lies on the boundaries of clusters, and it is not "representative" of other data in the same cluster. So knowing its label is unlikely to improve the performance of the clustering algorithm as a whole. This paper selects the data with max density from each cluster which is the result of MST clustering.

Since the dataset is imbalanced, the distribution of labeled data in a given dataset is not the same as the whole data space, and a data point and its *k-nearest* labeled data may not be in the same cluster, which leads to the result that most of the existing semisupervised learning algorithms cannot work well, especially when the size of labeled dataset is very small. However, in the whole data space, the label of a data point should be the same as that of most of its *k-nearest neighbors*. The proposed semisupervised clustering algorithm with label propagation is based on this idea. It expands the labeled dataset by labeling *k-nearest neighbors* of labeled dataset based on a threshold. Once an unlabeled data is labeled, it should be added into labeled dataset. If the difference of density between clusters is large in multidensity datasets, the expanding process cannot use the same threshold, and the threshold should be generated automatically according to the density of each cluster to which the labeled data point belongs. A new active semisupervised clustering algorithm, called active semisupervised clustering algorithm for imbalanced and multidensity datasets, is proposed based on the facts previously described. The presented algorithm tries to ensure that the selected data can cover as many clusters as possible in a given imbalanced and multidensity dataset. Those selected data are labeled by oracle, and they are viewed as the initial set of labeled data in the process of semisupervised clustering. The proposed algorithm expands the labeled dataset by propagating labels according to expanded threshold obtained automatically based on the character of each cluster which is obtained by running MST clustering algorithm. The proposed clustering algorithm mainly has the following two advantages in comparison with other semisupervised clustering methods.

(1) The proposed semisupervised clustering method utilizes MST clustering to select data points actively so as to avoid labeling data in the same cluster repeatedly. If we need m labeled data objects, we partition the given dataset into m clusters by using MST clustering and select actively only one data from each cluster. This method can reduce the number of labeled data points greatly without compromising the performance of clustering, and the selected data can cover as many clusters as possible.

(2) The proposed clustering algorithm achieves label propagation by using labeled data to expand their *k-nearest neighbors* according to the criterion that is automatically obtained based on the density of the cluster to which the labeled data point belongs, and the expanding model only requires one parameter.

The rest of this paper is organized as follows. Section 2 gives the proposed semisupervised clustering algorithm. In Section 3, three datasets from UCI Machine Learning Repository and one synthetic dataset are used to demonstrate the proposed algorithm. We summarize our work in Section 4.

2. Active Semisupervised Clustering for Imbalanced and Multidensity Datasets

The *k-nearest neighbors* algorithm is most often used for classification, and it gives the label of an unlabeled data by comparing it to the first k most similar objects in the

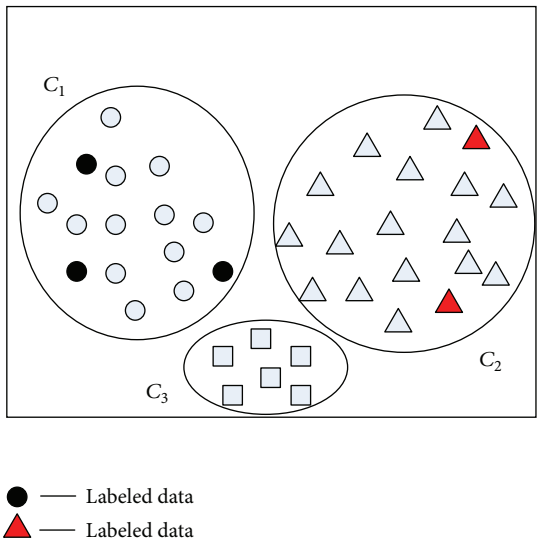

Figure 2: Low accuracy of KNN on an imbalanced dataset.

training set. Given a dataset $D = D_l \cup D_u$, where D_l is the labeled dataset and D_u is the unlabeled dataset, the *k-nearest neighbors* algorithm labels an unlabeled data y with the most frequent label among its *k-nearest* labeled neighbors. The label of an unlabeled data is given as follows:

$$y^l = \operatorname{argmax}_l \sum_{x \in \mathrm{KNN}(y, D_l)} \left[x^l == l\right], \tag{1}$$

where y^l and x^l are the labels of the data objects y and x, respectively, and the meaning of $\mathrm{KNN}(y, D_l)$ is defined as given in Definition 1.

Definition 1. $\mathrm{KNN}(x, C)$. Given one cluster C and one data object $x \in C$, $\mathrm{KNN}(x, C)$ is the set of *k-nearest neighbors* of x in C.

Each classification algorithm requires enough labeled data to achieve high classification accuracy. However, labeling data is quite expensive and time consuming in many real-world applications, and we can get a very small size of labeled dataset. For instance, there are 3 classes in Figure 2, and D_l contains 5 data objects (3 data objects in C_1, 2 data objects in C_2, and no data objects in C_3). The size of labeled dataset is very small compared with the whole dataset; suppose that we let $k = 1$ for *k-nearest neighbors* algorithm and use it to label the unlabeled data. All unlabeled data objects in C_3 and four unlabeled data objects in C_2 are assigned to C_1.

There are two problems for most of classifications and semisupervised clustering algorithms like *k-nearest neighbors* that lead those unlabeled data to wrong class when the size of labeled dataset is too small.

(1) The first one is that the whole dataset is imbalanced and the size of labeled dataset is too small, and using random method to select labeled data cannot guarantee that each class has more than one data object to be selected.

(2) The second is that the class label of some unlabeled data and that of its *k-nearest* labeled neighbors are not the same.

An active semisupervised clustering algorithm with label propagation for imbalanced and multidensity datasets is proposed to solve the previously mentioned problems. It uses MST clustering to partition the given dataset into clusters and selects one data object from each cluster as labeled data. This method for data selection can guarantee that the selected data can cover as many clusters as possible. Although the *k-nearest* labeled neighbors of each data in C_3 are not in C_3, the *k-nearest* neighbors are in C_3 (if $k \leq 4$). Since *k-nearest* neighbors of each data in C_3 are unlabeled, *k-nearest neighbors* algorithm has to find the nearest labeled neighbor from C_1 and C_2. The proposed algorithm selects more important data objects as labeled data and expands its label to its neighbors.

Some definitions are given as follows in order to describe the proposed active semisupervised clustering algorithm.

Definition 2. $dis\_\mathrm{KNN}(x, y, C)$. Given one cluster C, one data object $x \in C$, and $y \in \mathrm{KNN}(x, C)$, $dis\_\mathrm{KNN}(x, y, C)$ is the distance between x and y.

Definition 3. $k\_avgdis(C)$. Given one cluster C, $k\_avgdis(C)$ is defined as follows:

$$k\_avgdis(C) = \frac{\sum_{x \in C} \max dis\_\mathrm{KNN}(x, y, C)}{|C|}, \tag{2}$$

where $|C|$ is the number of data in cluster C.

Definition 4. $density(x, C)$. Given one cluster C and a data object $x \in C$, $density(x, C)$ is defined as follows:

$$density(x, C) = \frac{1}{\max dis\_\mathrm{KNN}(x, y, C)}. \tag{3}$$

The proposed active semisupervised clustering process can be divided into two algorithms: active data selection algorithm (Algorithm 1) and semisupervised clustering algorithm with label propagation (Algorithm 2). Algorithm 1 selects important data which do not lie in the boundaries of clusters and outputs those selected data after labeling them. Algorithm 2 expands the labeled datasets by propagating themselves labels to their neighbors.

If the dataset is imbalanced and we select small number of data points from this kind of datasets randomly, then there exist some clusters which have no data to be selected. Using these selected data as the labeled data to guide the process of clustering, the data objects in clusters which have no data being selected are assigned to other clusters forcibly. Thus, decreases the accuracy of semisupervised clustering algorithm, and the clustering results are unsatisfying. In order to make the selected data cover as many clusters as possible, an active mechanism of selecting data points is presented. It partitions a given dataset into m clusters by using MST clustering algorithm; here, m is the number of the data objects which will be selected, and only one data

(1) Let $m = |D| \times p$, m is the number of data points to be selected, $|D|$ is the size of dataset D.
(2) Use Prime method to construct MST of D.
(3) **Foreach** *edge* in MST **do**
(4) Compute edge's inconsistent value f.
(5) **End Foreach**
(6) Sort all *edges* in descending order according to f.
(7) Insert the sorted edges into a list: *edgesLst*.
(8) **Foreach** *edge* in *edgesLst* **do**
(9) Delete *edge* from *MST*
(10) Check the number of partitions in MST, *num*
(11) **If** *num* == m **then**
(12) Generate num clusters $T_1, T_2, \ldots, T_m$ from MST
(13) **Break**
(14) **End If**
(15) **End Foreach**
(16) **Foreach** cluster T in $T_1, T_2, \ldots, T_m$ **do**
(17) Compute density of each point in T
(18) Select one data with max density and add it to D_l
(19) **End Foreach**
(20) Query oracle about labels of data in D_l.
(21) Return $T_1, T_2, \ldots, T_m$ and D_l.

Algorithm 1: Selecting data by using MST clustering algorithm (*SelectDataPoint* (D, p)).

(1) Input the value of k.
(2) *SelectDataPoint* (D, p).
(3) Suppose that the number of different labels in D_l is p.
(4) Merge $T_1, T_2, \ldots, T_m$ into $C_1, C_2, \ldots, C_p$ according to labels of data in D_l.
(5) **Foreach** cluster C in $C_1, C_2, \ldots, C_p$ **do**
(6) **Foreach** data point x in C **do**
(7) Compute the KNN (x, C)
(8) **End Foreach**
(9) Compute $k_avgdis\,(C)$s
(10) **End Foreach**
(11) **Foreach** cluster C in $C_1, C_2, \ldots, C_p$ **do**
(12) *Expend*$(C, k\_avgdis\,(C), D_l)$
(13) **End Foreach**
(14) Label the rest unlabeled data according to KNN rule.
(15) Output the clustering results.

Algorithm 2: Semisupervised clustering algorithm with label propagation.

point is chosen in each cluster. Since only one data point in each cluster is selected, each of selected data should be the better representations of corresponding cluster, and the centers of clusters and the data with maximum density are two better representation of each cluster. This paper utilizes the method of label propagation to achieve a high accuracy of semisupervised clustering algorithm, and the data objects with maximum densities are chosen by us and are labeled by oracle. The details of selecting data points are shown in Algorithm 1.

Algorithm 1 has two parameters D and p. D is the dataset which will be clustered, and p is the percent of the selected data in D. Algorithm 1 uses the MST clustering to partition D into m clusters, and the value of m is larger than or equal to the real number of clusters in the dataset D. MST clustering algorithm used in Algorithm 1 is proposed by Zahn [36]. In the process of labeling the data, we should select the certain data objects which do not lie in the boundaries of clusters. Since the selected data are "representative" of other data in the same cluster, their labels are easy to be labeled, and this can reduce the required data labeling effort and increase the quality of the labeled data. The proposed semisupervised clustering algorithm requires very small number of labeled data, and even some cluster has only one data to be selected as labeled data. The data with max density in one cluster is easier to be labeled compared

(1) Get all the labeled data which belong to C from D_l.
(2) Let D_C^l denote these labeled data.
(3) **Foreach** x in D_C^l **do**
(4) Compute density of data x, $density\,(x, C)$
(5) **End Foreach**
(6) Sort D_C^l in descending order according to data density
(7) **While** (D_C^l is not null)
(8) Take out the first data x from D_C^l
(9) Compute KNN (x, C)
(10) **Foreach** y in KNN (x, C) **do**
(11) **If** $dis\_\text{KNN}\,(x, y, C) \leq avgdis$
(12) $y^l \leftarrow x^l$
(13) Insert y into D_C^l, and add y into D_l
(14) **End If**
(15) **End Foreach**
(16) Delete x from D_C^l
(17) **End while**
(18) Return D_l

Algorithm 3: $Expend(C, k\_avgdis\,(C)\,, D_l)$.

with the rest of data, so Algorithm 1 selects the data with max density in each cluster and labels them by querying the oracle about labels of the selected data.

How to use small number of labeled data to achieve a higher accuracy of clustering algorithm is a challenging work, especially when the dataset is imbalanced and multidensity. The semisupervised clustering algorithms should use the character of labeled dataset to guide their clustering process. In this paper, firstly, the clustering results of MST are merged according to the label of its labeled data (each cluster has and only has one labeled data). Since the density of each cluster is not unique and the densities of clusters may be different, we should not use the same expanding threshold when utilizing the method of label propagation to expand the labeled dataset. Secondly, the expanding threshold of each cluster should be obtained based on its density automatically, and it is used to expand the labeled dataset in one cluster. Finally, the rest of unlabeled data are assigned with the most frequent label among its *k-nearest* labeled neighbors. More detailed information is given in Algorithm 2.

The k in step 1 of Algorithm 2 is the parameter of *k-nearest neighbors*. Step 2 uses Algorithm 1 to select m data points. Since the value of m is not less than that of p and if m is larger than p, then some clusters in $T_1, T_2, \ldots, T_m$ are in the same cluster. Algorithm 2 can be divided into three stages. Firstly, Step 4 merges the clusters which should be in the same cluster into one. x_i and x_j are two data points in D_l, and $x_i^{\,l}$ and $x_j^{\,l}$ are the labels of them, respectively. If $x_i^{\,l} == x_j^{\,l}$, then Step 4 merges T_i and T_j into one. Secondly, different clusters may have different densities in multidensity datasets, which leads to the result that the process of label expanding cannot adopt the same expanding threshold on the whole data space when the difference of density between clusters is very large. It should adopt different expanding threshold according to its density of the cluster to which it belongs. Step 9 computes the expanding threshold for each cluster. In each cluster $C_i\,(1 \leq i \leq p)$, the labeled data which are in C_i expand their labels to their *k-nearest neighbors* based on the threshold which is obtained in C_i automatically, and function $Expend(C, k\_avgdis(C), D_l)$ uses the expanding threshold $k\_avgdis(C)$ to expand the labeled dataset D_l by propagating the labels of labeled data in cluster C, and the expanding process is given as Algorithm 3. Steps 5 to 13 complete the process of label propagating. Thirdly, since we use the expanding threshold $k\_avgdis(C)$ in the process of label propagation, then part of unlabeled data in cluster C is not be labeled. We should label these unlabeled data after the ending of label propagation and use the *k-nearest neighbors* rule to deal with the rest of unlabeled data.

Algorithm 3 expands the labeled data in cluster C by using the mechanism of label propagation. In cluster C, we find out the *k-nearest neighbors* in C for each data x in C. In cluster C, $k\_avgdis(C)$ is used as the expanding threshold, which is necessary in multidensity dataset. Steps 10 to 15 utilize $k\_avgdis(C)$ as the threshold to expand the labeled data in cluster C. Firstly, we take out one labeled data x which has not been used to expand its label to KNN(x, C) in C. For any data point y in KNN(x, C), if and only if $dis\_\text{KNN}(x, y, C)$ is less than $k\_avgdis(C)$, the label of x is assigned to y. After dealing with KNN(x, C), it takes another labeled data which has not been used to expand its *k-nearest neighbors* in C and uses the same method to label its *k-nearest neighbors*. If all of the labeled data in C have been used to label their *k-nearest neighbors*, Algorithm 3 returns the D_l as the result.

3. Experimental Results and Discussion

We use three standard datasets from UCI Machine Learning Repository [37]—IRIS, Wine, and Ecoli—and one synthetic dataset which is imbalanced and multidensity to demonstrate

the performance of the proposed algorithm. The Euclidean metric is employed to compute the distances between data objects. In order to prove that the proposed method has the ability of dealing with the imbalanced and multidensity datasets, we construct three imbalanced datasets by deleting data objects from IRIS, Wine, and Ecoli. Since the priori knowledge is given as the labeled data, we compare the proposed algorithm with SSDBSCAN and Constrained-Kmeans. We use the clustering accuracy to evaluate the clustering results. The notion of clustering accuracy (CA) of a dataset D is defined as follows:

$$\mathrm{CA} = \frac{|D'|}{|D|} \times 100\%, \tag{4}$$

where $|D|$ is the size of the unlabeled dataset D and $|D'|$ is the number of labeled data which are labeled correctly by clustering algorithms in D.

3.1. Standard Datasets. This subsection demonstrates the performance of the proposed semisupervised clustering algorithm in three UCI datasets: IRIS, Wine, and Ecoli. In order to test that the proposed algorithm has a higher accuracy compared with SSDBSCAN and Constrained-Kmeans in imbalanced and multidensity datasets, three datasets are constructed by deleting part of data from some clusters of IRIS, Wine, and Ecoli.

3.1.1. IRIS Dataset. The IRIS dataset, which contains 150 data objects, is perhaps the most well-known dataset in pattern recognition and data mining literature. IRIS contains 3 clusters of 50 data objects each. We turn IRIS into imbalanced and multidensity dataset by deleting 20 data objects from the second cluster randomly, and let modified IRIS denote this dataset. Since IRIS contains only 150 data objects, we select 2, 3, 4, 5, 6, 7, 8, 9, and 10 percents of the dataset from IRIS and the modified IRIS to be labeled datasets, respectively, and view the rest of the data as the unlabeled datasets. The experimental results are shown in Figures 3 and 4.

Figure 3 shows the experimental results of the 3 algorithms which run on the IRIS dataset. Figure 3 shows that the proposed algorithm has a higher accuracy compared with the SSDBSCAN and Constrained-Kmeans. In addition, the proposed algorithm is more stable than SSDBSCAN and Constrained-Kmeans, especially when the size of labeled dataset is very small. The proposed algorithm can reach stable state when selecting more than 3% of all data (there are only 4 labeled data). The accuracy of Constrained-Kmeans is very low when selecting 3% and 4% of all data, just because there is one cluster which has no data being selected, Constrained-Kmeans partitions IRIS dataset into 2 clusters forcibly, and SSDBSCAN has the same problem. The method of labeled data selection is based on MST clustering, and the experimental results show that the accuracy of clustering can be improved highly when using 4 labeled data to guide the process of clustering.

Figure 4 displays the experimental results of algorithms running on the modified IRIS dataset. The proposed algorithm has a much higher accuracy and more stable state

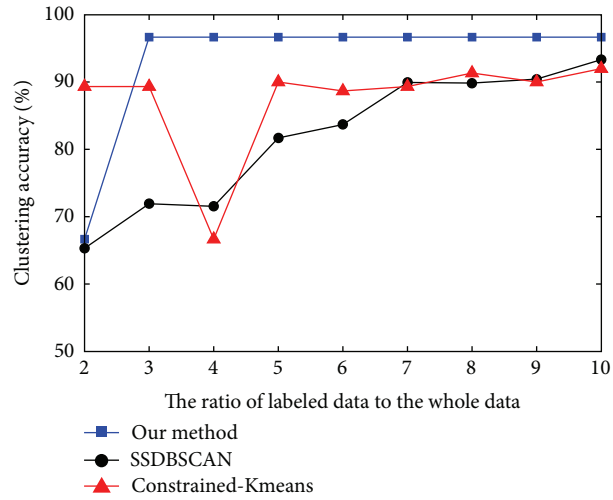

Figure 3: Clustering accuracy (%) obtained with the proposed algorithm and other algorithms on IRIS.

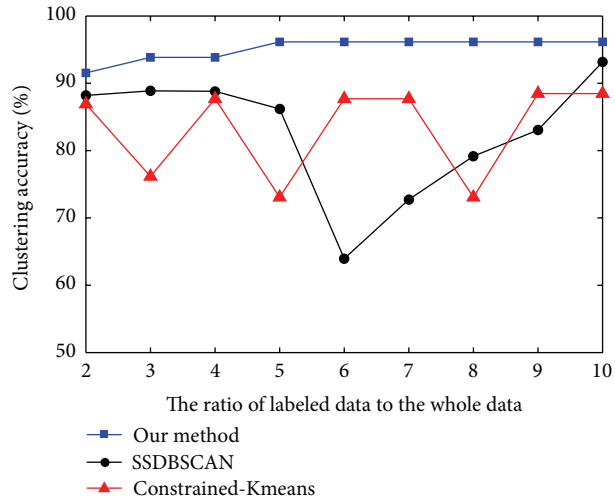

Figure 4: Clustering accuracy (%) obtained with the proposed algorithm and other algorithms on modified IRIS.

than SSDBSCAN and Constrained-Kmeans. When IRIS is modified to be imbalanced and multidensity, the labeled data which is selected by using random method cannot cover all clusters, which makes some clusters assigned to other clusters in force, and this is reflected in SSDBSCAN and Constrained-Kmeans, especially in Constrained-Kmeans. But the accuracy of the proposed algorithm is little influenced. The accuracy of the proposed algorithm reaches 93.8% when selecting 3% of all data, and the presented algorithm can reach stable state when selecting more than 7 labeled data.

3.1.2. Wine Dataset. Wine dataset contains 178 data objects, and these data can be assigned to 3 clusters whose sizes are 59, 71, and 48, respectively. We adapt the same method to turn Wine dataset into an imbalanced and multidensity dataset by removing 25 data objects from the first cluster randomly, and let modified Wine denote this dataset. We select 2, 3, 4, 5,

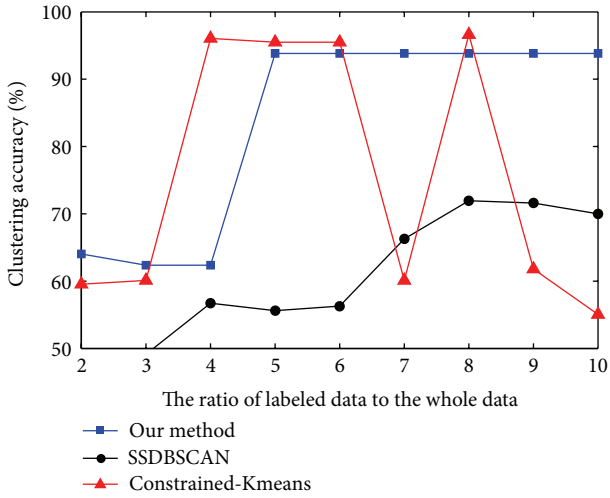

Figure 5: Clustering accuracy (%) obtained with the proposed algorithm and other algorithms on Wine.

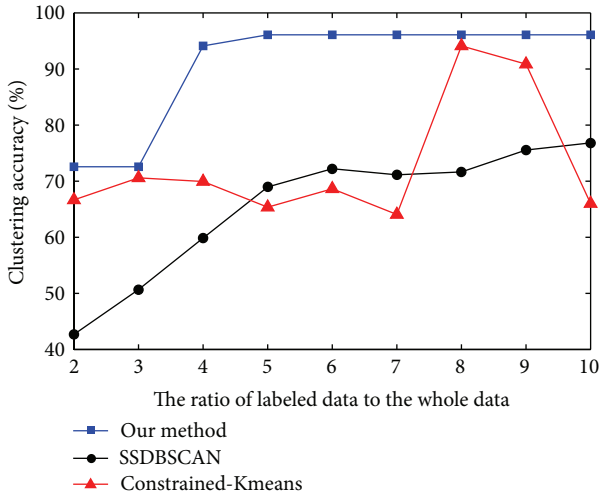

Figure 6: Clustering accuracy (%) obtained with the proposed algorithm and other algorithms on modified Wine.

6, 7, 8, 9, and 10 percents of the dataset from Wine and the modified Wine to be labeled datasets, respectively, and view the rest of the data as unlabeled datasets. Figures 5 and 6 show the changes of accuracy of the three algorithms.

Figure 5 shows that the proposed algorithm has a more stable state than SSDBSCAN and Constrained-Kmeans, and the accuracy of the proposed algorithm is much higher than that of Constrained-Kmeans. The proposed algorithm can reach stable state when selecting more than 5% of all data (there are only 9 labeled data). Since SSDBSCAN and Constrained-Kmeans use random method to select labeled datasets, there exists some cluster that has no data that can be selected as labeled data, and their accuracy fluctuates along with the change of percent of labeled data and this is also shown in Figure 5.

Figure 6 shows that the accuracy of Constrained-Kmeans and SSDBSCAN fluctuates much larger than that of the proposed method, and the proposed algorithm reaches a stable state when selecting only 5% of all data. When we select more than 4% of all data as the labeled data actively, the accuracy of the proposed method is 94.1%, and when the percent is more than 5, the accuracy is 96.1%. When the labeled data cover all clusters, Constrained-Kmeans has a high clustering accuracy which is close to that of the proposed method. But, in the 9 labeled datasets, only two labeled datasets cover all clusters, and the rest 7 labeled datasets miss some cluster. The accuracy of Constrained-Kmeans is less than 80% on the 7 labeled datasets. The accuracy of SSDBSCAN is less than 80% on all the labeled datasets.

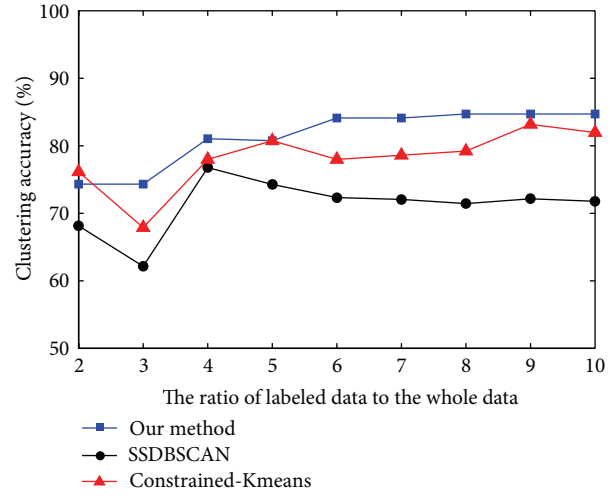

Figure 7: Clustering accuracy (%) obtained with the proposed algorithm and other algorithms on Ecoli.

3.1.3. Ecoli Dataset. The Ecoli dataset, which contains 336 data objects, has 8 clusters. The sizes of the 8 clusters are 143, 77, 52, 35, 20, 5, 2, and 2, respectively.

Since the data objects of the last three clusters are less than 6 and they can be viewed as noises, in the experiment, we delete these data. We select 2, 3, 4, 5, 6, 7, 8, 9, and 10 percents of the dataset from Ecoli dataset, respectively. The experimental results are shown in Figure 7. The results are similar to those in Figures 4 and 6. Figure 7 shows that the proposed algorithm has a much higher accuracy and more stable state than SSDBSCAN and Constrained-Kmeans. The accuracy of Constrained-Kmeans and SSDBSCAN fluctuates along with the difference of labeled data.

3.2. Synthetic Dataset. In this subsection, we generate 2500 data objects which have two attributes and are viewed as imbalanced and multidensity datasets, and these data can be partitioned into 4 clusters whose sizes are 1000, 100, 800, and 600, respectively. These data are shown in Figure 1. Ten subsets were selected from this synthetic dataset to demonstrate the three algorithms, and the ratios of them to the whole dataset are 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 percents, respectively. The experimental results are shown in Figure 8.

The accuracy of Constrained-Kmeans and SSDBSCAN depends on the labeled data seriously. Although we select

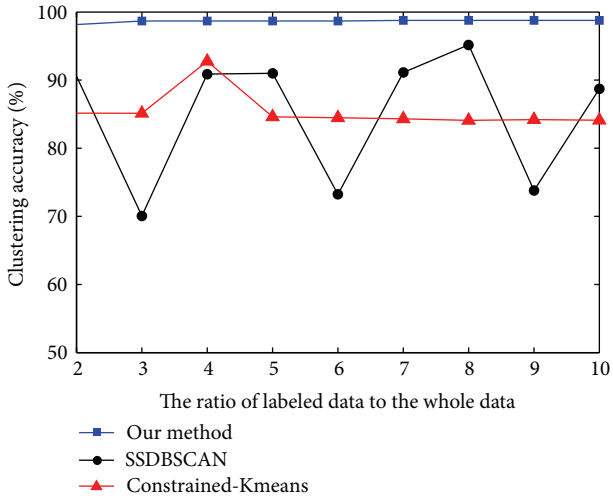

FIGURE 8: Clustering accuracy (%) obtained with the proposed algorithm and other algorithms on Synthetic.

10% of all data, the second cluster has no one data that can be selected as the labeled data, and, in the clustering results, the data objects in the second cluster have to be assigned to other clusters, and this phenomenon manifests in the clustering results of SSDBSCAN. In addition, even if Constrained-Kmeans selects labeled data from all of the clusters, it assigns some data objects from the rest of the three clusters to the second cluster, and this is the reason why the accuracy of Constrained-Kmeans is not improved as the percent of labeled data increases. Figure 8 also shows that the proposed algorithm has a much higher accuracy compared with SSDBSCAN and Constrained-Kmeans. The accuracy of the proposed algorithm exceeds 98% on the 10 subsets.

4. Conclusion

A new active semisupervised clustering algorithm is proposed which actively selects informative data by dealing with the clustering results of MST. Labeling these data and using them to label their *k-nearest neighbors* are based on an adaptive threshold. The experimental results show that the proposed semisupervised clustering can reach a stable state which only requires very small size of labeled dataset. However, the accuracy of the proposed semisupervised clustering is much lower in the dataset in which clusters overlap each other than that in the dataset in which the boundaries between clusters are not very vague. In the future, we plan to extend this work to the dataset in which clusters overlap each other. We will work on the data selection strategy in an active manner and the method of label propagation in the imbalanced and multidensity datasets in which clusters overlap each other.

Acknowledgments

This paper is supported by the Fundamental Research Funds for the Central Universities (lzujbky-2012-212) and is partially supported by the IBM 2010 X10 Innovation Awards Project.

References

[1] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *Proceedings of the 19th International Conference on Machine Learning (ICML '02)*, pp. 27–34, 2002.

[2] M. Bilenko, S. Basu, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 59–68, August 2004.

[3] M. Leng, X. Chen, and L. Li, "K-means clustering algorithm based on semi-supervised learning," *Journal of Computational Information Systems*, vol. 4, no. 5, pp. 2007–2013, 2008.

[4] Y. Dang, Z. Xuan, L. Rong, and M. Liu, "A novel initialization method for semi-supervised clustering," in *Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management*, pp. 317–328, 2010.

[5] S. Zhong, "Semi-supervised model-based document clustering: a comparative study," *Machine Learning*, vol. 65, no. 1, pp. 3–29, 2006.

[6] L. Lelis and J. Sander, "Semi-supervised density-based clustering," in *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM '09)*, pp. 842–847, December 2009.

[7] C. Böhm and C. Plant, "HISSCLU: a hierarchical density-based method for semi-supervised clustering," in *Proceedings of the 11th International Conference on Extending Database Technology (EDBT '08)*, pp. 440–451, March 2008.

[8] R. Guan, X. Shi, M. Marchese, C. Yang, and Y. Liang, "Text clustering with seeds affinity propagation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 4, pp. 627–637, 2011.

[9] M. Shiga and H. Mamitsuka, "Efficient semi-supervised learning on locally informative multiple graphs," *Pattern Recognition*, vol. 45, no. 3, pp. 1035–1049, 2012.

[10] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedel, "Constrained k-means clustering with background knowledge," in *Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, pp. 77–584, 2001.

[11] D. Klein, S. Kamvar, and C. Manning, "From instance-level constraints to space- level constraints: making the most of prior knowledge in data clustering," in *Proceedings of the 19th International Conference on Machine Learning (ICML '02)*, pp. 307–314, 2002.

[12] C. Ruiz, M. Spiliopoulou, and E. Menasalvas, "Density-based semi-supervised clustering," *Data Mining and Knowledge Discovery*, vol. 21, no. 3, pp. 345–370, 2010.

[13] X. Wang and I. Davidson, "Flexible constrained spectral clustering," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*, pp. 563–572, July 2010.

[14] I. S. Dhillon, Y. Guan, and B. Kulis, "Weighted graph cuts without eigenvectors a multilevel approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 11, pp. 1944–1957, 2007.

[15] Z. Lu and M. Á. Carreira-Perpiñán, "Constrained spectral clustering through affinity propagation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, pp. 1–8, June 2008.

[16] C. Alzate and J. A. K. Suykens, "A regularized formulation for spectral clustering with pairwise constraints," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '09)*, pp. 141–148, June 2009.

[17] W. Pedrycz and J. Waletzky, "Fuzzy clustering with partial supervision," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 27, no. 5, pp. 787–795, 1997.

[18] A. Bouchachia and W. Pedrycz, "A semi-supervised clustering algorithm for data exploration," in *Proceedings of the 10th International Fuzzy Systems Association World Congress Conference on Fuzzy Sets and Systems*, pp. 328–337, 2003.

[19] J. Gao, P.-N. Tan, and H. Cheng, "Semi-supervised clustering with partial background information," in *Proceedings of the 6th SIAM International Conference on Data Mining (SDM '06)*, pp. 489–493, 2006.

[20] N. Grira, M. Crucianu, and N. Boujemaa, "Semi-supervised fuzzy clustering with pairwise-constrained competitive agglomeration," in *Proceedings of the 14th IEEE International Conference on Fuzzy Systems*, pp. 867–872, May 2005.

[21] W. Pedrycz, V. Loia, and S. Senatore, "P-FCM: a proximity-based fuzzy clustering," *Fuzzy Sets and Systems*, vol. 148, no. 1, pp. 21–41, 2004.

[22] Y. Yan, L. Chen, and W.-C. Tjhi, "Fuzzy semi-supervised co-clustering for text documents," *Fuzzy Sets and Systems*, vol. 215, pp. 74–89, 2013.

[23] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3–12, 1994.

[24] A. Culotta and A. McCallum, "Reducing labeling effort for structured prediction tasks," in *Proceedings of the 20th National Conference on Artificial Intelligence and the 17th Innovative Applications of Artificial Intelligence Conference (AAAI '05)*, pp. 746–751, July 2005.

[25] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pp. 1069–1078, 2008.

[26] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA '01)*, pp. 309–318, 2001.

[27] R. Hwa, "Sample selection for statistical parsing," *Computational Linguistics*, vol. 30, no. 3, pp. 253–276, 2004.

[28] T. Scheffer and S. Wrobel, "Active learning of partially hidden markov models," in *Proceedings of ECML/PKDD Workshop on Instance Selection*, 2001.

[29] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *Proceedings of the 19th International Conference on Pattern Recognition (ICPR '08)*, pp. 1–4, December 2008.

[30] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proceedingsof the 21st International Conference on Machine Learning (ICML '04)*, pp. 623–630, July 2004.

[31] V.-V. Vu, N. Labroche, and B. Bouchon-Meunier, "Active learning for semi-supervised K-means clustering," in *Proceedings of the 22nd International Conference on Tools with Artificial Intelligence (ICTAI '10)*, pp. 12–15, October 2010.

[32] W. Zhao, Q. He, H. Ma, and Z. Shi, "Effective semi-supervised document clustering via active learning with instance-level constraints," *Knowledge and Information Systems*, vol. 30, no. 3, pp. 569–587, 2012.

[33] N. Grira, M. Crucianu, and N. Boujemaa, "Active semi-supervised fuzzy clustering," *Pattern Recognition*, vol. 41, no. 5, pp. 1834–1844, 2008.

[34] X. Wang and I. Davidson, "Active spectral clustering," in *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM '10)*, pp. 561–568, December 2010.

[35] R. Huang, W. Lam, and Z. Zhang, "Active learning of constraints for semi-supervised text clustering," in *Proceedings of the 7th SIAM International Conference on Data Mining (SDM '07)*, pp. 113–124, April 2007.

[36] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Transactions on Computers*, vol. 20, no. 1, pp. 68–86, 1971.

[37] Uci datasets, http://archive.ics.uci.edu/ml/datasets.html.