The Research on Clustering Ensembles Selection Algorithm based on Semi-supervised K-means Clustering

Xiaoxuan Wu 1,2,*, Changjian Guo1, Tanglei Hu3
1School of Artificial Intelligence and Big Data, Hefei University, Hefei, China
2Anhui Engineering Lab of Big Data Technology Application for Urban Infrastructure, Hefei University, Hefei, China
3Institute of Business Intelligence, Hefei University of Technology, Hefei, China

*Corresponding author e-mail: kexinyufan@163.com

Abstract. Selective clustering ensemble algorithm can eliminate the inferior quality clustering member’s influence and can achieve a better clustering solution relative to the clustering ensemble algorithm. For high dimensional data clustering, in this paper, a novel selective ensemble algorithm based on semi-supervised K-means clustering is proposed. In this paper, through a large number of experiments to verify the validity of the proposed algorithm for dealing with high dimensional data clustering. The new algorithm can achieve statistically significant performance improvement over other clustering algorithms.

1. Introduction

Clustering is the process of classifying data objects so that the data objects in the same class are as similar as possible and the data objects between different classes are as similar as possible. The function of clustering is to gather the data with a large degree of similarity into one category in the case of unsupervised learning, that is, without knowing the category in advance, to facilitate people to study and analyze the properties and characteristics of each type of data[1]. Due to the uncertainty of clustering, how to select an appropriate clustering algorithm for specific data has always been the focus of clustering analysis[2]. Clustering fusion is the fusion of the clustering results generated by the design of the common knowledge function to maximize the existing clustering results to share information, to obtain more accurate and stable mining results than the single clustering algorithm[3-5].

Diversity among cluster members is the key to the success of clustering fusion, especially the greater the diversity among cluster members, the better the fusion result[6]. It indicates that different cluster members have different impacts on the fusion result. High-quality cluster members can improve the quality of the fusion, on the contrary, inferior cluster members can also affect the fusion result. To eliminate the interference of inferior clustering members, a selective clustering fusion method is produced. Compared with the traditional clustering fusion algorithm, which integrates all the clustering members, the selective clustering fusion algorithm selects some clustering members to participate in the fusion according to certain selection strategies and obtains better results, which improves the accuracy and robustness of the algorithm.

Literature [7] puts forward the idea of selective integration and proves that partial integration can achieve better results than total integration. Literature [8-10] has confirmed that the quality and diversity of cluster members are two crucial factors influencing the integrated learning system. Literature [9] puts forward the concept of selective clustering integration and puts forward three kinds...
of selective fusion algorithm clustering: JC (be Criterion), CAS (cluster and select) and Convex Hull. Literature [6,11] proposed a random projection clustering method for high-dimensional data sets. Literature [12] proposed a discriminant semi-supervised clustering algorithm based on pair constraints, which solved the violation of pair constraints and the dimensionality reduction of high-dimensional data. Literature [13] proposed an adaptive selective clustering fusion method. Literature [14] proposed an optional clustering integration algorithm based on Bagging. By using mutual information to calculate the selection strategy of clustering fruit weight, individual learners with better quality were selected for integration, and the integration results were better. In literature [15], a new selection and fusion strategy is proposed. Firstly, pairwise fusion technology is used to substitute all clustering results for clustering member selection, and then weighted clustering member fusion based on attributes is adopted. In literature [16], a new clustering fusion algorithm is proposed, which takes into account the clustering quality and membership diversity, adopts a new similarity measure, and clips the clustering members first, then groups and selects them according to the degree result. Finally, a new weighting function is designed based on the contribution of each cluster member to each category.

Present, some existing selective clustering fusion algorithms cannot take into account the quality and diversity of clustering members, and the accuracy of the algorithm for high-dimensional data clustering needs to be further improved. For these two reasons, this paper puts forward a new selective clustering algorithm to realize the high-dimensional data clustering, a semi-supervised clustering method using a small amount of tag data to improve the performance of clustering algorithms, has gradually become a research focus in the pattern recognition and related areas. At the same time, a selection strategy based on reference members is proposed, and the quality and diversity of cluster members are taken into account.

The organization structure of the rest of this paper: section 2 gives the description and related definitions of selective clustering fusion algorithm; Section 3 introduces the framework of selective clustering fusion algorithm based on semi-supervised k-means clustering; Section 4 tests the performance of the algorithm through experiments, and analyzes the experimental results. Section 5 is a summary of this paper.

2. Description and Definition of the Selective Clustering Fusion Algorithm

2.1 Description of the Selective Clustering Fusion Process

Definition 1(selective clustering fusion\(^9\)): given a set of members of a cluster, the purpose of selective clustering fusion is to select some members of the cluster. Fusing some members of the cluster will generally achieve better clustering results than fusing all members of the cluster.

According to the above definition, compared with the traditional clustering fusion algorithm, selective clustering fusion algorithm increases the process of "selecting part of the clustering members", and then fuses the selected part of the clustering members to obtain the final clustering result. Therefore, the implementation of a selective clustering fusion algorithm includes three parts:(1) generation of clustering members;(2) selection strategy and selection of some clustering members;(3) merge the cluster members in step (2).

The selective clustering fusion problem can be described as: let \(X = (x_1, x_2, \ldots, x_n)\) be a set of \(n\) data points, we take \(H\) times different clustering algorithm or the same clustering algorithm with different parameters to cluster datasets, and then obtain the result \(l = \{l_1, l_2, \ldots, l_H\}\), where the \(i\)-th result is \(l_i = \{C_{i1}, C_{i2}, \ldots, C_{ik}\}\) \((i = 1, 2, \ldots, H)\), in which the data set are gathered into \(k\) clusters, select some cluster members from \(l\) to form a subset \(l' = \{l'_1, l'_2, \ldots, l'_{H'}\}\) \((l' \subseteq l)\) by some selection strategy \(D\); finally, get the final clustering results \(\Gamma\) by ensemble the cluster members \(l'\) using Co-association matrix \(G\). The fusion process of selective clustering is shown in Figure 1.
2.2 semi-supervised k-means clustering based on the projection matrix

In the application of data mining and machine learning, especially for unsupervised learning, the high dimension of data sets is often faced with “dimension disaster”. One manifestation of “disaster” is that, in higher dimensions, almost all pairs of points are roughly the same distance apart; Another way to think about it is that almost any two vectors are almost orthogonal to each other \[17\]. Therefore, it is often necessary to preprocess the high-dimensional data set and delete redundant and irrelevant attributes, to reduce the dimension of the data set and simplify the data set. Inspired by literature \[12\], this paper proposes a semi-supervised k-means clustering method based on the projection matrix. Firstly, a projection matrix is obtained by using the given pair constraint set of must-link and cannot-link. In the projection space, the semi-supervised k-means algorithm is used to reduce the dimension and cluster the high-dimensional data set to generate cluster members.

It is necessary to high-dimensional data into low-dimensional space to cluster data in high-dimensional space. Suppose \(d\)-dimensional data set \(X = \{x_1, x_2, \ldots, x_n\}\), \(W_{d,l} = [W_1, W_2, \ldots, W_l]\) is a projection matrix, which contains \(d\) dimensional orthogonal unit vectors. To project the original data into a low-dimensional space:

\[
y_i = W^T x_i \in \mathbb{R}^l, \quad l < d
\]

The projection matrix should not only preserve the structure of the original data as much as possible in the projection space, but also maximize the distance between the point pairs in the cannot-link constraint set and minimize the distance between the point pairs in the must-link constraint set.

Definition 2 Define the objective function \(J(W)\),

\[
J(W) = \frac{1}{2|C|} \sum_{(i,j) \in C} \left\| W_x^T - W_y^T \right\| - \frac{1}{2|M|} \sum_{(i,j) \in M} \left\| W_x^T - W_y^T \right\|^2
\]

\[
= \sum_{i} W_i^T \left[ \frac{1}{2|C|} \sum_{(i,j) \in C} (x_i - x_j)(x_i - x_j)^T - \frac{1}{2|M|} \sum_{(i,j) \in M} (x_i' - x_j')(x_i' - x_j')^T \right] W_i
\]

(2)

Where, \(|C|\) and \(|M|\) represent cannot-link and must-link as the number of point pairs in the set.

Let \(D = \frac{1}{2|C|} \sum_{(i,j) \in C} (x_i - x_j)(x_i - x_j)^T - \frac{1}{2|M|} \sum_{(i,j) \in M} (x_i' - x_j')(x_i' - x_j')^T\), \(W_i^T W_i = \begin{cases} 1, & i = j \\ 0, & \text{otherwise} \end{cases}\).

The Lagrange function is constructed by KT theorem:

\[ L(W) = J(W_1, W_2, \ldots, W_l) - \sum_{i=1}^l \lambda_i (W_i^T W_i - 1) \]

Take the part of \(L(W)\):
\[
\frac{\partial L}{\partial W_i} = 2D W_i - 2\lambda W_i = 0, \forall i = 1, \cdots, I
\]
\[
D W_i = \lambda W_i, \forall i = 1, \cdots, I
\] (3)

The optimal projection matrix \( W_{std} = [W_1, W_2, \cdots, W_I] \) can be solved from equation (3), which are composed by the eigenvectors corresponding to the \( I \) maximum eigenvalues of matrix \( D \). Then, formula (1) projects the original data into a low-dimensional space to reduce the dimensionality of high-dimensional data; furthermore, the semi-supervised k-means clustering method is used to realize the clustering of data in low-dimensional space. In the low-dimensional space, a semi-supervised k-means clustering method is repeatedly used to cluster the target data to generate clustering members to participate in the subsequent fusion. The specific process of semi-supervised k-means clustering algorithm\(^{[18]}\) is as follows:

**Step1.** Initialize \( k \) clustering centers \( C_1, \cdots, C_k \).

**Step2.** For each data \( x_i \) in dataset \( X \), assign it to the closest class \( C_j \), and make sure the return value of the function \( \text{Violate constraint}(x_j, C_j, ML, CL) \) is false. If no such class is found, call the function \( \text{Process violation}(x_i, CL) \).

**Step3.** Calculate the mean value of data points allocated to each class \( C_i \), and update the center point of class \( C_i \).

**Step4.** Repeat Step2 and Step3 until the algorithm converges or reaches the iteration number \( T \).

**Step5.** Return the final clustering result \( C_1, C_2, \cdots, C_k \).

\( \text{Process violation}(x_i, CL) \):

1. Find the class \( C \) which closest to data \( x_i \);
2. For each sample that has a cannot-link constraint relationship with \( x_i \), find the sample that has been assigned to class \( C \) from this sample and save it in set \( S \);
3. Cancel the class labels of all data samples in set \( S \) and assign \( x_i \) to class \( C \);
4. For each sample \( x_s \) in set \( S \), reassign it to the nearest class except class \( C \), and require that the cannot-link constraint is still violated after \( x_s \) is assigned to the class.

2.3 Reference to Members and Selection Strategy

To select cluster members that can improve the fusion quality, the key points of the selection strategy are as follows: (1) design a method to measure the quality of cluster members and the degree of difference between them; (2) design suitable selection principle\(^{[19]}\). When measuring the quality and difference degree of cluster members, it is necessary to consider what is the standard for measurement, that is, the selection of reference members. In many previous studies, the preliminary fusion results obtained by the simple clustering fusion algorithm of all cluster members are usually taken as the reference members. The disadvantage of this method is that it fails to eliminate the interference of inferior cluster members, and the fusion results obtained by referring to the selected cluster members are not ideal.

In this paper, a new selection method of reference members is proposed based on the effectiveness evaluation method of clustering, and the selection strategy is designed based on this method.
2.3.1 Reference Member. A good cluster member has a high degree of separation\[20\] (between different clusters) and tightness (relative to the inside of the cluster), so choosing a good cluster member as a reference member can eliminate the interference of inferior members and improve the quality of the cluster. This paper introduces the concepts of cluster density and separation degree\[21-23\] to evaluate the quality of cluster members.

Let $(x_1, x_2, \ldots, x_n)$ be a set of $n$ data points, we take $H$ times different clustering algorithms or the same clustering algorithm with different parameters to cluster data sets, and then obtain the result $l = \{l_1, l_2, \ldots, l_H\}$, where the $i$-th result is $l_i = \{C_i^1, C_i^2, \ldots, C_i^k\} (i = 1, 2, \ldots, H)$, in which the data set are gathered into $k_i$ clusters, then the average value of the data set is called $\bar{X}$, 

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i;$$

where the standard deviation of the data set is called $\text{Var}(X)$ and it can be defined as

$$\text{Var}(X) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} d^2(x_i, \bar{X})}.$$ 

The definitions of compactness and separateness of cluster members are given below.

Definition 3 The compactness of the $i$-th clustering result $l_i = \{C_i^1, C_i^2, \ldots, C_i^k\}$ is given by:

$$\text{Cmp}_i = \frac{1}{k_i} \sum_{j=1}^{k_i} \frac{\text{Var}(C_j^i)}{\text{Var}(X)}$$

Where $k_i$ is the number of clusters in the clustering member $l_i$, $\text{Var}(C_j^i)$ is the variance of the $j$-th cluster $C_j^i$ in the $l_i$. The bigger the $\text{Cmp}_i$, the closer the cluster members in the class $l_i$ are, and the better the clustering result is.

Definition 4 The separation of the $i$-th clustering result $l_i = \{C_i^1, C_i^2, \ldots, C_i^k\}$ is given by:

$$\text{Sep}_i = \frac{1}{k_i(k_i-1)} \sum_{j=1}^{k_i} \sum_{p=j+1}^{k_i} \exp \left( -\frac{d^2(\bar{X}_{C_j^i}, \bar{X}_{C_p^i})}{2\sigma^2} \right)$$

Where $d$ is Gauss constant (setting $2\sigma^2 = 1$), $C_p^r$ is the $p$-th cluster of $l_i$; $\bar{X}_{C_p^r}$ is the center of $C_p^r$, $d(\bar{X}_{C_j^i}, \bar{X}_{C_p^i})$ is the distance of the center of $C_j^i$ and the center of $C_p^i$. The smaller the $\text{Sep}_i$, for the class $l_i$, the higher the discrete degree among the different classes, and the better the clustering result is.

To validate the quality of clusters objectively, we combine the compactness and separation degree of clustering members to evaluate the result.

Definition 5 The comprehensive evaluation quality of the $i$-th clustering result $l_i = \{C_i^1, C_i^2, \ldots, C_i^k\}$ can be evaluated by comprehensive evaluation function as follows:

$$\text{TQC}_i = 1 - [\alpha \ast \text{Cmp}_i + (1 - \alpha) \ast \text{Sep}_i]$$

Where $\alpha$ is the weight factor, which can measure the weights of clustering members’ compactness and separation degree. The smaller the $\text{TQC}_i$ is, the bigger the compactness of $l_i$ is, the smaller the separation of $l_i$ is, and the better the quality of the clustering member is.

Utilize Eqs. (4)-(6) to calculate the comprehensive evaluation function values of all available clustering members, and then rank these values in descending order. The clustering member $l^*$ as the reference partition, which corresponds to the minimum value $\text{TQC}_i$.

2.3.2 Selection Strategy. A good selection strategy is necessary to take into account the quality of clustering, but also should consider diversity. Normalized mutual information (NMI) among the clustering members partly depicts the compactness of the clustering between individuals\[14\]. In this paper, we assess the quality and diversity of clustering members by NMI and reference partition.
Definition 6 A data set \( X = (x_1, x_2, \ldots, x_n) \) is divided into two clustering members \( l^{(a)} = \{C_1^{(a)}, C_2^{(a)}, \ldots, C_k^{(a)}\} \) and \( l^{(b)} = \{C_1^{(b)}, C_2^{(b)}, \ldots, C_k^{(b)}\} \) after two times clustering. Suppose that there are \( n_l \) and \( n_j \) data points included in \( C_i^{(a)} \) and \( C_j^{(b)} \) respectively, and there are \( n_{ij} \) same data points in \( C_i^{(a)} \) and \( C_j^{(b)} \), then the NMI between \( l^{(a)} \) and \( l^{(b)} \) is:

\[
\Phi_{NMI}(l^{(a)}, l^{(b)}) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} \log \left( \frac{n_{ij}}{n_i n_j} \right)
\]

(7)

Definition 7 A data set \( X = (x_1, x_2, \ldots, x_n) \) is divided into \( H \) clustering members \( l = \{l_1, l_2, \ldots, l_H\} \) after \( H \) times clustering, where the \( i \)-th clustering result is \( l_i = \{C_1^i, C_2^i, \ldots, C_k^i\} \) \((i=1,2,\ldots,H)\), then the quality \( Q(l_i) \) of each clustering member can be defined as:

\[
Q(l_i) = \frac{1}{H} \sum_{i=1}^{H} \Phi_{NMI}(l_i, l^*)
\]

(8)

The bigger the \( Q(l_i) \), the more similar \( l_i \) and reference partition \( l^* \), and the better the clustering quality of \( l_i \) is.

Definition 8 The diverse degree \( D(l_i) \) of the \( i \)-th clustering result \( l_i = \{C_1^i, C_2^i, \ldots, C_k^i\} \) \((i=1,2,\ldots,H)\) can be defined as:

\[
D(l_i) = 1 - \frac{\sum_{j=1}^{k} \text{NMI}(l_i, l_j)}{H-1}
\]

(9)

The higher the \( D(l_i) \), the bigger the diverse degree of \( l_i \) with other clustering members is.

Definition 9 The comprehensive evaluation function joining the clustering quality and diverse degree of clustering member is defined as:

\[
F(l_i) = \beta * Q(l_i) + (1-\beta) * D(l_i)
\]

(10)

where \( \beta \) is a balance factor, for convenience, we set \( \beta = 0.5 \).

Based on Eqs. (7)-(10), we can calculate \( F(l_i) \) for all clustering members. By ranking all clustering members in descending order and select first \( H' \) best ones to build the ensemble committee. Since the qualities of these \( H' \) clustering members are different, we propose the weights for these \( H' \) clustering members.

Definition 10 For ensemble committee set \( l' = \{l'_1, l'_2, \ldots, l'_{H'}\} \) \((l'_i \subseteq l)\), the weight of each clustering member \( l'_i \) is defined as:

\[
\omega_i = \frac{F(l'_i)}{\sum_{i=1}^{H'} F(l'_i)}
\]

(11)

2.4 Weighted Co-association Matrices

We apply an ensemble algorithm to calculate the selective clustering member set \( l' = \{l'_1, l'_2, \ldots, l'_{H'}\} \) \((l'_i \subseteq l)\), and get the final clustering results. In this paper, we use a weighted Co-association matrix to achieve fusion.

Definition 11 Suppose that \( Co \) is a weighted Co-association matrix and \( Co(i,j) \) is the element of \( Co \):
$Co(i, j) = \frac{\sum_{p=1}^{k} g(i, j, p) \omega_p}{H'}$  \hspace{1cm} (12)

$g(i, j, p) = \begin{cases} 1, & i \in C^p_i \text{ and } j \in C^p_j \\ 0, & i \notin C^p_i \text{ or } j \notin C^p_j \end{cases}, \ m = 1, 2, \cdots, k$ \hspace{1cm} (13)

where $g(i, j, p) = 1$ means that data point $i$ and $j$ belong to the same cluster $C^p$, otherwise, $g(i, j, p) = 0$. $m$ is the number of clusters, $p$ is the number of selective cluster members, and $\omega_p$ is the corresponding weight. The weight of clustering members as a measurable index can produce better ensemble results relative to the traditional Co-association matrix \cite{24}.

3. Selective Clustering Fusion Algorithm Framework based on Semi-supervised K-means Clustering

In this paper, a selective clustering ensemble algorithm for high-dimensional data is proposed. The idea of this algorithm is as follows: firstly, a projection matrix is obtained by using the given pair constraint set of must-link and cannot-link. In the projection space, the semi-supervised k-means algorithm is used to reduce the dimension and cluster the data set, and generate the cluster members. According to the evaluation criteria of reference members and selection strategy, select some cluster members with better quality. Finally, the weighted Co-association matrix is used to realize the fusion, and we can obtain the final clustering result.

The framework of this algorithm is as follows:

**Step1.** First, the objective function $J(W)$ is constructed based on the point pairs of the pair constraint set in the combination of cannot-link and must-link.

**Step2.** Solve the objective function $J(W)$ according to the KT theorem, and get the optimal projection matrix $W_{td} = [W_1, W_2, \cdots, W_H]$.

**Step3.** Use equation (1) to project the original data into a low-dimensional space.

**Step4.** In the low-dimensional space, use a semi-supervised k-means clustering algorithm to cluster the data set $X$ (runs $H$ times) and get a collection $l = \{l_1, l_2, \cdots, l_H\}$ which contains $H$ clustering partitions.

**Step5.** Calculate the comprehensive evaluation function values $TQC_l$ of each clustering member according to Eqs. (4)-(6), choose the minimum value of the corresponding clustering members as a reference partition $l'$.

**Step6.** Calculate the values $F(l_i)$ of each clustering member according to Eqs. (7)-(10), and then choose the first $H'$ cluster members $l' = \{l'_1, l'_2, \cdots, l'_H\} (l' \subseteq l)$.

**Step7.** Calculate the weights $\omega_i$ of the clustering members which have been selected according to Eq. (11).

**Step8.** Calculate the weighted Co-association matrix according to Eqs. (12)-(13), and set the similarity threshold $t$. For each pair of points $(i, j)$, if $Co(i, j) > t$, then point $i$ and $j$ belong to the same cluster, combine clusters when two points in different clusters; if $Co(i, j) \leq t$, then two points can’t be classified as a cluster, isolated point separately forms a cluster. Get the final clustering results by comparing the relationship between all pairs of points with the threshold value.

4. Experimental design and result analysis

4.1 Experimental environment

The hardware environment tested in this paper is Inter(R) Core(TM)2 Duo CPU E8300 2.83GHz 1.98GB, the operating system is Windows 7, programming tool is Matlab(R2011a) for experimental testing of the algorithm.
4.2 Data sets

To validate the feasibility and effectiveness of the proposed algorithm, there are five data sets available from the UCI Machine Learning Repository. The five data sets are both different in quantity and properties which can reflect the advantages and disadvantages of the proposed algorithm both in data volume and attribute. All data sets have referenced the class labeling and the detailed information on experimental data sets can be seen in Table 1. TSE describes the data set of Turkiye Student Evaluation; WLEMIMU represents the data set of Weight Lifting Exercises monitored with Inertial Measurement Units.

| Data sets     | Data set characteristics       | Instances | Features | Classes |
|---------------|--------------------------------|-----------|----------|---------|
| Wine          | Multivariate                   | 178       | 13       | 3       |
| Waveform(version1) | Multivariate, Data-Generator | 5000      | 21       | 3       |
| TSE           | Multivariate                   | 5820      | 33       | 3       |
| Libras Movement | Multivariate, Sequential      | 360       | 91       | 15      |
| WLEMIMU       | Multivariate                   | 39242     | 152      | 5       |

In the experiment, the constraint set of pair points is obtained by random extraction. When the constraint set is empty, it is expressed as unsupervised clustering integration.

4.3 Evaluation criteria

In this paper, the classic Micro-precision evaluation criterion is used as follows:

\[ \text{micro} - p = \frac{1}{n} \sum_{i=1}^{k} a_i \]  

Where \( k \) represents the number of clusters in the data set; \( a_i \) is the number of data points that are correctly assigned to the \( i \)-th cluster; \( n \) is the number of points in the data set. The bigger the value of \( \text{micro} - p \), the better the clustering performance.

4.4 Analysis of experimental results

This paper makes an experimental test of the selective clustering ensemble algorithm based on semi-supervised K-means clustering. By being compared to the experimental test results of the other two algorithms simple k-means clustering algorithm and clustering ensemble algorithm based on k-means, it proves the feasibility and effectiveness of the proposed algorithm.

Experiment 1: Running k-means algorithm and the Semi-supervised k-means clustering algorithm based on projection matrix on five data sets to validate the effectiveness of using a semi-supervised clustering algorithm to cluster the high dimensional data by comparing the accuracy of the clustering results. Running each algorithm 10 times on five data sets which were randomly selected 10% pair constraints, then calculate the average results of the 10 times, the results of the experiment are indicated in Table 2. Where k-means represents a simple k-means algorithm, and PSKC represents Semi-supervised k-means clustering algorithm based on projection matrix.

| Data sets     | Wine     | Waveform(version1) | TSE     | Libras Movement | WLEMIMU   |
|---------------|----------|--------------------|---------|-----------------|-----------|
| K-means       | 0.6322   | 0.6025             | 0.6067  | 0.7345          | 0.5899    |
| PSKC          | 0.6933   | 0.6850             | 0.7276  | 0.8534          | 0.7211    |

By comparing the \( \text{micro} - p \) value of the two algorithms, we can observe that: (1) Compared with k-means clustering, PSKC can significantly improve the clustering performance. (2) Since the data volume and attributes of each data set are distinct, the increasing rate of accuracy of different data set is different. As attributes of data sets Wine and Waveform (version 1) are smaller, the increasing rate
of clustering accuracy is less improved when running PSKC. But for the data sets TSE, Libras Movement and WLEMIMU, the rate of clustering accuracy is greatly improved with the increasing of the attribute value. Therefore, PSKC is feasible and effective, especially for the high dimensional data.

Experiment 2: Making an experimental test for the algorithm proposed in this paper and comparing the accuracy of the clustering results with the K-means algorithm, clustering ensemble algorithm based on the K-means, clustering ensemble algorithm based on semi-supervised K-means clustering, selective clustering ensemble algorithm based on semi-supervised K-means clustering (without reference partition) and proposed the algorithm by this paper. The first two algorithms belong to unsupervised clustering, while the last three algorithms belong to semi-supervised clustering.

The parameters of this experiment are set as follows: the maximum number of iteration is 100. The error threshold is 1e-5. Running each algorithm 10 times respectively on five data sets which were randomly selected 10% pair constraints, then calculate the average results of the 10 times, the scale of the cluster members is placed at 100 and set ensemble scale to 50. The results are presented in Table 3 and Figure 2. Where k-means represents simple k-means algorithm, K-means ensemble represents clustering ensemble algorithm based on the K-means, SKC ensemble represents clustering ensemble algorithm based on semi-supervised K-means clustering, SKC SE(without reference partition) represents selective clustering ensemble algorithm based on semi-supervised K-means clustering (without reference partition) and SKCSE represents selective clustering ensemble algorithm based on semi-supervised K-means clustering.

| Data sets      | K-means | K-means ensemble | SKC ensemble | SKCSE(without reference partition) | SKCSE |
|----------------|---------|------------------|--------------|-----------------------------------|-------|
| Wine           | 0.6624  | 0.6578           | 0.7077       | 0.7233                            | 0.7734|
| Waveform(version) | 0.5956  | 0.6022           | 0.6856       | 0.7289                            | 0.7344|
| TSE            | 0.5903  | 0.5989           | 0.7128       | 0.7356                            | 0.7505|
| Libras Movement | 0.7440  | 0.7721           | 0.8530       | 0.8728                            | 0.8890|
| WLEMIMU        | 0.5638  | 0.5901           | 0.7146       | 0.7657                            | 0.7529|

Figure 2. Comparison of the clustering results with five kinds of algorithms.

From Table 3 and Figure 2, we can note that the results generated by the proposed algorithm run on the five data sets have improved in different amplitudes and the results are preferable to the other four algorithms. Particularly, compared to the results generated by SKCSE(without reference partition), it can be found that the selection of reference members is increased in this algorithm, which is effective and feasible to improve the accuracy of clustering results.

Experiment 3: This experiment is mainly to test the impact of diverse ensemble size for the
algorithm presented in this paper. Supposing the scale of clustering member is 100, we test the effect of ensemble sizes 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100 by calculating the accuracy of the clustering ensemble algorithm based on semi-supervised K-means clustering (SKC ensemble), selective clustering ensemble algorithm based on semi-supervised K-means clustering (without reference partition) (SKC selective ensemble) and the algorithm of this paper as shown in Figure 3.

![Figure 3. The influence of ensemble size on the performance of the clustering ensemble](image)

We can observe some interesting phenomenon from Figure 3: (1) By the clustering accuracy curve graph on five data sets, the results of the selective clustering ensemble algorithm especially the algorithm in this paper is superior to the clustering ensemble if the appropriate ensemble size is selected. (2) When the ensemble size is small, the results of the selective clustering ensemble are not better than cluster all members, as the ensemble size is very small, it cannot contain the information of all clustering members, the information loss more, so the clustering accuracy is lower. (3) Achieving the maximum accuracy, the ensemble size may be different for different data sets using different clustering algorithms. For example, the data set Wine uses SKC selective ensemble, the clustering accuracy is the highest when the integration scale is 50; if the data set Wine uses this algorithm for clustering, the clustering accuracy is the highest when the integration scale is 40. (4) For data set WEITEMMU, when the ensemble size achieves 50, the accuracy of the proposed algorithm is lower than SKC selective ensemble. It may be related to the effect of the data set itself, clustering algorithm or the ensemble size.

5. Conclusions

The selective clustering ensemble algorithm based on semi-supervised K-means clustering which is fit for the high-dimensional data has been proposed in this paper. This algorithm uses the given pairs constraints set of must-link and cannot-link to get a projection matrix, and projects the high-dimensional space data to the low-dimensional space, uses the semi-supervised K-means clustering algorithm in projection space for data collection, improve the accuracy of the initial cluster member. In the design of the selection strategy, the algorithm put forward by the new selection strategy based on the reference members which eliminates the inferior quality clustering member’s influence distinctly and obtains the clustering members with higher accuracy. It’s noted that although clustering accuracy is proved, the algorithm’s time complexity is also high. So the future work is about how to optimize the selective clustering ensemble algorithm and reduce the time complexity of the algorithm so that the algorithm has better applicability.

6. Acknowledgments

The work was partially supported by grants from the Key Research Project of Anhui Education Department (No. KJ2019A0836) and the Talent Research Fund Project of Hefei University (No. 16-17RC16). The authors would like to thank the comments and suggestions from the reviewers.
References

[1] Zhiwei Ni, Li-ping NI, Huiting Liu, and Ruiyu Jia. Dynamic data mining, 2010.
[2] YANG Lin-yun, ZHOU Hai-jing, ZHUO Qing, and WANG Wen-yuan. Weighted cluster ensemble based on significance of attribute [J]. Computer Science, 4(2009) pp. 067.
[3] Lin-yun Yang and Wen-yuan Wang. Clustering ensemble approaches: An overview. Application Research of Computers, 12(2005) pp. 8-14.
[4] Jiang Shengyi. Cluster fusion algorithm based on majority voting mechanism. Journal of Chinese Computer Systems, 2(2007) pp. 306-309.
[5] Zhao Ziyi Zou, Yuanqiang, and Li Guowei. New clustering algorithm based on combination of genetic algorithm and ant colony algorithm. Science Technology and Engineering, 6(2006) pp. 4700-4704, 4713.
[6] Fern X, Brodley C. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach[J]. Proceedings of ICML- 2003. Washington DC, 2003,186-193.
[7] Zhou Z H, Wu J, Tang W. Ensembling neural networks: Many could be better than all[J]. Artificial Intelligence. 2002, 137(1-2):239-263.
[8] Hadjitodorov S, Kuncheva L I, Todorova L P. Moderate Diversity for Better Cluster Ensembles[J]. Information Fusion Journal, 2006, 7(3): 264-275.
[9] Fern X Z, Lin W. Cluster ensemble selection[J]. Statistical Analysis and Data Mining. 2008, 1(3):128-141.
[10] Zhang X, Jiao L, Liu F, Bo L, Gong M. Spectral clustering ensemble applied to SAR image segmentation[J]. Geoscience and Remote Sensing, IEEE Transactions, 2008, 46(7):2126-2136.
[11] Fern X, Brodley C. Clustering ensembles for high dimensional clustering: An empirical study[J]. Technical report CS06-30-02, Oregon State University.
[12] Yin xuesong, Hu enliang, Chen songcan. Discriminative Semi-Supervised Clustering Analysis with Pairwise Constraints[J]. Journal of Software, 2008, 19(11):2791-2802.
[13] Javad, Azimi, XiaoliFern. Adaptive Cluster Ensemble Selection[C]//Proc of the 21st International Joint Conference on Artificial Intelligence, IJCAI 2009, pp. 993-997.
[14] Tang Wei, and Zhou Zhihua. Bagging-based selective cluster ensemble. Journal of Software, 16(4): 496-502, 2005.
[15] Xiaoping Fan, Rongfen Sheng, Zhifang Liao, and Limin Liu. Selective and weighted clustering fusion algorithm. Computer Engineering and Applications, 2012, 48(22): 195-200.
[16] Zhi-Zhou Kong and Zi-Xing Cai. Sub-grouping and selecting method of cluster fusion. Control and Decision, 2013, 27(3): 369-373.
[17] Anand Rajaraman, Jeffrey David Ullman. Big data: Internet large-scale data mining and distributed processing[M]. Beijing: Posts and Telecom Press, 2012.
[18] Tan wei. Research on clustering ensemble and semi-supervised clustering in data mining[D]. ChengDu: Computer application technology, Southwest jiaotong university, 2010.
[19] Liu limin. Study on Clustering Ensemble Selection Algorithm[D]. ChangSha: School of information science and engineering, central south university, 2013.
[20] Kuncheva L I, Hadjitodorov S T. Using diversity in cluster ensemble[C]//IEEE International Conference on System, Man and Cybernetics, 2(2004) pp. 1214-1219.
[21] HE Ji, TAN A H, TAN C L. Modified ART 2A growing network capable of generating a fixed number of nodes [J]. IEEE Trans on Neural Networks, 2004, 15(3):728-737.
[22] HE Ji, TAN A H, TAN C L, etal. On quantitative evaluation of clustering systems[C]//Proc of Information Retrieval and Clustering Boston, MA: Kluwer, 2003.
[23] Yang yan, Jin fan. KAMEL M. Survey of clustering validity evaluation [J]. Application Research of Computers, 2008, 25(6):1630-1638.
[24] Zhang zhen, Liang yongquan, Zhang xinglin. Community Mining In Dynamic Complex Network: Selective Clustering Fusion Algorithm [J]. Computer & Digital Engineering, 2013,41(3):388-390.