A Semi-supervised Text Clustering Algorithm Based on Pairwise Constraints*

Jiang Zhong\textsuperscript{a,d,*},  Gaofeng Dong\textsuperscript{a}, Ying Zhou\textsuperscript{b}, Xue Li\textsuperscript{c}
Longhai Liu\textsuperscript{a}, Qiang Chen\textsuperscript{b}

\textsuperscript{a}College of Computer Science, Chongqing University, Chongqing 400044, China
\textsuperscript{b}Chongqing Management Center of Housing Provident Fund, Chongqing 400015, China
\textsuperscript{c}School of Information Technology and Electrical Engineering, Queensland University Queensland QLD 4072, Australia
\textsuperscript{d}Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing 400044, China

Abstract

In this paper, an active learning method which can effectively select pairwise constraints during clustering procedure was presented. A novel semi-supervised text clustering algorithm was proposed, which employed an effective pairwise constraints selection method. As the samples on the fuzzy boundary are far away from the cluster center in the clustering procedure, they can be easily divided into the wrong clusters. Therefore, we choose the pairwise constraint points from the fuzzy boundary to guide the clustering process towards appropriate partition. The experimental results show that the proposed algorithm can effectively improve the text clustering results by using the same amount of pairwise constraints.

Keywords: Text Clustering; Semi-supervised Learning; Pairwise Constraints Selection; Fuzzy Boundary

1 Introduction

Clustering is a process of dividing data set into different clusters according to the features of the data objects and making objects with high similarity into the same cluster, while making objects with high dissimilarity into different clusters. As a kind of unsupervised learning, clustering algorithms don’t employ any supervised information drawn from labeled samples. On the contrary, supervised learning, such as classification algorithms only use labeled samples and could not utilize those information implied in the unlabeled samples. The significant merit of semi-supervised

\*Project supported by the Chinese National Natural Science Foundation Key Program (No. 90818028), the Natural Science Foundation Project of CQ CSTC (2010BB2046, 2009BB2184) and the Third Stage Building of “211 Project” (Grant No. S-10218).

*Corresponding author.

Email address: zhong.jiang@cqu.edu.cn (Jiang Zhong).

1548–7741/ Copyright © 2011 Binary Information Press
June 2011
learning could take advantages of labeled and unlabeled samples. The Semi-Supervised Kernel Fuzzy C-Mean (SSKFCM) algorithm [1] was proposed by Huaxiang Zhang et al, which could optimize semi-supervised clustering for both labeled and unlabeled data.

Semi-supervised clustering algorithms are trying to use some supervised information to improve the clustering performance [2]. The supervised information in the semi-supervised clustering may be labeled samples or constraint information acquired from whether a pair of samples having the same class labels or not. Comparing with the unlabeled samples data, the supervised information is difficult to get in real applications. On the other hand, the performance of semi-supervised clustering algorithm mainly depends on the supervised information. So the selection of supervised information is very critical for active learning. However, the traditional semi-supervised clustering algorithms face some challenges. The adaptive semi-supervised clustering kernel method based on metric learning (SCKMM) [3] faces the problem that the computational complexity of solving metric matrix is expensive. The pairwise constrained k-means algorithm (PCKmeans) [4] relies on suitable parameters selection, which needs domain knowledge. The constrained k-means algorithm (CKmeans) [5] requires a lot of constraint information to obtain satisfied results.

This paper presented an active learning algorithm employing an effectively pairwise constraints selection method. In order to demonstrate the efficiency of the active learning algorithm, a semi-supervised text clustering method based on pairwise constraints was proposed. During the active clustering algorithm, the pairwise constraints selection and clustering procedure are executed alternately. The experimental results show that the proposed method can effectively improve the text clustering results by using the same amount of pairwise constraints. The organization of the rest of the paper is as follows. In Section 2, the active text clustering method is introduced, which includes the key idea of pairwise constraints selection method. Some experimental results are given and compared with other clustering methods in Section 3. At last, conclusions and future work are presented.

2 Active Semi-supervised Text Clustering Method Based on Pairwise Constraints

2.1 PCCA Clustering Algorithm

The Pairwise-Constrained Competitive Agglomeration (PCCA) algorithm [2, 6] is extended by Grira et al, which is based on the Competitive Agglomeration (CA) algorithm [7]. CA algorithm is a fuzzy clustering algorithm, and the idea of the algorithm is: firstly, divide the data set into many smaller clusters. As the clustering progresses, adjacent clusters compete on the sample points, and the failing clusters in the competition gradually become smaller until they disappear. At last, the number of clusters gradually decreases to an optimal number. PCCA algorithm introduces the pairwise constraints into the CA algorithm, expecting to use constraint information to guide the clustering process and further improve the performance of fuzzy clustering algorithm. According to the prior knowledge of whether two samples belong to the same class or not, creating two data sets $M$ and $D$. $M$ represents the point set, in which a pair of samples belong to the same class; $D$ represents the point set, in which a pair of samples belong to the different classes. That is, if $(x_i, x_j) \in M$, then $x_i$ and $x_j$ should be divided into the same class; If $(x_i, x_j) \in D$, then $x_i$ and $x_j$ should be divided into the different classes. The literature [2] gives the objective function of
The iteration formula of cluster center is as following:

\[ J(V, U) = \sum_{k=1}^{C} \sum_{i=1}^{N} u_{ik}^2 d^2(x_i, \mu_k) - \beta \sum_{k=1}^{C} \left[ \sum_{i=1}^{N} u_{ik} \right]^2 + \]

\[ \alpha \left( \sum_{(x_i, x_j) \in M} \sum_{k=1}^{C} \sum_{l=1,l \neq k}^{C} u_{ik} u_{jl} + \sum_{(x_i, x_j) \in C} \sum_{k=1}^{C} u_{ik} u_{jk} \right) \]

\[ s.t. \sum_{k=1}^{C} u_{ik} = 1, i = 1, 2, ..., N \]

In Eq. (1) and Eq. (2), \( C \) is the number of clusters; \( V = \{ \mu_k \mid k \in \{1, 2, ..., C\} \} \) is the cluster center points set; \( U = \{u_{ik} \mid i = 1, 2, ..., N; k = 1, 2, ..., C\} \) is the fuzzy membership matrix, \( u_{ik} \in [0, 1] \) is the fuzzy membership of the sample \( i \) to the cluster \( k \).

The objective function of PCCA algorithm consists of three items: the first item is FCM, which is the sum of weighted square distances. The second item is a competition item, which uses competition cardinality among the clusters to make the algorithm automatically get an optimal number of clusters. Cardinality refers to the sum of sample membership, which belongs to a class. For example, the cardinality of cluster \( k \) is \( N_k = \sum_{i=1}^{N} u_{ik} \). The first two items of Eq. (1) constitute the objective function of CA algorithm. The third item is the penalty cost item, that is, penalty cost of violating the known constraints in the clustering process, which includes two parts: a) the penalty cost of violating must-link, that is, when two samples in \( M \) are divided into the different clusters, the penalty cost is the product of membership of the two samples belonging to the different clusters; b) the penalty cost of violating cannot-link, that is, when two samples in \( D \) are divided into the same cluster, the penalty cost is the product of membership of the two samples belonging to the same cluster.

The literature [2] gives the definition of \( \alpha \) and \( \beta \), namely:

\[ \alpha = \frac{N}{M} \sum_{k=1}^{C} \sum_{i=1}^{N} u_{ik}^2 d^2(x_i, \mu_k) \]

\[ \beta(t) = \frac{\gamma(t) \alpha}{\sum_{j=1}^{C} \left[ \sum_{i=1}^{N} u_{ij} \right]^2} \sum_{j=1}^{C} \sum_{i=1}^{N} u_{ij}^2 d^2(x_i, \mu_j) + \]

\[ \alpha \left( \sum_{(x_i, x_j) \in M} \sum_{k=1}^{C} \sum_{l=1,l \neq k}^{C} u_{ik} u_{jl} + \sum_{(x_i, x_j) \in C} \sum_{k=1}^{C} u_{ik} u_{jk} \right) \]

\( \alpha \) is the weight coefficient of constraint item, reflecting the importance of constraint term in the objective function, which adjusts the importance of constraint item by means of the normalized performance index. The better normalized degree is, the smaller \( \alpha \) is; the worse normalized degree is, the bigger \( \alpha \) is. \( M \) is the number of pairwise constraints. \( \beta \) reflects the importance of competition item with respect to the other two items in the objective function. \( \beta \) changes as \( t \) changes, and \( t \) is the number of iterations. Using Lagrange multiplier method to minimize the objective function, we can obtain the iteration formulas of cluster center and membership matrix. The iteration formula of cluster center is as following:

\[ \mu_k = \frac{\sum_{i=1}^{N} u_{ik}^2 x_i}{\sum_{i=1}^{N} u_{ik}^2} \]
The iteration formula of membership matrix is as following:

\[ u_{rs} = u_{rs}^{FCM} + u_{rs}^{Bias} + u_{rs}^{Constraints} \]  \hspace{1cm} (6)

In Eq. (6),

\[ u_{rs}^{FCM} = \frac{1}{\sum_{k=1}^{C} \frac{1}{d^{2}(x_{r}, \mu_{k})}} \]  \hspace{1cm} (7)

\[ u_{rs}^{Constraints} = \frac{\alpha}{2d^{2}(x_{r}, \mu_{s})}(\overline{C\nu_{r}} - C\nu_{rs}) \]  \hspace{1cm} (8)

\[ u_{rs}^{Bias} = \frac{\beta}{d^{2}(x_{r}, \mu_{k})}(N_{s} - N_{r}) \]  \hspace{1cm} (9)

In Eq. (8),

\[ C\nu_{rs} = \sum_{(x_{r}, x_{j}) \in M} \sum_{l=1/l \neq s}^{C} u_{jl} + \sum_{(x_{r}, x_{j}) \in C} \mu_{js} \]  \hspace{1cm} (10)

\[ \overline{C\nu_{r}} = \frac{\sum_{k=1}^{C} \sum_{(x_{r}, x_{j}) \in M} \sum_{l=1/l \neq s}^{C} u_{jl} + \sum_{(x_{r}, x_{j}) \in C} u_{jk}}{\sum_{k=1}^{C} \frac{1}{d^{2}(x_{r}, \mu_{k})}} \]  \hspace{1cm} (11)

In Eq. (9),

\[ N_{s} = \sum_{i=1}^{N} u_{is}, \quad N_{r} = \frac{\sum_{k=1}^{C} N_{k}}{\sum_{k=1}^{C} \frac{1}{d^{2}(x_{r}, \mu_{k})}} \]

From Eq. (6), it can be known that the membership of PCCA algorithm consists of three parts. \( u_{rs}^{FCM} \) represents the membership of fuzzy C means; \( u_{rs}^{Bias} \) represents the membership of competition item. These two items constitute the membership of CA algorithm. In Eq. (9), \( N_{s} \) is the cardinality of cluster \( s \), \( N_{r} \) is the average cardinality of all the clusters that sample \( r \) belongs to. \( u_{rs}^{Constraints} \) represents the membership of constraint item, \( C\nu_{rs} \) is the penalty cost of sample \( x_{r} \) belongs to cluster \( s \), and \( \overline{C\nu_{r}} \) is the average penalty cost of all the clusters that sample \( x_{r} \) belongs to. If \( C\nu_{rs} > \overline{C\nu_{r}} \), then \( u_{rs}^{Constraints} \) is negative, constraint item makes the membership of cluster \( s \) that sample \( x_{r} \) belongs to become smaller; If \( C\nu_{rs} < \overline{C\nu_{r}} \), then \( u_{rs}^{Constraints} \) is positive, constraint item makes the membership of cluster \( s \) that sample \( x_{r} \) belongs to become bigger. If sample \( x_{r} \) is neither in \( M \) nor in \( D \), then \( u_{rs}^{Constraints} = 0 \). \( u_{rs}^{Constraints} \) can increase or decrease the total membership of clusters that sample \( x_{r} \) belongs to, so on the basis of the membership of CA algorithm, we can adjust CA algorithm by overlapping the membership of constraint item.

### 2.2 Active Selection of Pairwise Constrains Method

Semi-supervised clustering algorithm studies how to use a small amount of supervised information to improve the performance of unsupervised clustering algorithm. Improving the clustering performance in semi-supervised clustering algorithm largely depends on the supervised information. In the clustering process, if the clustering algorithm can discover the supervised information provided by the users, such as the two pairs of constraints shown in the Fig. 1 (a), then it is very difficult to improve clustering performance by the supervised information [6].
In order to get with more useful guidance from pairwise constrains, we should focus on those fuzzily divided clusters, in which the distribution of sample points is not closely and the boundary between the clusters is not obvious. More specifically, we should concern about those fuzzily divided clusters and their common boundary between the adjacent clusters. In Fig. 1 (b), the cluster surrounded by a solid line is the most fuzzily divided cluster. The rectangular area called fuzzy boundary is the common boundary of the fuzzy cluster and its adjacent clusters. As the samples on the fuzzy boundary are far away from the cluster center, they can be easily divided into the wrong clusters. Therefore, the constraint information selected from the fuzzy boundary will have more useful guidance to the clustering algorithm. In Fig. 1 (c), the solid samples represent the samples on the fuzzy boundary, and the pair of samples surrounded by a rectangle is a pair of constraints selected from the fuzzy boundary, which should have belonged to the same cluster, while is divided into the different clusters. In the clustering process, by adjusting their memberships, they eventually are divided into the correct cluster, as shown in Fig. 1 (d).

Fig. 1: (a) the two pairs of samples connected by straight line are two pairs of constraint information provided by the user; (b) cluster $b$ is a fuzzy cluster, the area surrounded by the rectangle is fuzzy boundary; (c) a pair of constraint information surrounded by a rectangle is the constraint information selected from the fuzzy boundary; (d) the clustering result obtained under the guidance of constraint information.

Select the constraint information before each iteration of the clustering algorithm: firstly, find the current most fuzzily divided cluster, then select three samples on the boundary. Finally, for each selected sample, find the sample in its adjacent cluster, which is the nearest to it. The two samples constitute a pair of constraint information, and ask the user whether they belong to the same cluster or not. The specific steps of actively selecting pairwise constrains are as follows:

1. Find the most fuzzily divided cluster by the Fuzzy Hyper Volume (FHV) [8]. FHV is proposed by Gath et al, whose role is to evaluate the tightness of a fuzzy clustering. The smaller the value of FHV is, the closer the cluster is to the samples in the cluster; On the contrary, the greater the value of FHV is, the farther away from clustering center the samples in the cluster is. In each iteration, it will find the most fuzzily divided cluster with the greatest FHV. FHV is calculated as:

$$FHV = |C_k|$$  \hspace{1cm} (12)

In Eq. (12), $|C_k|$ represents the determinant value of covariance matrix $C_k$, where the covariance matrix $C_k$ is calculated as follows:

$$C_k = \frac{\sum_{i=1}^{N} u_k^2 (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^{N} u_k^2}$$  \hspace{1cm} (13)
In Eq. (13), \( N \) is the number of samples, \( u_{ik} \) indicates the membership of sample \( x_i \) which belongs to the cluster \( k \), \( \mu_k \) is the clustering center of cluster \( k \). In Fig. 1 (b) and Fig. 1 (c), the cluster whose boundary surrounded by a solid line has the largest FHV.

(2) After finding the most fuzzily divided cluster, select three samples on its boundary. According to the principle of maximum membership, if \( u_{ik} \) was the greatest membership of sample \( x_i \), then sample \( x_i \) should be divided into cluster \( k \). For the cluster \( k \), the samples located on its boundary refer to those samples having the smallest membership in the cluster. In each iteration, we select three samples on the boundary of the most fuzzily divided cluster. That is, select three samples having smallest membership from the most fuzzily divided cluster. In Fig. 1 (c), the samples shown in solid points are those samples having the smallest membership in the corresponding cluster.

(3) After selecting three points on the boundary, for each point selected on the boundary, find the nearest point in its adjacent cluster. The adjacent cluster of sample \( x_i \) is the cluster that the second largest membership of sample \( x_i \) corresponds. In Fig. 1 (c), cluster \( b \) is the most fuzzily divided cluster, where the sample \( x_i \) is the selected sample on the boundary and its adjacent cluster is cluster \( a \), where the sample \( x_j \) is the nearest to sample \( x_i \). Then the two samples constitute a pairwise constraints and ask the user whether \( (x_i, x_j) \) belongs to the same cluster or not. If they belong to the same cluster, then put them into the set \( M \); otherwise, put them into the set \( D \).

It can be seen from the above process, each iteration can bring out at most three pairwise constrains. After selecting the constraint information, we can use them to guide the clustering process.

### 2.3 Active Semi-supervised Text Clustering Based on Pairwise Constraints Method

By using the method of actively selecting pairwise constraints, an active semi-supervised text clustering method based on pairwise constraints was proposed. The detail of this method is described as following.

**Step 1** Preprocess the texts, and then use the Latent Semantic Indexing (LSI) to reduce the dimensionality of the text, which is represented by the Vector Space Model (VSM). VSM is a common representation model for text data, which has some disadvantages including the dimensionality of text feature vector is too high and ignoring the semantic links among words. LSI is a method of deducing the latent semantic link between terms from the co-occurrence term information. Therefore, this paper uses LSI to map the natural language text into a low-dimensional latent semantic space.

**Step 2** Initialize the PCCA algorithm: let \( C \) be the maximum number of clusters, randomly select \( C \) samples as the initial cluster centers, initialize the membership of each sample with equal probability and calculate the sum of membership of each cluster, which is the cardinality of each cluster.

**Step 3** Select the constraint information by the method proposed in section 2.2.
Step 4 Use Eq. (3) to calculate $\alpha$.

Step 5 Use Eq. (4) to calculate $\beta$.

Step 6 Use Eq. (6) to calculate the membership of each sample.

Step 7 Calculate the cardinality $N_s(s = 1, 2, ..., C)$ of each cluster; if the cardinality of a cluster is smaller than a given threshold value, then delete the cluster, update the number of clusters $C$ and re-normalize the membership matrix.

Step 8 Use Eq. (5) to calculate the cluster center of each cluster.

Step 9 Repeat Step 3 to Step 8 until $|PCCA(t) - PCCA(t - 1)| < 10^{-5}$, where $t$ represents the number of iterations.

Algorithm 1: The outline of semi-supervised text clustering algorithm based on pairwise constraints

The PCCA algorithm is not sensitive to the initial value of $C$, when the value of $C$ is greater than the correct number of clusters, the algorithm will generally converge to the correct number of clusters in the iterative process. In theory, $C$ can take a prodigious value. But the greater $C$ is, the more time converging to the correct number of clusters will cost.

3 Experimental Results and Analysis

In this paper, we adopted Chinese corpus of Fudan University as a data set. The corpus includes ten categories. They are environment, transportation, computer, education, military affairs, economy, medical, sports, art and politics. We randomly selected 2000 texts of the corpus as the test set of clustering.

Firstly, preprocess the text set such as doing word segmentation, then use LSI to reduce the dimensionality of the text and select 100 features from the feature set to constitute the feature vector, finally use the active semi-supervised text clustering method based on pairwise constraints to cluster the texts. In the experiment, $t_0 = 5$, $\eta_0 = 0.8$, $C = 15$.

The common evaluation indexes of clustering quality include F1 measure, average entropy, Mutual Information (MI) and clustering accuracy and so on. In this paper, we used clustering accuracy to evaluate the clustering quality.

To illustrate the effectiveness of the active semi-supervised text clustering method based on pairwise constraints, firstly, we compared the method proposed with the unsupervised KMEANS algorithm and CA algorithm. The experimental results are shown in Fig. 2. The results shown in Fig. 2 and Fig. 3 are the average results of several experiments.

It can be seen from Fig. 2, KMEANS is an unsupervised clustering algorithm, in the experiment let $K$ be equal to the correct number of clusters, the average clustering accuracy of this algorithm was 57.6%; CA algorithm is an unsupervised fuzzy clustering algorithm, the average clustering accuracy of which in the experiment was 61.3%. Compared with KMEANS algorithm and CA algorithm, the novel method achieved a higher clustering accuracy. The semi-supervised clustering algorithm can use a small amount of supervised information to guide the clustering process and improve the clustering performance. Under the guidance of the supervised information,
the clustering accuracy of the method proposed could reach to 79.4%, which was 21.8% higher than the accuracy of KMEANS clustering algorithm and 18.1% higher than the accuracy of CA clustering algorithm.

![Fig. 2: The comparison among active semi-supervised text clustering method based on pairwise constraints, unsupervised KMEANS algorithm and CA algorithm](image)

To further illustrate the effectiveness of the method, we have compared this method with the Pairwise Constraint Kmeans (PCKmeans) [9] algorithm. The PCKmeans method proposed by Basu is a variation of k-means clustering algorithm based on pairwise constraints, which adopts farthest-first traversal strategy to select k data points to initialize the K-means algorithm. The k data points are furthest away from the current data set. The experimental results are shown in Fig. 3.

![Fig. 3: The comparison between active semi-supervised text clustering method based on pairwise constraints and PCKmeans algorithm](image)

It can be seen from Fig. 3, compared with the PCKmeans algorithm, the method proposed could get a higher accuracy. In the experiment, the clustering accuracy of PCKmeans algorithm was from 57.6% to 66.8%, which was increased by 9.2%. The clustering accuracy of the method proposed was from 61.3% to 79.4%, which was increased by 18.1%.

4 Conclusions and Future Work

The performance of semi-supervised clustering algorithm based on pairwise constraints mainly depends on the supervised information which is included in the pairwise constraints. This paper presented an effectively heuristic pairwise constraints selection method by the useful guidance
for clustering results. Using this method, an active semi-supervised text clustering method was proposed. The experimental results show that the proposed method can effectively improve the text clustering results by using the same amount of pairwise constraints. In the subsequent work, in order to use a small amount of supervised information to get a better clustering result, we will further study how to choose supervised information with more guidance.

References


