Improving Semi-supervised Constrained k-Means Clustering Method Using User Feedback

Kavan Fatehi a,∗
Arastoo Bozorgi b
Mohammad Sadegh Zahedi c
Ehsan Asgarian d

a Information Processing and Knowledge Discovery Lab, Yazd University, Yazd, Iran.
b Faculty of Electrical and Computer Engineering, Shahid Beheshti University, G.C., Tehran, Iran.
c Faculty of Electrical and Computer Engineering, Tehran University, Tehran, Iran.
d Faculty of Electrical and Computer Engineering, Ferdowsi Mashhad University, Mashhad, Iran.

ARTICLE INFO

Article history:
Received: 08 February 2014
Revised: 28 June 2014
Accepted: 28 September 2014
Published Online: 27 February 2015

Keywords:
Clustering, Semi-supervised using user feedback, Active learning, Boundary data.

ABSTRACT

Recently, semi-supervised clustering methods have been considered by many researchers. In this type of clustering, there are some constraints and information about a small portion of data. In constrained k-means method, the user (i.e. an expert) selects the initial seeds. In this paper, a constraint k-means method based on user feedback is proposed. With the help of the user, some initial seeds of boundary data obtained from clustering were selected and then the results of the user feedback were given to the constrained k-means algorithm in order to obtain the most appropriate clustering model for the existing data. The presented method was applied to various standard datasets and the results showed that this method clustered the data with more accuracy than other similar methods.

© 2014 JComSec. All rights reserved.

1 Introduction

Data clustering, known as a popular pattern recognition techniques, has been used in a wide variety of fields, ranging from web mining, machine learning, image segmentation, and biometric recognition, to electrical engineering, mechanical engineering, remote sensing, and genetics [1]. Many attempts have been made during past years to improve the previous work in the clustering context [2]. Learning methods are divided into three categories: supervised learning, unsupervised learning, and semi-supervised learning. In supervised learning methods, the number of the categories and the samples from each category are clear from the beginning [3]; thus, the information about the target variable is predefined. In these methods, a collection of labeled data is specified as training data, based on which a model is created to predict the label of the other sets of data. In other words, the main goal of supervised methods is to ensure that the created model is adequately trained in the classifier section in order to place new samples in appropriate categories. In unsupervised methods, the aim is to cluster a set of unlabeled data based on a distance measure criterion (or similarity between data) in a way that the data within each cluster have the most similarity and the data distance between the clusters is maximum [3]. Therefore, no target variable has been pre-defined,
rather, the correlation and structure in the existing data are studied.

As the name suggests, semi-supervised learning is somewhere between unsupervised and supervised learning methods [4]. In this model, for creating a learner model, both labeled and unlabeled data are used; as a result, semi-supervised clustering improves the performance of clustering by learning from the labeled data [5]. In other words, semi-supervised learning is applied to both classification and clustering methods. In semi-supervised classification, the training data include both labeled and unlabeled data; the numbers of unlabeled data are much more than that of the labeled ones [5]. In semi-supervised clustering, the existing data are unlabeled; but, there are some constraints and information about the clusters. These information can be so-called must-link constraints, in which two instances $x_i$ and $x_j$ must be in the same cluster, and cannot-link constraints, in which $x_i$ and $x_j$ cannot be in the same cluster [4].

In another categorization, the learning methods are divided into active and passive groups. In active learning methods, which have great semantic similarity to supervised and semi-supervised learning, the goal is to design and analyse learning algorithms which effectively choose the samples to be labeled by an expert [6]. In active learning methods, training data consists of unlabeled data. The goal of the active learner is the same as that of a passive learner [7]. In active learner methods, the learner model can ask for the labels of the training data; actually, by selecting data for labelling, it chooses more accurate data for training. The key idea behind active learning is that the algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns [6].

Semi-supervised clustering with user feedback is closely related to active learning [9]. In active learning, the learning system attempts to select which data points, if labeled, would be most informative. In semi-supervised clustering, a human selects the data points and puts on them a wide array of possible constraints, instead of labels [9]. Indeed, the goal is to let the user guide the clustering process so that the best clustering model for the data is achieved.

In the current semi-supervised learning methods, the selection of initial seeds for the clustering algorithm is done randomly by the user, which may cause the data to be selected not from all the clusters; as a result, achieving a clustering model with a high degree of accuracy is not possible. Innovation in this paper is the more accurate way of selecting the initial seeds for the constrained k-means algorithm which results in increased accuracy of the algorithm. In this work, initial clustering was done on the data by applying k-means method and, then using the proposed method, boundary data were appropriately identified from each cluster and selection of the first seeds was done by the user on the boundary data. Results of the user feedback were submitted to constrained k-means algorithm and the clustering steps were guided in a perfect way to reach the most appropriate model for clustering the data.

The rest of the paper is organized as follows:

A review of clustering algorithms based on k-means is introduced in the second section of the paper. In the third section, the proposed clustering method is presented. Evaluation criteria for comparing different algorithms, profile of the datasets used for comparing the quality of the proposed clustering methods and the results of evaluating the implemented methods are explained in Section 4. Finally, the last section includes conclusion and suggestions for further research.

# 2 A review on semi-supervised clustering algorithms

The clustering algorithms studied in this section are semi-supervised and based on k-means algorithm [10]. These algorithms are organized into two major categories [11]. The first category includes the methods which apply a small number of labeled data for clustering. The other category includes the algorithms which cluster the data according to the stated constraints. Finally, active learning algorithm is studied.

## 2.1 Algorithms based on labeled data

In this category of algorithms, labeled data are used to initialize the clusters and the restrictions from the data are applied for guiding the process of clustering.

### 2.1.1 Seeded k-Means algorithm

In seeded k-means algorithm [12], data labeled by the user are used to provide the first number of cluster centers. The cluster center $i$ is the average of the points which possess label $i$. Therefore, in this method, the seed clustering is only used for initialization and the seeds are not used in the following steps of the algorithm.

### 2.1.2 Constrained k- Means algorithm

In the algorithm presented in [12] called constrained k-means algorithm, the data labeled by the user are applied to initialize the k-means algorithm. Next, during the k-means procedure and in the process of assigning data to the clusters, the cluster labels of seeded
Another algorithm is COP k-Means algorithm [21] which is the same as k-Means algorithm for which Must-Link and Cannot-Link constraints are implemented on its data points. These two types of constraints for sample pairs are defined as follows:

- Must-Link between two data points: this constraint indicates that two samples must be placed in a single cluster.
- Cannot-Link between two data points: this constraint shows that two samples cannot exist in one single cluster.

In this method, initial centers are obtained through the use of the stated constraints. Later, any data segment related to those centers for which must-link constraint has been implemented cannot be selected as a center for another cluster. During the stages of assigning data to the cluster, each point is assigned to the nearest cluster which does not violate its constraints. If it is not assigned in that way, the algorithm will terminate.

2.2.3 Pairwise Constrained k-Means algorithm

Like previous methods, Pairwise Constrained k-Means algorithm [12] has two types of constraints and, it takes data set \( X \), Must-Link constraints sets between two data points \( (M) \), Cannot-Link constraint sets between two data points \( (C) \), weight of data \( (W) \) and the number of clusters \( (k) \) as its input data and returns the independent cluster \( k \).

PCk-Means algorithm actually consists of three main steps: initialization step, cluster assignment step and centroid estimation step. In the first phase of implementing PCk-Means algorithm, it is assumed that the constraints are consistent. Suppose that the number of connected components in set \( M \) equals \( \lambda \), and which is used for creating \( \lambda \) neighboring sets \( \{Np\}_{p = 1}^{\lambda} \). Now for each neighbor pairs \( Np \) and \( Np' \) which have at least one Cannot-Link constraint between two data segments among them, we insert Cannot-Link constraint between all the pairs of these two \( Np \) and \( Np' \) sets. In this method, the neighboring \( Np \) sets have been obtained through Must-Link constraint and are constant during different implementations of the algorithm; They are independent from the created cluster sets and will be updated at each iteration of the algorithm. After this initialization step the neighbor set \( \lambda \{Np\}_{p = 1}^{\lambda} \) is used as the cluster center. If \( \lambda >= K \), then choose the biggest \( K \) neighbor sets which their centers are considered as the center of the clusters. If \( \lambda <= K \), \( \lambda \) number of cluster centers are chosen from the set of neighbors and the rest of the cluster centers are chosen from other data by using Cannot-Link constraint. In the second step of this algorithm, each point \( x \) is set in a cluster such that the distance from this point to the cluster center and the number of violated constraints by the cluster is minimum. The centroid estimation phase acts in a
similar way to k-Means algorithm to update the center of the clusters. Finally, the second and third steps are repeated until the algorithm converges. (Proof of convergence of the algorithm is shown in [11]).

2.3 Active learning

In [22], a method was proposed which is carried out in two steps. In the first step, this algorithm looks for \( k \) data which are located next to each other and each one is in a separate cluster. For this purpose, the farthest-first method is used. At the end of this phase, there must remain at least one data segment from each cluster and these data will be labeled by the user. Later in step two, for each data not existing in the obtained set of neighboring points in phase one, at most \( k - 1 \) search is done to find the cluster label of that data. It is achieved based on the closeness of each data to the members of different clusters specified in phase one.

Another algorithm for active learning methods which works according to graphs was given in [23]. This algorithm is a combination of constrained spectral clustering and k-means clustering algorithms. It is placed in a category of spectral clustering algorithm, called spectral graph transducer. In order to obtain more information in this method, unlabeled data and data obtained from testing a semi-supervised method are used. Using spectral clustering algorithm can be due to making no assumption about the shape and size proportion of the clusters. According to the number of labels, the algorithm divides the learning points into the following four cases:

- If the labels of the points are converging,
  - Run the approximate spectral clustering algorithm based on k-Means on the data set.

- If \( L < \theta \)
  - Run k-Means algorithm on the unlabeled data and data obtained from the test.
  - With the help of labeled data and the center of clusters set up a new data set.
  - Run STG on the new data set.

- If \( L \geq \theta \)
  - Run k-Means algorithm on the unlabeled data and test data.
  - Run k-Means algorithm on the labeled data separately on each of the two classes.
  - With the help of the obtained clusters centers, create a new data set.
  - Perform STG on the new data set.

- If \( L = u \)
  - Run the set with Linear SVM.

In this algorithm: \( L \) is the labeled data, \( u \) is the unlabeled data, \( \theta \) is the threshold limit on the labeled points and SGT is the spectral graph transducer.

3 The proposed algorithm

In this paper, constrained k-means semi-supervised clustering method is introduced. The way of selecting the initial seeds has an important impact on the effectiveness of the semi-supervised clustering methods. In constrained k-means clustering method, the user (i.e. an expert) selects the first seeds. The expert user selects some data for the initial seed and is sure about its placement in a particular cluster while the clustering method most probably will also place the data in the correct cluster. The problem with clustering methods is in setting the boundary data existing between the clusters, in the right cluster; thus, the user is asked to select the initial seeds on the boundary data. The proposed algorithm is shown in algorithm 1.

**Algorithm 1 Constrained k-means with user feedback**

**Input:** set of data points \( X = \{x_1, ..., x_n\} \), number of clusters \( K \)

**Output:** disjoint \( K \) partitions \( \{X_l\} \) of \( X \)

1: initialize: do the K-Means algorithm on \( X = \{x_1, ..., x_n\} \) data points and make the initial clusters \( H = \{h_1, ..., h_k\} \)

2: boundary data: repeat until convergence

3: for each \( h_i \in H \) do

4: \( x_i \in h_i, w = \text{distance of } x_i \text{ from } h_i, y = \text{ distance of } x_i \text{ from } h_j, j \neq i, z_i = w - y \)

5: calculate boundary data: set of boundary data points \( B = \min z_i \) in each cluster \( h_i \)

6: boundary data selection:

7: for each \( h_i \in H \) do

8: select \( \frac{1}{2} \) boundary data from \( B \)

9: end for

10: user feedback: call the user to label the selected boundary data

11: do the Constrained K-Means algorithm

12: end for

Clustering data in the proposed method was done in three steps. In the first step, i.e. preprocessing step, the aim was to make a set of clusters as initial clusters which were used in the next steps. The algorithm, used in this step to cluster the data, was k-means algorithm and the number of the clusters was its input.

In the second step, the aim was to recognize the boundary data and to seed from the initialized clusters. Suppose that there are some data which are near their cluster centers and also their distance to the center of other clusters is short; in other words, there is small difference between the distance to their cluster center and the center of other clusters. Such data are the boundary ones. To find the boundary data, the distance of data \( s \) in a cluster to its cluster center was calculated as \( x \). Also, the median of distances
of $s$ to all other cluster centers was calculated as $y$. Then, the difference between $x$ and $y$ was determined as $z$ for $s$. This process was done for all data in the present dataset. Finally, in each cluster, the data with minimum $z$ were the boundary data. After recognizing the boundary data, an important point in selecting the boundary data was that the algorithm may not select boundary data from all the clusters similarly or there might be a cluster, from which no boundary data could be selected. This problem resulted in lacking adequate information about all the clusters and, consequently, decreased accuracy of final clustering. To address this problem, the number of boundary data selected from each cluster was calculated in proportion of the total data. The number of data selected from cluster $j$ followed the next equation.

$$\frac{P.N_j}{N_j}$$

In equation 1, $N$ is the total number of data, $N_j$ is the number of data in cluster $j$, and $P$ is the number of data the user can label. Thus, boundary data were selected from all clusters uniformly. Then, an expert user was asked to determine the labels of the selected boundary data. In the last step of this approach, constrained k-means algorithm was used to do the final clustering process over the remaining data.

4 Results and Evaluation

4.1 Data set

To evaluate the effectiveness of clustering algorithms, Standard Data Set UCI\(^1\) was used. The data set profile is displayed in Table 1.

4.2 Evaluation criteria

To compare and evaluate the clustering methods in this research, Accuracy, ARI and NMI were used which are explained in this section.

4.2.1 Accuracy criterion

The accuracy criterion \([22]\) identifies the data rate which has been correctly clustered. For this purpose, the two criteria of sensitivity and specificity are used. Sensitivity indicates the ratio of true positive values (positive values which are truly identified) to all positive values and specificity shows the ratio of true negative values (negative values which are correctly recognized) to all negative values.

$$sensitivity = \frac{t_{pos}}{pos} \quad (2)$$

$$specificity = \frac{t_{neg}}{neg} \quad (3)$$

In Equation 2, $t_{pos}$ is the number of true positive values and $pos$ shows the number of positive values.

In Equation 3, $t_{neg}$ is the number of true negative values. According to equation 4 it can be said that accuracy is a function of sensitivity and specificity.

$$Accuracy = (sensitivity * \frac{pos}{pos + neg}) + (specificity * \frac{neg}{pos + neg}) \quad (4)$$

4.2.2 ARI (Adjusted Rand Index) criterion

To compare the results obtained from clustering base on an external criterion, ARI criterion \([22, 24]\) is used. If $U$ is the external criterion, $V$ the results of clustering, $n_{ij}$ the number of existing objects in both $u_i$ group and $v_j$ cluster, and $n_i$ and $n_j$ the number if objects in $u_i$ group and $v_j$ cluster, then the ARI criterion is calculated according to Equation 5.

$$\frac{\sum \binom{n_i}{2} - [\sum \binom{n_i}{2} \cdot \sum \binom{n_j}{2}] / \binom{N}{2}}{\frac{1}{2} [\sum \binom{n_i}{2} + \sum \binom{n_j}{2} ] - [\sum \binom{n_i}{2} \cdot \sum \binom{n_j}{2} ] / \binom{N}{2}} \quad (5)$$

4.2.3 NMI

This criterion specifies the amount of statistical information generated by random variables that represent cluster initialization and group labeling done by the user. NMI measure calculates the extent to which the algorithm correctly clusters the labeled data.

$$NMI = \frac{I(C; K)}{(H(C) + H(K))/2} \quad (6)$$

In Equation 6, $C$ is the random variable which represents initializations of the points inside the cluster, and $K$ is the random variable representing the label of the groups.

$I(X; Y) = H(X) - H(X|Y)$ represents the mutual information between variables $X$ and $Y$. $H(X)$ is the Shannon expansion of variable $X$ and $H(X|Y)$ is the Shannon expansion of $X$ if $Y$.

To evaluate the proposed method and compare it with the two methods of constrained k-means and k-means, it was applied to the described data sets and the results relating to CCI, NMI, AND ARI have been

\(^1\) http://archive.ics.uci.edu/ml/datasets.html
displayed in Figures 1, 2, 3.

As shown in Figure 1, based on CCI criterion, the proposed method has a better performance in all data sets except in glass. The maximum improvement in relation to constrained k-means was 3.53% in ionosphere data set, and the maximum improvement relating to k-means was 38.70% in waveform data set.

Based on NMI standard, the highest rates of improvement in relation to constrained k-means and k-means were 11.22% and 42.5% respectively in Iris data set. Also, improvement of the proposed method based on ARI criterion relating to constrained k-means and k-means were calculated 5.45% and 40.15% respectively in Iris data set. Figures 2 and 3 illustrate the results related to NMI and ARI criteria.

Figures 4, 5 and 6 present the results of comparing the accuracy of proposed approach with semi-supervised clustering methods. Compared to other methods the proposed method achieved better results on most of the data sets and has the best performance of all the methods in CCI criterion on vowel, ionosphere, waveform and wine data sets. For NMI, it achieves the best results in vowel, ionosphere and iris.
Accordingly, effect of selecting initial seeds can be observed. Seeded K-Means and Constrained K-Means selects the initial seeds randomly, but this method selects the initial seeds by user feedback and ended in better results. This method has better results than Seeded K-Means and MPC-K-Means in all datasets. As can be seen in the results shown in Figures 1 - 6, the proposed method has better results than Seeded K-Means and Constrained K-Means on all datasets, because Seeded K-Means and Constrained K-Means selects the initial seeds randomly, but this method selects the initial seeds by user feedback and ended in better results. Accordingly, effect of selecting initial seeds can be observed.
Figure 6. Comparison of the results obtained from implementation of semi-supervised methods and the proposed method based on ARI criterion

5 Conclusion

In this paper, various data clustering methods were studied and attempts were made to present a method for applying the user feedback to improve clustering results. The focus of this study was on semi-supervised learning with user feedback which had a close relationship with active learning method. The goal of active learning is to specify data samples and label them by an expert user to obtain more information about data clusters and models; however, in semi-supervised method using the user feedback, the user is allowed to provide the system with only the information and constraints about the data without knowing about the labels of the data. The aim was to let the user guide the clustering process in a way to obtain the most appropriate clustering model for the existing data. As a result, the boundary data were identified first and then the user was asked to label them; after labeling them by the user, a semi-supervised clustering method was applied for the final clustering. In fact, the main goal of this study was not to present a new approach of semi-supervised clustering, rather the innovation is in providing a method to maintain an interaction between the user and clustering system for achieving better results. The results indicated that implementing user feedback led to improvement in CCI, NMI, and ARI criteria on different standard datasets. Thus, the study paved the way for further studies on applying user feedback in clustering methods. Due to the similarity between semi-supervised methods with user feedback and active learning methods, it can be said that active learning methods can be used in the same way as well. In addition, more accurate ways can be used for selecting the boundary data.

References


[23] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining: Concepts and techniques, (the morgan kaufmann series in data management systems). 2006.


Kavan Fatehi

Received MS degree in artificial intelligence from University of Yazd. He is currently member of Information Processing and Knowledge Discovery Lab in University of Yazd. He is interested in data mining, machine learning and information retrieval.

Arastoo Bozorgi

Received the BS degree in computer engineering from Shahid Bahonar University of Kerman in 2010, and the MS degree in software engineering from the Shahid Beheshti University in 2013, Tehran. His current research interests include social network analysis, graph mining and information retrieval.

Mohammad Sadegh Zahedi

Received the BS degree in Software Engineering (2010) from the University of Zanjan and MS degree in Software Engineering (2013) from University of Tehran, Department of Electrical and Computer Engineering. He is currently member of Database Research Group (DBRG) in University of Tehran. His research interests include information retrieval, social network analysis and data mining.

Ehsan Asgarian

Received the BS degree (first class Honors) in computer engineering from Azad University of Mashhad in 2006, and the MS degree in software engineering from the Sharif University of Technology(SUT) in 2008, Tehran. Currently, he is working toward the PhD degree in computer science, College of Engineering at the Ferdowsi University of Mashhad(FUM) under supervision of Prof. Kahani. His early research focused on bioinformatics, and haplotype data analysis. His current research interests include integration of machine learning and natural language processing for sentiment analysis and information extraction from the web.