Improving semi-supervised constraint k-means clustering method by using user feedback

Kavan Fatehi[1] , Arastoo Bozorgi[2] , Mohammad Sadegh Zahedi[3] and Ehsan Asgarian[4]

[1]University of Yazd
kavan.fatehi@stu.yazd.ac.ir

[2]University of Shahid Beheshti
a.bozorgi@Mail.sbu.ac.ir

[3]University of Tehran
s.zahedi@ece.ut.ac.ir

[4]Quchan Institute of Engineering and Technology
asgarian@alum.sharif.edu

Abstract: Recently, semi-supervised clustering methods have been taken into consideration by many researchers. In this type of clustering, there are some constraints and information about a small portion of the data. In constraint k-means method, selecting the initial seed is on the user, i.e. an expert. In this paper, this clustering method has been developed based on user feedback. With the help of the user some initial seeds of boundary data obtained from a first clustering are selected and then the results of the user feedback are given to the constraint k-means algorithm in order to obtain the most appropriate clustering model for the existent data. The presented method has been applied on various standard data sets; the results show that this method clusters the data with more accuracy compared to other similar methods.

Keywords: clustering, semi-supervised using user feedback, active learning, boundary data

1 Introduction

Data clustering, known as one of the popular pattern recognition technique, 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]. 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, thus the information about the target variable is pre-defined. In these methods, a collection of labeled data are specified as training data based on which a model is created to predict the label of other sets of data; in other words, the main goal of supervised methods is to insure that the created model is adequately trained in the classifier section in order to place new samples in

appropriate categories. In unsupervised method, the aim is to cluster a set of unlabeled data, i.e. we divide the data without having any knowledge about them into a set of clusters 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. Therefore, no target variable has been pre-defined earlier and the correlation and structure in the existing data are studied.

As the name suggests, semi-supervised learning is somewhere between unsupervised and supervised learning [2]. In these models, 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 objects [3]. In other words, semi-supervised learning is applied in both classification and clustering methods. In semi-supervised classification, the training data include both labeled and unlabeled data which the numbers of unlabeled data are much more than the labeled ones [3]. 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, that two instances x_i, x_j must be in the same cluster; and cannot-link constraints, that x_i, x_j cannot be in the same cluster [2].

In another categorization, the learning methods are divided into active and passive groups. In active learning methods which have a lot of semantic similarity to supervised and semi-supervised learning, the goal in active learning is to design and analyze learning algorithms that can effectively choose the samples for which they ask the teacher for a label [4]. In active learning methods, training data are consisted of unlabeled data. The goal of the active learner is the same as that of a passive learner [5]. In this method, the learner model can ask for the labels of the training data; actually by selecting data for labeling, chooses more accurate data for training. The key idea behind active learning is that 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 [7]. In active learning, the learning system attempts to select which data points, if labeled, would be most informative. In semi-supervised clustering, the human selected the data points, and puts on them a wide array of possible constraints instead of labels [7]. 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 existent semi-supervised learning methods, the selection of initial seeds for the clustering algorithm is done randomly by the user, and this may cause the data to be selected not from all the clusters, as a result, reaching 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 constraint k-means algorithm which has resulted in the increased accuracy of the algorithm; in this case, an initial clustering is done on the data by applying k-means method and then using the proposed method, boundary data are appropriately identified from each cluster, and selection of the first seeds is done by the user on the boundary data. The results of the user feedback are submitted to constraint k-means algorithm and the clustering steps are 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 whose base on k-Means is introduced in the second section of the paper, and in the third part the proposed clustering method is presented.

Evaluation criteria for comparing different algorithms, the profile of the data sets used for comparing the quality of the proposed clustering methods and the results of evaluating the implemented methods are explained in section four of the paper. And finally, the last part includes conclusion and suggestion for further research.

2 A review on clustering algorithms

The clustering algorithms studied in this section are semi-supervised based on k-Means algorithm [8]. These algorithms are organized into two major categories. The first category is 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 for the initial amount of the clusters, and the restrictions from the data are applied for guiding the process of clustering.

2.1.1 Seeded k-Means algorithm [9]

In this algorithm, data labeled by the user are used to provide the first amount of cluster centers. The cluster center i is the average amount 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 [9]

In this algorithm, the data labeled by the user are applied for providing the initial amount of k-Means algorithm. Next, during the k-Means procedure, the cluster label of seeded data points in the process of assigning data to the clusters, remain unchanged and only the label of unseeded data points are allowed to change.
Consequently, like the previous method, constrained k-Means algorithm is provided with the initial amount by the user's labeled data; these labels remain unchanged in the algorithm process while they may face change in the Seeded k-Means method. Constrained k-Means method is appropriate when the initial seed labeling is noise-free or even when the data labels remain unchanged; on the other hand, Seeded k-Means are good when the initial labeled data are noisy.

2.2 Algorithms based on constraints about data

In this class of algorithms, there are some constraints among the existing data in the data sets, such as Must Link or Cannot Link restrictions between two pieces of data; these constraints are applied in the clustering process.

2.2.1 Metric Pairwise Constrained k-Means algorithm [10]

This algorithm is implemented in two stages:Initialization of clusters using determined restrictions by the user.

- Satisfying all the constraints at each iteration of the algorithm.

In other words, this algorithm applies dual constraints for creating initial clusters and guiding the algorithm throughout its iterations. During the iterations, the algorithm is alternating between cluster initialization in E-Step, set center estimate and module learning in M-Step. In E-Step, each x data is placed in a cluster such that the total amount of x distance to the cluster center is minimum, and maximum constraints are met. In M-Step, sets center is double estimated using data from the cluster data set.

2.2.2 COP k-Means algorithm [11]

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

In this method, initial centers are obtained through the use of 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 [9]

Like previous methods, this method has two types of constraints and, it takes data set X, Must-Link constraints sets between two data segments (M), Cannot-Link constraint sets between two data segments (C), weight of data (W) and the number of clusters, i.e. *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 λ which is used for creating λ neighboring sets {Np} p=1. λ. 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 achieved through Must-Link constraint and is constant during different implementations of the algorithm; it is independent from the created cluster sets and will be updated in each iteration of the algorithm. After this initialization step, then neighbor set λ, λ {Np} p = 1 is used as the cluster center. If $\lambda \geq K$, then neighbor K has been selected larger than the set of neighbors and the center of these neighbors is considered as the center of the clusters. If $\lambda \leq K$, in this case first the number of λ cluster center is calculated from the set of neighbors

and then the rest of the clusters are calculated from other data by using Cannot-Link constraint. In the second step of this algorithm, each x point is set in a cluster such that the total 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 and updates the center of the clusters. And finally, the second and third steps are repeated until the algorithm is converging. (Proof of convergence of the algorithm is shown in [9]).

2.3 Active learning

One of the active learning methods [12] is carried out in two phases. In the first phase, this algorithm is looking for k data which are located next to each other and each is in a seperate 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 phase two of the algorithm, for each data not existing in the obtained set of neighboring points in phase one, at most k-1 search is done to realize the cluster label of that data. It is achieved based on the distance closeness from each data to different clusters data specified in phase one.
Another algorithm for active learning methods which works according to graphs is given in [13]. This algorithm is a combination of constrained spectral clustering and k-Means clustering algorithm. It is placed in a category of spectral clustering algorithm, called spectral graph transducer. Inorder to acheive more information in this method, unlabeled data and data obtained from testing a semi-supervised method are used. The reason for 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 learining 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 based
- 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 seperately 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,

θ is the thereshold limit on the labeled points and,
SGT is the spectral graph transducer.

3 The proposed algorithm

In this paper constraint k-means semi-supervised clustering method has been developed. The way of selecting the initial seeds has an important impact on the effectivity of the semi-supervised clustering methods. In constraint k-means clustering method, the selection of the first seeds is on the user, i.e. an expert; The expert user selects some data for the initial seed and is sure about its placement in a particlular cluster while the clustering method most probably will also place the data in the correct cluster. The problem with clustering method is in setting the boundary data existing between the clusters in the right cluster, thus, it is asked from the user to selecet the initial seeds on the boundary data. Boundary data are data whose distance to some neighboring cluster centers is very minimal.

Algorithm:Our approach
Input: Set of data points $X = \{x_1, \ldots, x_N\}$, Number of clusters K
Output: Disjoint K partitioning $(X_l)_{l=1}^{K}$ of X such that
---
Method:
1. initialize: do the k-Means algorithm on $X = \{x_1, \ldots, x_N\}$ data points and make the initial clusters $H = \{h_1, \ldots, hk\}$
2. bounded data: Repeat until convergence
2a. For $h_i \in H$
2b. $x_i \in h_i$, w = distance of x_i from h_i, y = distance of x_i from h_j, $j \neq i$, $z_i = w - y$
2c. calculate bounded data: set of bounded data points B = minimum z_i in each cluster h_i
3. bounded data selection: For $h_i \in H$, select $\frac{p \times N_i}{N}$ bounded data from B
4. user feedback: call the user to label the selected bounded data
5. do the Constrained K-Means algorithm

Fig. 1. The proposed method

In the proposed method which is shown in figure 1, at first k-Means clustering algorithm is implemented on the data set; the purpose for initial clustering is to recognize the boundary data from the rest of the data. Therefore, the distance between each data and its cluster center is calculated, and then the result is compared with the average distance between that data and the near cluster centers. The small difference between the two values of each data shows the boundary rate of that data. These operations are done for all data within each cluster separately; the difference values within each cluster are compared with all the data in that cluster and the data known to have the least difference are considered boundary data.
After recognizing the boundary data, the user is asked to select the initial seeds among them. However, the imortant point in selecting among the boundary data is that the

algorithm may not select boundary data from all the clusters similarly, or there might be a cluster of which no boundary data has been selected. This problem results in lacking adequate information about all the clusters and consequently, decrease in the accuracy of final clustering. To address this problem, the number of boundary data selected from each cluster is calculated in proportion of the total data. The number of data selection from cluster j follows the next equation.

$$\text{The proportion of cluster j from the boundary data} = \frac{P.N_J}{N} \quad (1)$$

In equation (1):
N is the total number of data,
N_j is the number of data in each cluster j, and
P is the number of data the user can label.
Thus, boundary data is selected from all clusters uniformly. Next, the user's boundary data labeling is applied at the last phase of Constrained k-Means algorithm to determine the label of the rest of the data. The stages involved in the proposed method has been shown in figure 1.

4 Results and Evaluation

4.1 Data set

To evaluate the effectivity of clustering algorithms, Standard Data Set [16] UCI was applied. The data set profile has been displayed in table 1.

Table 1: Different data set profile

Data set		Iris	Glass	wavefo rm	vowel	Ionosph ere	Wine
data set profile	Number of data	150	214	5000	990	351	178
	Number of features (dimensions)	4	9	40	13	34	13
	Number of Categories	3	6	3	11	2	3

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 criteria [12]

This criterion identifies the data rate which have been clustered correctly. For this purpose, the two criteria of *sensitivity* and *specifity* are used. *Sensitivity* indicates the

rate of true positive values (the ratio of positive values which were identified truely) and specificity shows the rate of true negative values (the ratio of negative values which were recognized correctly).

$$\text{sensitivity} = \frac{t\_pos}{pos} \tag{2}$$

$$\text{specificity} = \frac{t\_neg}{neg} \tag{3}$$

In equation no. 2:
t_pos is the number of true positive values and *pos* shows the positive values.
In equation (3) also:
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.*

$$\text{Accuracy} = Sensitvity\ \frac{pos}{pos+neg} + Specificity\ \frac{neg}{pos+neg} \tag{4}$$

4.2.2 ARI (The Adjusted Rand Index) criteria [13]

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

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{n_i}{2}\cdot\sum_j \binom{n_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \frac{\left[\sum_i \binom{n_i}{2}\cdot\sum_j \binom{n_j}{2}\right]}{\binom{n}{2}}} \tag{5}$$

4.2.3 NMI

This criterion specifies the amount to 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.

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

In equation 6, C is the random variable which represents initializations of the points inside the cluster, and K is the rendom 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 constraint k-means and k-means, it was implemented on the described data sets and the results relating to CCI, NMI, AND ARI have been displayed in figures 2,3, and 4.

As it is clear from figure 2, the proposed method in CCI criterion, has had a better performance in all data sets except in glass. The maximum improvement in relation to constraint k-means was 3.53% in ionosphere data set, and the most improvement relating to k-means was 38.70 % in waveform data set.

The highest rates of improvement in NMI standard in relation to constraint k-means and k-means were 11.22 % and 42.5% respectively in Iris data set. Also, the proposed method improvement in ARI criterion relating to constraint k-means and k-means were calculated 5.45 % and 40.15% respectively in Iris data set. Figures 3 and 4 illustrate the results related to NMI and ARI criteria.

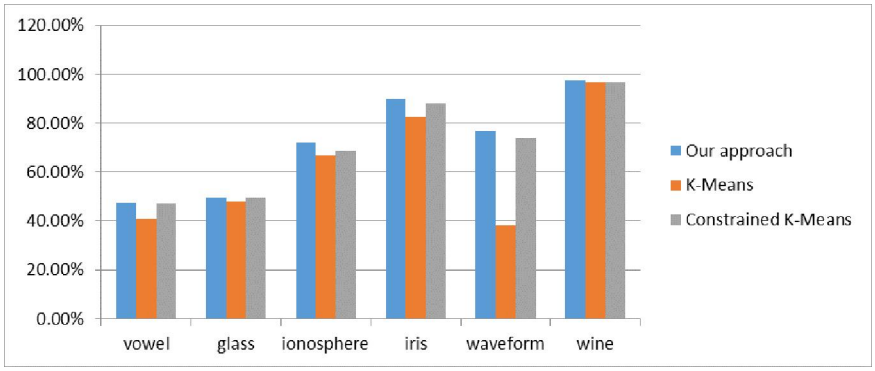

Fig. 2. Comparing the results obtained from implementing constraint k-means and k-means methods by the proposed method based on CCI criterion

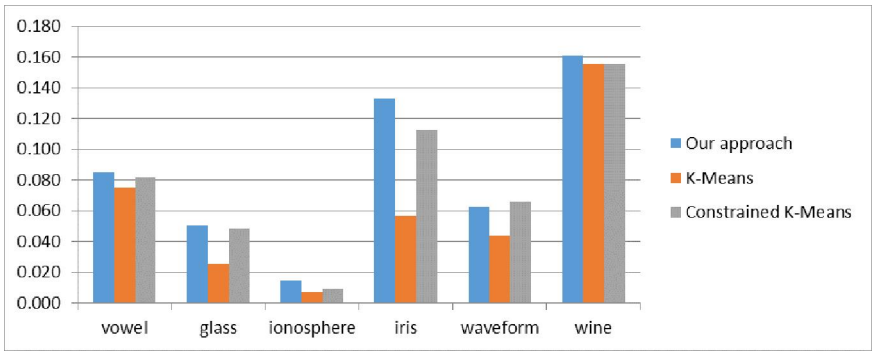

Fig. 3. Comparing the results obtained from implementing constraint k-means and k-means methods by the proposed method based on NMI criterion

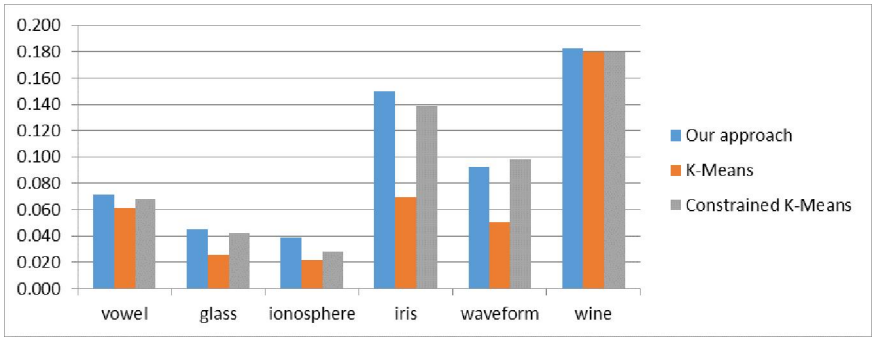

Fig. 4. Comparing the results obtained from implementing constraint k-means and k-means methods by the proposed method based on ARI criterion

The comparison of the accuracy of the proposed method with other semi-supervised clustering methods are represented next in figures 5, 6, and 7. Compared to other methods the proposed method achieved better results on most of the data sets; it had the best performance of all the methods in CCI criterion on vowel, ionosphere, waveform and wine data sets. For NMI, it achieved the best results in vowel, ionosphere and iris and in other data sets there was a very slight difference between the proposed method and the best method. Also, in ARI criterion, it achieved the best results for most of the data sets in comparison with other methods.

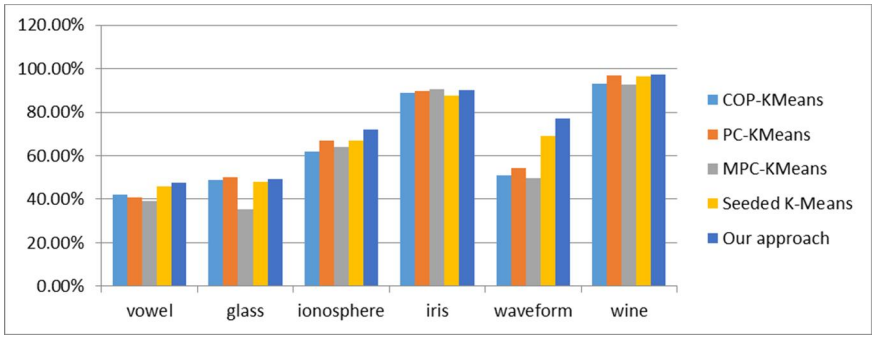

Fig. 5. Comparison of the results obtained from implementation of semi-supervised methods and the proposed method based on CCI criterion

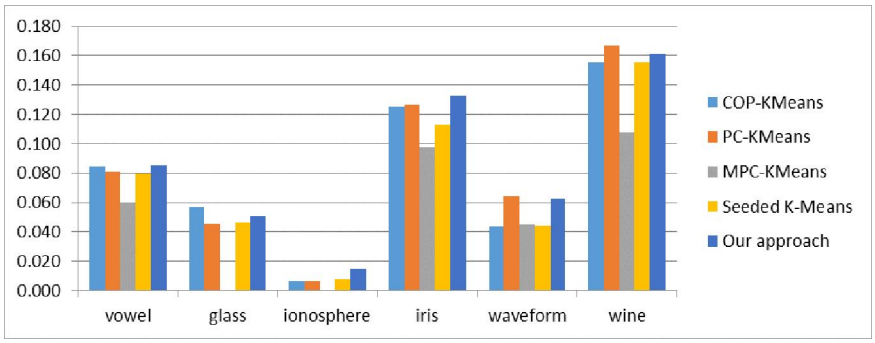

Fig. 6. Comparison of the results obtained from implementation of semi-supervised methods and the proposed method based on NMI criterion

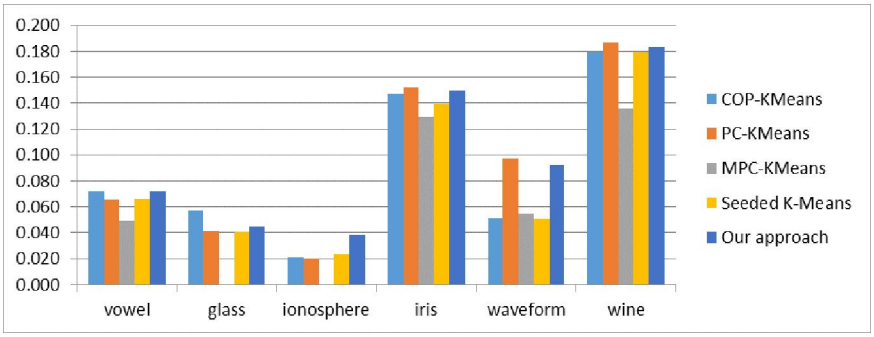

Fig. 7. Comparison of the results obtained from implementation of semi-supervised methods and the proposed method based on ARI criterion

5 Conclusion

In this paper, various data clustering methods were studied, and it was attempted to present a method for applying the user's feedback to improve clustering results. The focus of the study was on semi-supervised learning with user's feedback which has 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 achieve more information about the data clusters and models, however, in semi-supervised method using the user's 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 is 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 are identified first and then the user is asked to label them; after labeling them by the user, a semi-supervised clustering method is applied for the final clustering. In fact, the main goal of the study was not to present a new approach of semi-supervised clustering, rather the innovation in providing a method to maintain an interaction between the user and clustering system for better result achievement was most considered. The results indicated that implementing user's feedback led to improvement in CCI, NMI, and ARI criteria on different standard data sets. Thus, the study has paved the way for further researches on applying user's feedback in clustering methods.

Due to the similarity between semi-supervised methods with user's 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 implemented for selection of boundary data.

References

1. Gu, L., Lu, X.: Semi-supervised subtractive clustering by seeding; In IEEE International Conference on Fuzzy Systems and Knowledge Discovery, pp. 738-741, (2012)

2. Zhu, X., Goldberg, A. B.: Introduction to semi-supervised learning; Synthesis lectures on artificial intelligence and machine learning, 3, 1, pp.1-130, (2009)

3. Leng, M. W., Chen, X. Y., Cheng, J. J., Li, L. J.: Semi-Supervised Clustering Algorithm Based on Small Size of Labeled Data; Applied Mechanics and Materials, 121, pp.4675-4679, (2012)

4. Baram, Y., El-Yaniv, R., Luz, K.: Online choice of active learning algorithms; The Journal of Machine Learning Research, 5, pp.255-291, (2004)

5. Hsu, D. J.: Algorithms for active learning; Doctoral dissertation, The University of California, San Diego (2010)

6. Settles, B.: Active learning literature survey; Technical Report 1648, University of Wisconsin–Madison, USA (2010), Updated on: January 26, 2010

7. Cohn, D., Caruana, R., McCallum, A.: Semi-supervised clustering with user feedback. Constrained Clustering; Advances in Algorithms, Theory, and Applications, 4, 1, pp.17-32 (2003)

8. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, pp.281-297 (1967)

9. Basu, S., Banerjee, A., Mooney, R. J.: Semi-supervised Clustering by Seeding In: Proceedings of the 19th International Conference on Machine Learning, pp. 27-34 (2002)

10. Bilenko, M., Basu, S., Mooney, R. J.: Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the 21th international conference on Machine learning, (2004)

11. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: Proceedings of the 18th international Conference on Machine Learning, pp. 577-584, (2001)

12. Basu, S.: Semi-supervised clustering: Learning with limited user feedback. Doctoral dissertation, The University of Texas at Austin, (2003)

13. Bodó, Z., Minier, Z., Csató, L.: Active learning with clustering. Workshop on Active Learning and Experimental Design, (2011)

14. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques; the Morgan Kaufmann Series in Data Management Systems, USA, (2012)

15. Yeung, K. Y., & Ruzzo, W. L.: Details of the adjusted Rand index and clustering algorithms, supplement to the paper "An empirical study on principal component analysis for clustering gene expression data, Bioinformatics, 17, 9, pp. 763-774, (2001)

16. Machine Learning Repository, Retrieved from http://archive.ics.uci.edu/ml/datasets.html, accessed 5October, (2012)