FAST CONVEGENCE CLUSTERING ENSEMBLE

Javad Azimi
Iran University of Science and Technology
Farjam , Tehran , Iran
Tel: +98-2177341148
Fax: +98-2133868481
Email: Ja_azimi@comp.iust.ac.ir

S.Reza Davoodi
Iran University of Science and Technology
Farjam , Tehran , Iran
Tel: +98-2177341148
Fax: +98-2133868481
Email: SR_davoodi@comp.iust.ac.ir

Morteza Analoui
Iran University of Science and Technology
Farjam , Tehran , Iran
Tel: +98-2177341148
Fax: +98-2133868481
Email: Analoui@iust.ac.ir

ABSTRACT
Clustering ensemble combines some clustering outputs to obtain better results. High robustness, accuracy and stability are the most important characteristics of clustering ensembles. Previous clustering ensembles usually use k-means to generate ensemble members. The main problem of k-means is initial samples which have high effect on final results. Refining initial samples of k-means increases the complexity of algorithm significantly. In this paper we try to predict initial samples, especially for clustering ensemble, without any increasing in time complexity. In this paper we introduce two approaches to select the initial samples of k-means intelligently to generate ensemble members. The proposed methods increase both accuracy and the speed of convergence without any increasing in time complexity. Selecting one sample from each cluster of previous result and selecting k samples which have minimum similarity to each other from co-association matrix are the two proposed method in refining initial samples of k-means. Clarity, simplicity, fast convergence and higher accuracy are the most important parameters of proposed algorithm. Experimental results demonstrate the effect of proposed algorithm in convergence and accuracy of common datasets.

Keywords
Clustering ensembles, K-means, Initial samples, Fast convergence.

1. INTRODUCTION
Clustering is used to divide data into similar groups. The members of each group have maximum similarity with each other and have maximum dissimilarity with the ensemble members of other groups. Because of the characteristic of each data set, there is not a perfect clustering algorithm which performs well in all conditions. Therefore, clustering ensembles is used as a powerful method which can obtain better accuracy that a single clustering algorithm. Clustering ensembles mainly consist of three steps: generating some bootstrap samples or sub samples of input data, applying different clustering algorithms on these bootstrap samples to obtain partitions as the results and using a consensus function to obtain a final partition. Clustering ensembles can offer better solutions in terms of robustness, novelty and stability [1, 2, and 3].

Previous studies mainly focused on 4 approaches to improve the results, using different clustering algorithms to produce partitions for combination [4] , changing initialization or other parameters of a clustering algorithm [3,5], using different features via feature extraction for subsequent clustering [1,6,7] and partitioning different subsets of the original data [8,9,10,11,12,13].

All above introduced mechanisms try to produce higher accuracy and faster convergence from different aspects. But these mechanisms usually can not gather both high accuracy and fast convergence. In this paper we introduce an algorithm which has both high accuracy and fast convergence.

Previous clustering ensemble algorithms usually use k-means as first clustering algorithm which generates the ensemble members. The simplicity and clarity of k-means made it popular in clustering.

The major problem of k-means algorithm is initial samples. It has been reported that the solutions obtained from the k-means are dependent on the initial samples [14, 15]. In the first step of k-means algorithm, we must select k initial samples which k is the number of clusters. There are many methods which select the initial samples intelligently [14, 16, 17]. They study the whole feature spaces to select k initial samples. They usually study the feature space and select the initial samples using probabilistic method. Therefore, they increase the complexity of their algorithms considerably. In this paper we try to predict initial samples, especially for clustering ensemble, without any increasing in time complexity. We try to combine simplicity and accuracy with each other. We think that the simplicity of each algorithm make it popular between algorithms.

Two algorithms which select the initial samples intelligently are introduced in this paper. Clustering ensemble consists of two major steps, generating partitions (ensemble members) and consensus function which obtains the final partition. Previous clustering ensemble algorithms usually use some independent runs of k-means in generating partitions and the initial samples of the k-means runs are selected at random. The result of each k-means has not any effect in others. In previous study the ensemble members obtained completely independent. The propose methods in this paper use the previous result of ensemble members to generate the next ensemble members. Selecting one sample from each cluster of previous result and selecting k samples which have

minimum similarity to each other from co-association matrix are the two proposed method in this paper.

Clarity, simplicity, fast convergence and higher accuracy are the most important parameters of proposed algorithm.

Experimental results demonstrate that clustering ensemble results based on these ensemble members are more accurate and faster than standard k-means with random initial samples in common datasets.

2. Fast Convergence Clustering Ensemble

In k-means algorithm; after each execution of k-means loop, if there is not a sample whose cluster has been changed during the previous loop, we stop k-means algorithm. The number of loops of each execution of k-means algorithm has a high effect on the speed of k-means algorithm. One of the most important parameters which effects on the speed of k-means is initial samples of k-means. The experience demonstrates that the initial samples of k-means have a great effect not only in the number of k-means loops but also in the accuracy of clustering.

It has been reported that the solutions obtained from the k-means are dependent on the initialization of cluster centers [14,15]. In the first step of k-means algorithm, we must select k initial samples which k is the number of clusters. If there are k real clusters, then the chance of selecting one sample from each cluster is small. The chance is relatively small when the number of the clusters is large.

If k clusters have equal samples (n), then the chance of selection of one sample from each cluster is:

$$p = \frac{\alpha}{\beta} = \frac{k!n^k}{(kn)^k} = \frac{k!}{k^k} \quad (1)$$

Where α , is the number of ways to select one samples from each cluster and β is the number of ways to select k samples.

There are many methods which select the initial samples intelligently [14, 16, 17]. They study the whole feature spaces to select k initial samples. In previous studies, they tried to refine initial samples for one execution of k-means algorithm. They should study the feature space and select the initial samples using probabilistic method. Therefore, they increase the complexity of their algorithms considerably.

In proposed algorithm we introduce two methods for selecting the initial samples of k-means intelligently. The proposed methods can only be used in clustering ensembles methods. In clustering ensembles, some independent execution runs of k-means are done, which are named as ensemble members, and the final partitions are obtained from the ensemble by using some deterministic algorithms, such as majority vote or average linkage.

Clustering ensemble has a higher accuracy than single clustering algorithm, such as k-means. The most negative aspect of clustering ensemble is time complexity. Previous studies tried to decrease the time complexity by decreasing the number of ensemble members. The method which has a high accuracy with lower number of ensemble members is the goal of researches.

The two proposed method support both high accuracy and high speed without decreasing the number of ensemble members and without increasing the complexity time in clustering ensembles.

The proposed methods refine the initial samples of k-means algorithm without any increasing in time complexity. Using previous k-means results and using co-association values are the two proposed method.

2.1. Selecting initial samples from previous results

Clustering ensemble consists of two major steps, generating partitions (ensemble members) and consensus function which obtains the final partition. Previous clustering ensemble algorithms usually use some independent runs of k-means in generating partitions. They usually use k-means to generate partitions. The initial samples of the k-means runs are selected at random. We introduce the method which selects the initial samples according to the previous k-means results.

In generating ensemble members, the first ensemble member uses the standard k-means with random initial samples. But other ensemble members use the previous result of k-means algorithm to select the initial seed points. The initial points for execution i is selected from the results of execution i-1 of k-means result. The initial points are selected from different clusters of previous result. After each execution of k-means we select one sample from each cluster at random for next k-means.

Experimental results demonstrated that higher accuracy and higher speed are obtained by proposed algorithm.

Input: *Di (Data points) , K (Number of Cluster) and N(The Number of Partitions)*
Output: *N Ensemble Members*
For i=i , i <N
  if i = 1 **then**
    begin
     seed points = select K initial samples at random;
     do standard k-means;
    end
  else
    begin
      seed points = select K initial points from
      previous k-means results;
      do standard k-means;
    end
end

Figure 1.The proposed algorithm.

2.2. Selecting initial samples from co-association matrix

The last step of clustering ensembles is consensus function. There are many types of consensus function such as *hypergraph partitioning* [1, 6], *voting approach* [8, 18, 19], *quadratic mutual*

information algorithm [20] *and co-association based functions* [2, 21, 22].

*In co-association based functions (*also pair wise approach) the consensus function operates on the co-association matrix. Let *D* be a data set of *N* data points in *d*-dimensional space. The input data can be represented as an $N \times d$ pattern matrix or $N \times N$ dissimilarity matrix, potentially in a non-metric space. Suppose that $X=\{X1\ldots XB\}$ is a set of bootstrap samples or sub samples of input data set *D*. A chosen clustering algorithm is run on each of the samples in *X* that results in *B* partitions $P = \{P1,\ldots, PB\}$. Each component partition in *P* is a set of clusters $P_i = \{C_1^i, C_2^i, \ldots C_{K(i)}^i\}$, $X_i = C_1^i \bigcup C_2^i \ldots . \bigcup C_{k(i)}^i$ and k(i) is the number of clusters in the i-th partition .

$$\text{Co-association (x, y)} = \frac{1}{B}\sum_{i=1}^{B}\varphi(P_i(x), P_i(y))$$

$$\varphi(a,b) = \begin{cases} 1 \text{ if a = b} \\ 0 \text{ if a} \neq \text{b} \end{cases} \qquad (2)$$

Similarity between a pair of objects simply counts the number of clusters shared by these objects in the partitions $\{P1,\ldots, PB\}$. Numerous hierarchical agglomerative algorithms (criteria) can be applied to the co-association matrix to obtain the final partition, including Single Link (SL), Average Link (AL), Complete Link (CL) and voting k-means.

As mentioned above, after each execution of k-means, the co-association matrix is updated. High value in co- association matrix means more similarity and vise versa. Therefore we select k samples which have low similarity with each other. The low similarity or low value in co-association matrix guaranties that the selective samples are not in same clusters. Experimental results demonstrated that higher accuracy and higher speed are obtained by proposed algorithm.

3. Experimental results

The experiments were performed on several data sets, including, three data sets from the UCI repository, "Iris", "Wine", "Soybean". A summary of data set characteristics is shown in Table1.

Table-1 (Datasets which use in our article)

Dataset	Number of samples	Number of cluster	Feature count
iris	150	3	4
Soybean	47	4	35
Wine	178	3	13

In this step, the standard k-means is run for 100 times with random initial samples and intelligent initial samples. The number of loops and the number of misclassified samples of each execution are saved. Table 2 shows the experimental results of the standard k-means on proposed data sets. The total number of k-means loops and the average error rate of k-means for 100 independent executions when initial samples has been selected at random or when initial samples has been selected from previous results (SIPR) are reported in Table 2.

The results in Table 2 are individual accuracy of k-means results in both situations. Table 2 shows that not only the ensemble members in SIPR are more accurate than standard k-means but the ensemble members' loops are less than standard k-means. For example in Iris data set the loop count has been reduced to 37.5 from 46.2.

Table2.the average of miss classification samples and loop count of 100 individual ensemble members

Dataset		Standard k-means	SIPR
Iris	Loop count	46.2	37.5
	Miss classification Samples	34	18
Wine	Loop count	46.1	40.2
	Miss classification Samples	59	50
Soybean	Loop count	24.8	22.6
	Miss classification Samples	17	13

Therefore it can be expected that the clustering ensemble based on these ensemble members is more accurate than previous models.

Table 3 shows both the number of loops and the number of miss classification samples in proposed data set when clustering ensemble is applied to different ensemble members. Three different ensemble members are proposed, standard k-means which generates the ensembles members with random initial samples, the k-means whose initial samples are selected from previous results (SIPR) and the k-means whose initial samples are selected from co-association matrix (SICM). Co-association matrix along with average linkage has been used as consensus function.

Table 3 shows that the clustering ensemble with SICM members is more accurate than others and clustering ensemble with SIPR ensemble members is more accurate than standard k-means. This means that we have obtained more accurate and faster clustering. Since SICM uses all previous results to select the initial samples, the SICM has the better results than previous proposed methods.

Clarity, simplicity, fast convergence and higher accuracy are the most important parameters of proposed algorithm. Time complexity in large data sets with high dimension and real time systems is the vital parameter. Therefore, the proposed algorithm can be applied to these data sets in different situations to improve

the final results significantly. The proposed algorithm can only be used in clustering ensemble algorithms

Table3. The results of proposed method

Dataset		Standard k-means	SIPR	SICM
Iris	Loop count	46.2	37.5	31.3
	Miss classification Samples	33.33	17.66	14.33
Wine	Loop count	46.1	40.2	35.2
	Miss classification Samples	52	49	44
Soybean	Loop count	24.8	22.6	14.3
	Miss classification Samples	15	12.66	11

4. REFERENCES

[1] A. Strehl & J. Ghosh,. Cluster ensembles—a knowledge reuse framework for combining partitionings, in: Proc. of 11-th National Conf. on Artificial Intelligence, Edmonton, Alberta, Canada, 2002, pp. 93–98.

[2] A.L.N. Fred & A.K. Jain : Data Clustering Using Evidence Accumulation, Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR 200, Quebec City ,2002,pp.276 – 280.

[3] A. Topchy & A.K. Jain, & W. Punch.: Combining Multiple Weak Clusterings, Proc. 3d IEEE Intl. Conf. on Data Mining, 2003, pp.331-338.

[4] X. Hu, I. Yoo.: Cluster ensemble and its applications in gene expression analysis, in: Y.-P.P. Chen (Ed.), Proc. 2-nd Asia-Pacific Bioinformatics Conference, Dunedin, New Zealand, 2004, pp. 297–302.

[5] X.Z. Fern & C.E. Brodley.: Random projection for high dimensionaldata clustering: a cluster ensemble approach, in: Proc. 20thInternational Conference on Machine Learning, ICML, Washington,DC, 2003, pp.186–193.

[6] A. Strehl & J. Ghosh.: Cluster ensembles a knowledge reuse framework for combining multiple partitions. Journal on Machine Learning Research, 2002, pp. 583-617.

[7] D. Greene & A. Tsymbal & N. Bolshakova & P. Cunningham.: Ensemble clustering in medical diagnostics, in: R. Long et al. (Eds.), Proc. 17th IEEE Symp, on Computer-Based Medical Systems, 2004, pp. 576– 581.

[8] S. Dudoit & J. Fridlyand.: Bagging to improve the accuracy of a clustering procedure, Bioinformatics 19, 2003, pp.1090–1099.

[9] B. Fischer & J.M. Buhmann.: Bagging for path-based clustering,IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, pp.1411–1415.

[10] A.L.N. Fred & A.K. Jain.: Robust data clustering, in: Proc. IEEEComputer Society Conference on Computer Vision and PatternRecognition, CVPR ,USA, 2003, vol. II, pp. 128–136.

[11] B Minaei & A. Topchy & W. F. Punch.: Ensembles of Partitions via Data Resampling , in Proc. Intl. Conf. on Information Technology, ITCC 04, Las Vegas, 2004.

[12] S. Monti & P. Tamayo & J. Mesirov & T. Golub.: Consensus clustering: a resampling based method for class discovery and visualization of gene expression microarray data, Machine Learning 52, 2003, pp.91–118.

[13] A. Topchy & B. Minaei-Bidgoli & A.K. Jain & W. Punch.: Adaptive Clustering ensembles In Proc. Intl. Conf on Pattern Recognition,ICPR'04, Cambridge, UK, 2004, pp.272-275.

[14] P. Bradley, U. Fayyad, "Refining initial points for k-means clustering", Proceedings 15th International Conf, on Machine Learning, San Francisco, CA, 1998, pp. 91-99.

[15] J. Pena, J. Lozano and P. Larranaga, "An Empirical comparison of four initialization methods for the k-means algorithm", Pattern Recognition Letters Vol. 20, 1999, pp. 1027-1040

[16] G. Babu and M. Murty, "A near optimal initial seed value selection in k-means algorithm using a genetic algorithm", Pattern Recognition Letters Vol. 14, 1993, pp. 763-769.

[17] Y. Linde, A. Buzo and R. Gray," An algorithm for vector quantizer design", IEEE trans. Comm. Vol. 28, 1980, pp. 84-95.

[18] X.Z. Fern, C.E. Brodley, Random projection for high dimensionaldata clustering: a cluster ensemble approach, in: Proc. 20thInternational Conference on Machine Learning, ICML, Washington,DC, 2003, pp. 186–193.

[19] A. Weingessel& E. Dimitriadou & K. Hornik.: An ensemble method for clustering, working paper, 2003.

[20] A. Topchy & A.K. Jain & W. Punch.: A mixture model for clustering ensembles, in: Proceedings of SIAM Conference on Data Miningpp, 2004, pp.379–390.

[21] R.O. Duda & P.E. Hart & D.G. Stork.: Pattern Classification. 2nd Edition, John Wiley & Sons Inc. New York NY, 2001.

[22] A. FredF & Roli & J. Kittler: Finding consistent clusters in data partitions, in: (Eds.), Proc. 2nd International Workshop on Multiple Classifier Systems, MCS_01, Lecture Notes in Computer Science, vol. 2096, Springer-Verlag, 8, Cambridge, UK, 2001, pp. 309–31.