A Novel Semi-Supervised Fuzzy C-Means Clustering Method

Kunlun Li[1], Zheng Cao[1], Liping Cao[2], Rui Zhao[1]

1. College of Electronic and Information Engineering, Hebei University, Baoding 071002,China

E-mail: likunlun @hbu.edu.cn, cao-zheng2008@163.com

2. Department of Electrical and Mechanical Engineering, Baoding Vocational and Technical College, Baoding 071002, China

Abstract: In this paper we propose a novel semi-supervised fuzzy c-means algorithm. We introduce a seed set which contains a small amount of labeled data. First, generating an initial partition in the seed set, we use the center of each partition as the cluster center and optimize the objective function of FCM using EM algorithm. Experiments results show that, our method can avoid the defect of fuzzy c-means that is sensitive to the initial centers partly and give much better partition accuracy.

Key Words: Semi-supervised; Fuzzy c-means; EM.

1 INTRODUCTION

As an important method in data analysis, clustering has been used widely in computer vision, information retrieval, data mining and other fields. Basically, the two main approaches to clustering are hierarchical clustering and partitioning clustering. Most partitioning clustering methods iteratively update the cluster centers, and as such, are often referred as center-based clustering methods. Depending on the way data points are assigned to clusters, partitioning clustering methods are usually classified as: hard and soft. Hard clustering produces a disjoint partition of the data, using a binary strategy so that each data point belongs exactly to one of the partitions. One of the most widely used hard clustering algorithms is the classical k-means. Soft clustering is a relaxation of the binary strategy used in hard clustering, and allows for overlapping of the partitions. Fuzzy c-means is one of the most popular soft clustering methods, which is proposed by Dunn [1], and have been used in many fields such as: medical treatment [2], image segmentation [3], biological information process [4] and so on.

Clustering algorithms try to find the structure information in unlabeled data to construct a classifier, which need no prior knowledge and be seen as one of unsupervised learning usually. Because of no labeled information on data distribution, when the objective function is unsuitable to data set, clustering methods may give useless partition results in practical problems. Semi-supervised clustering using little prior knowledge to improve the performance of clustering algorithms became a novel hotpot in machine learning recent years. Existing semi-supervised clustering methods fall into three general categories that called constrained-based, distanced-based and combination of the former two methods. The supervised information used in semi-supervised clustering can be labels of data points or pairwise constraints that two instances belong to one cluster or not. Bilenko et al considered that the proposed approaches aids unsupervised clustering by incorporating labeled data in the three ways in [18]: First, improved initialization, where initial cluster centroids are estimated from the neighborhoods induced from constraints. Second, Constraint-sensitive assignment of instances to clusters, where points are assigned to clusters so that the overall distortion of the points from the cluster centroids is minimized, while a minimum number of must-link and cannot-link constraints are violated. Third, iterative distance learning, where the distortion measure is re-estimated during clustering to warp the space to respect user-specified constraints as well as to incorporate data variance. Based on k-means, Basu et al proposed a semi-supervised k-means method [5], using a small amount of labeled data. A seed dataset is introduced which contain a little labeled data and assume that all k clusters are covered. In each cluster there is typically at least one seed point. Assigning the instances in seed set to k clusters in order to give an initial partition, optimize the objective function using EM algorithm and gain much better results than classical k-means. Wagstaff et al introduced two types of instances-level constraints named must-link and cannot-link [6] to assistant clustering. Must-link constraints specify that two samples should be assigned into one cluster and cannot-link constraints specify two samples should be assigned into different clusters.

Due to the defects of being sensitive to the initial centers, fuzzy c-means may be trapped into local optimum in practical problems. Semi-supervised clustering strategy may be good choices to solve such kind of problems. Grira et al proposed an active fuzzy constrained clustering method [7], in which pairwise constraints are used. In this paper based on Basu's idea, we propose a semi-supervised fuzzy clustering method: Constrained Fuzzy C-Means algorithm. We utilize the seed dataset, which contains

This work is supported by the National Natural Science Foundation of China under Grant No.60773062, the Science and Technology Supporting Program of Hebei Province under Grant No.072135188 and Scientific research plan projects of Hebei Educational Bureau: 2008312.

978-1-4244-2723-9/09/$25.00 ©

some labeled instances to give an initial partition, and then the classical fuzzy c-means are used to partition the whole dataset.

2 K-Means and Fuzzy C-Means

As classical clustering algorithms, k-means and fuzzy c-means have been widely used in many fields. To the integrality of our paper, firstly, we give a short review of them.

2.1 K-Means Clustering

Let X is the dataset of N samples and D dimensions, $\chi = \{x_i, x_2, ..., x_n\}, x_i \in R^d$. Our goal is to assign the data points into K partitions. Assume that the K centers are $m_1, m_2, ..., m_K$ and in cluster k there is N_k instances. So we get $m_k = \frac{1}{N_k}\sum_{i=1}^{N_K} x_i$, for $k = 1,..., K$ 。Based on squared Euclidean distances and the criteria of within-groups sum of squared error, the objective function of k-means clustering can be presented as:

$$J = \sum_{k=1}^{K}\sum_{i=1}^{N_K}\left\|x_i - m_k\right\|^2 \qquad (1)$$

2.2 Fuzzy C-Means Clustering

As a generalization of classical k-means clustering, the fuzzy c-means was first proposed by Dunn [1] and then expanded by Bezdek [8]. By introducing the membership of each sample belongs to different centers, we compute the membership other than the label that means one instance belongs to a cluster.

Given a dataset $\chi = \{x_i, x_2, ..., x_n\}, x_i \in R^d$ and the number of clusters C. Assume that the centers of all clusters are given as $v_1, v_2, ..., v_C$, $u_{ik}(i = 1,2,...n, k = 1,2,..., C)$ indicates the membership of the data point *i* belong to the cluster k. The objective function of fuzzy c-means can be presented as:

$$\text{Min } J_m(U, v) = \sum_{i=1}^{n}\sum_{k=1}^{C} u_{ik}^m d_{ik}{}^2 \qquad (2)$$

$$\text{s.t. } \sum_{k=1}^{C} u_{ik} = 1, i = 1,..., n \qquad (3)$$

$U = \{u_{ik}\}$ designates the partition matrix, V represents the set of the prototypes associated with clusters. The superscript m is the degree of fuzziness associated with the partition matrix (m>1). If m=1, the soft clustering will degenerate to hard case. Usually, we set m=2. d_{ik} indicates the distance between the sample and the cluster i.

$$d_{ik} = \left\|x_k - v_i\right\|_A^2 = (x_k - v_i)^T A(x_k - v_i) \qquad (4)$$

A is a symmetry positive define matrix, when A=I, d_{ik} will be Euclidean distance. Due to the constraint (3), when the dataset is not ideally, the FCM may give unexpected result. For instance, a data point is far away each cluster center, the membership it belongs to each cluster should be small, but owning to the normalization constraint (3), the FCM will give much bigger membership, which it belongs to each cluster. To improve the robustness of the algorithm, we usually use a relaxation constraint that set the sum of membership is n, such as:

$$\sum_{i=1}^{n}\sum_{k=1}^{C} u_{ik} = n \qquad (5)$$

The FCM can be presented in detail as follows:

1) Given the number of clusters C, the degree of fuzziness m, the threshold to stop the iteration ε, set up the iteration counter t=0.
2) Initialize the cluster centers $V = \{v_1, v_2, ..., v_C\}$.
3) Compute the membership matrix $u_{ik} = 1/\sum_{j=1}^{C}\left(\frac{\left\|x_k - v_i\right\|_A}{\left\|x_k - v_j\right\|_A}\right)^{2/(m-1)}$.
4) Update the cluster center using $v_k = \sum_{i=1}^{n}(u_{ik})^m x_i \Big/ \sum_{i=1}^{n}(u_{ik})^m$, update the u_{ik} to $u_{ik}^{(t)}$, t=t+1.
5) If $\max\left(u_{ik}^{(t+1)} - u_{ik}^{t}\right) \le \varepsilon$, then stop the iteration and output the membership matrix U, or go to 3) and continue.

3 SEMI-SUPERVISED FUZZY C-MEANS CLUSTERING

3.1 Semi-Supervised Clustering Methods

Semi-supervised clustering algorithms have been widely used in many fields. Huang & Pan proposed an expanded method of k-medoids using gene ontology [9]. Lam et al proposed an active learning method used in semi-supervised text categorization [10]. Erman et al used semi-supervised clustering to classify p2p data packets in networks real-time [11].

Several taxonomies of semi-supervised clustering algorithms have been suggested. By focusing on the overall taxonomy we envision three general categories: constrained-based, distanced-based and combination of the two former methods. Constrained-based methods rely on user-provided labels or constraints to guide the algorithm towards a more appropriate data partitioning. This is done by modifying the objective function for evaluating clustering so that is includes satisfying constraints [12], enforcing constraints during the clustering process [13], or initializing and constraining the clustering based on labeled examples [5].

Based on the objective function of classical k-means, Demiriz & Bennett [12] added a measure term of cluster impurity in it:

$$\min \beta * cluster\_dispersion + \alpha * cluster\_impurity \tag{6}$$

Where two cluster dispersion measures were examined: mean square error and Davis-Bouldin index. For the cluster impurity measure, the Gini index is used. Since the objective function is highly discontinuous with many locally minima, genetic algorithm is used.

Wagstaff et al proposed a semi-supervised clustering algorithm [13] in which the pairwise constraints were used as supervised information. Their method was called cop-k-means.

In distance-based approaches, an existing clustering algorithm that used a particular clustering distortion measure is employed; however, it is trained to satisfy the labels or constraints in the supervised data. Several adaptive distance measures have been used for semi-supervised clustering, including string-edit distance trained using EM [14], KL divergence trained using gradient descent [15], Euclidean distance modified by a shortest-path [16], or Mahalanobis distance trained using convex optimization [17].

Basu et al proposed a unified approach to semi-supervised clustering called MPC-k-means [18]. They modified the objective function using pairwise constraints and then proposed a probabilistic framework based on hidden markov random fields that combined the constraint-based and distance-based approaches in a unified model [19].

3.2 Semi-Supervised K-Means

Based on the classical k-means, Basu et al proposed a semi-supervised k-means method [5], using a small amount of labeled data. A seed dataset is introduced which contain a little labeled data and assume that all k clusters are covered. In each cluster there is typically at least one seed point. Assign the instances in seed set to k clusters to give an initial partition, optimize the objective function using EM algorithm and gain much better result than classical k-means.

Using seed set, Basu given two semi-supervised k-means algorithms: Constrained-k-means and Seeded-k-means. In Constrained-k-means, the cluster memberships of the data points in the seed set are not re-computed and thus the cluster labels of the seed data are kept unchanged, and only the labels of the non-seed data are re-estimated. In Seeded-k-means, the user-specified labeling of the seed data may be changed. Take Constrained-k-means as an instance, the algorithm can be presented as follows:

Input: Data set $\chi = \{x_1, ..., x_N\}, x_i \in R^d$, the number of clusters k, initial seed set $S = \cup_{l=1}^{K} S_l$.

Output: K partition $\{\chi_l\}_{l=1}^{K}$ of X such that the objective function is minimized.

Method:1.Initialize $\mu_h^{(0)} \leftarrow \frac{1}{|S_h|}\sum_{x \in S_h} x$ for $h = 1,...,K; t \leftarrow 0$.

2. Repeat until convergence

2a. If $x \in seed\ set$, then assign x to cluster h;

If $x \notin seed\ set$, then assign x to the cluster h^*, for $h^* = \arg\min \left\| x - \mu_h^{(t)} \right\|^2$.

2b. Update the cluster center using $\mu_h^{(t+1)} \leftarrow \frac{1}{\left|\chi_h^{(t+1)}\right|}\sum_{x \in \chi_h^{(t+1)}} x$.

2c. $t = t + 1$.

3.3 Constrained Fuzzy C-Means

As a center-based clustering algorithm, FCM may be effected by the initial center and trapped into locally minimum. In this paper, we modify the classical FCM by using semi-supervised strategy. We use a small amount labeled data as supervised information to improve the initial center artificially. Based on Basu's idea, we introduce a seed set which contain some labeled data. First, generating an initial partition in the seed set, we use the center of each partition as the cluster center and optimize the objective function of FCM using EM algorithm. We assume that the seed set involve in all c clusters and there are typically at least one seed points in each cluster. Our Constrained FCM can be descript in detail as follows:

Input: Data set $\chi = \{x_1, ..., x_N\}, x_i \in R^d$, seed set S, the number of clusters C, the threshold to stop the iteration ε, the degree of fuzziness m, set up the iteration counter t=0.

Output: The final membership matrix U.

Steps of Constrained Fuzzy C-Means:

1. Compute the c centers $V' = \{v'_1, v'_2, ..., v'_C\}$ in seed set S.
2. Initiate a membership matrix U' randomly.
3. Compute the distance of data points in χ to C cluster centers of $V' = \{v'_1, v'_2, ..., v'_C\}$ and gain the initial objective function value and new membership matrix U.
4. Repeat until convergence

4a.Update the cluster centers using $v_k = \sum_{i=1}^{n} (u_{ik})^m x_i \Big/ \sum_{i=1}^{n} (u_{ik})^m$, for k=1…C

4b.Compute the objective function value using $J_t = \sum_{i=1}^{n}\sum_{k=1}^{C} u_{ik}^m \left\| x_i - v_k \right\|^2$.

4c.Update membership using $u_{ik} = 1 / \sum_{j=1}^{C} \left(\frac{\left\| x_k - v_i \right\|_A}{\left\| x_k - v_j \right\|_A} \right)^{2/(m-1)}$, for k=1…C

and i=1…N, t=t+1.

4d.If $\left|J_{t+1}-J_t\right|<\varepsilon$, then stop iteration and output U, or continue.

4 EXPERIMENTS

In experiments, we use two data sets from UCI [20]: Iris and Wine. Iris data set contains 150 vectors of four dimensions and 3 classes of 50 instances each, where each class refers to a type of iris plant. Wine data set contains 178 vectors of 13 dimensions. The number of each class is 59,71 and 48. Set the threshold to stop the iteration ε=1e-6, the degree of fuzziness m=2.

4.1 Experiment Results

To show the effect of the number of labeled data points to the performance of our method, we introduce the labeled rate that is the ratio of the number of labeled data to all data points in dataset. In the experiments, we set the labeled rate to be 10%, 20%, 30%, 40% and 50%, and compare our method to the classical FCM, the results are shown in table 1 and table 2. The seed points are selected randomly from data set and used to compute the initial centers only; the labels are kept unchanged in the iteration. To show the generalization of our methods, the results in the tables are the average of 1000 runs.

To show the advantage of our method, we compare our Constrained FCM (abbreviation for cfcm) and Constrained-k-means (abbreviation for ckm) proposed by Basu[5]. In different labeled rate, we note four indexes: maximum number of correctly labeled (max), minimum number of correctly labeled (min), average number of correctly labeled (average) and average clustering accuracy (accuracy) which is the radio of average number to the number of data set. The results are the average of 1000 runs shown in table 3 and table 4. Particularly, the comparison results on clustering accuracy of two algorithms are shown in figure 1.

Table 1. Compare with FCM on Iris

	FCM	Constrained-FCM				
Labeled rate	-	10%	20%	30%	40%	50%
Average	134.06	136.08	137.97	139.82	141.40	143.05
Accuracy	0.894	0.907	0.920	0.932	0.943	0.954

Table 2. Compare with FCM on Wine

	FCM	Constrained-FCM				
Labeled rate	-	10%	20%	30%	40%	50%
Average	123.29	125.23	133.73	139.84	145.31	150.92
Accuracy	0.793	0.704	0.752	0.786	0.817	0.848

Table 3. Comparisons with Constrained-k-means on Iris

Labeled rate	10%		30%		50%	
	ckm	cfcm	ckm	cfcm	ckm	cfcm
Max	17	16	17	15	14	13
Min	10	9	3	4	1	2
Average	135.27	136.08	139.28	139.82	142.93	143.05
Accuracy	0.902	0.907	0.928	0.932	0.953	0.954

Table 4. Comparisons with Constrained-k-means on Wine

Labeled rate	10%		30%		50%	
	ckm	cfcm	ckm	cfcm	ckm	cfcm
Max	91	89	71	47	41	38
Min	42	40	27	24	17	13
Average	125.22	127.70	139.83	141.44	150.91	152.29
Accuracy	0.703	0.717	0.785	0.794	0.847	0.855

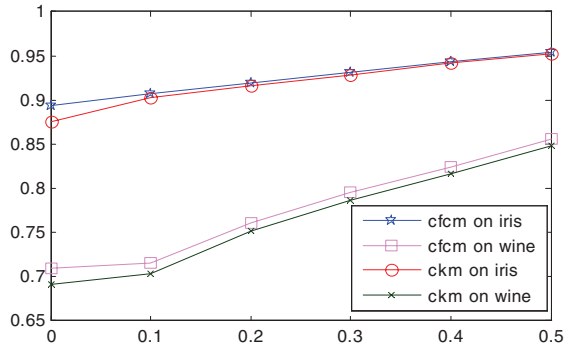

Figure 1. Comparisons of clustering accuracy. The figure shows the performance of the Constrained-k-means and our Constrained FCM on both data sets according to the labeled rate.

4.2 Analysis of Results

From table1 and table 2, we can see that the average clustering accuracy of our Constrained FCM method is better than standard FCM in each labeled rate. The performance is much better when the number of labeled data increase. The results show that using the idea of semi-supervised clustering in fuzzy clustering can exactly improve the classification accuracy. It is because that the initial center in classical FCM is selected randomly, thus the algorithm may be trapped in minimum locally. But in our method, the seed set which contains some labeled data determines the initial center. So we obtain much better results. When the labeled rate increases, the center in seed set is more nearer the center in data set, so the accuracy improves along with the labeled rate.

From table 3 and table 4, we can see that our Constrained FCM gives better performance in both datasets than Constrained-k-means in average accuracy. The results show that in both datasets the fuzzy clustering method is better than hard clustering method and the advantages is kept in semi-supervised clustering. We can gain these results obviously form the curve in figure 1. Note that when there is no data point in seed set, which means labeled rate reduces to 0, the two algorithms become classical k-means and fuzzy c-means.

5 CONCLUSION AND FUTURE WORK

Semi-supervised clustering which use extra information to improve the performance of unsupervised clustering algorithms, gains more attention and become a novel host point in machine learning. Semi-supervised strategy is an effective instrumentality to promote the performance of fuzzy clustering algorithm.

In this paper, we propose a novel semi-supervised fuzzy c-means method—Constrained FCM. We introduce a seed set which contain some labeled data, generate a initial partition in the seed set, use the center of each partition as the cluster center and optimize the objective function of FCM using EM algorithm. Experiments on UCI data sets show that our method can improve the performance of clustering accuracy exactly. When the number of labeled data increase the performance is better. The results proved that using semi-supervised strategy could improve the performance of clustering algorithms effectively. But we do not consider the instance that there is noise in seed set, so we will work on this problem in future.

REFERENCES

[1] Dunn J C. A Fuzzy Relative of the ISODATA Process and its use in detecting compact, well-separated clusters. [J] Cybern, 3 (1): 32-57,1974.

[2] Chen Weijie, MSC MARYELL EN L. A fuzzy C-means (FCM) based approach for computerized segmentation of breast lesions in dynamic contrast enhanced MR images [J]. Academic Radiology, 13 (1): 63-72,2006.

[3] Muneeswaran K, Ganesan, Arumugam S. Texture image segmentation using combined features from spatial and spectral distribution [J]. Pattern Recognition letters, 27: 755-764. 2006

[4] Tari L et al., Fuzzy c-means clustering with prior biological knowledge,[J] Biomed Inform (2008), doi:10.1016/j.jbi. 2008.05.009

[5] Sugato Basu, Arindam Banerjee, Raymond Mooney: Semi-supervised Clustering by Seeding. International Conference on Machine Learning.19-26, 2002

[6] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In: Proc. of the 17th International Conference on Machine Learning, 1103-1110. 2000

[7] Nizar Grira, Michel Crucianu, Nozha Boujemaa. Active semi-supervised fuzzy clustering. Pattern Recognition 41.1834-1844.2008

[8] Bezdek J C.Pattern Recognition with Fuzzy Objective Function Algorithms [M]. New York: Plenum, 1981.

[9] Huang D, Pan W. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics 2006; 22(10): 1259-68.

[10] R. Huang, W. Lam, An active learning framework for semi-supervised document clustering with language modeling, Data & Knowledge. Engineering. doi: 10. 1016/ j.datak. 008.08.008.

[11] Jeffrey Erman, Anirban Mahanti, Martin Arlitt, Ira Cohen, and Carey Williamson. Offline/Realtime traffic classification using semi-supervised learning. Perform. Evaluation, 64(9-12): 1194–1213.2007

[12] A. Demiriz, K. P. Bennett, and M. J. Embrechts. Semi-supervised clustering using genetic algorithms. In Artificial Neural Networks in Engineering (ANNIE-99), 809–814.1999

[13] K. Wagstaff, C Cardie, S Rogers and S. Schroedl, Constrained K-means Clustering with Background Knowledge, In: Proc. of 18th International Conference on Machine Learning, 577-584. 2001

[14] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In: Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2003), 39–48. 2003.

[15] D. Cohn, R. Caruana, and A. McCallum. Semi-supervised clustering with user feedback. Technical Report TR2003-1892, Cornell University, 2003.

[16] D. Klein, S. D. Kamvar, and C. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: Proc. of the 19th International Conference on Machine Learning (ICML-2002), 307–314. 2002

[17] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In Advances in Neural Information Processing Systems 15 Cambridge, MA, 2003. MIT Press. 505–512.

[18] S. Basu, M. Bilenko, and R. J. Mooney. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering. In: Proc. of the ICML-2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, 42–49. 2003

[19] Basu, S, Bilenko, M., Mooney, R. J. A probabilistic framework for semi-supervised clustering. International Conference on Knowledge Discovery and Data Mining, 59-68.2004

[20] UCI repository of machine learning databases http://www.ics. uci.edu/mlearn/MLRepository.html