A New Ensemble Clustering Method Based on Dempster-Shafer Evidence Theory and Gaussian Mixture Modeling

Yi Wu, Xiabi Liu, and Lunhao Guo

Beijing Lab of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, Beijing 100081, China
{wuyi,liuxiabi,guolunhao}@bit.edu.cn

Abstract. This paper proposes a new method based on Dempster-Shafer (DS) evidence theory and Gaussian Mixture Modeling (GMM) technique to combine the cluster results from single clustering methods. We introduce the GMM technique to determine the confidence values for candidate results from each clustering method. Then we employ the DS theory to combine the evidences supplied by different clustering methods, based on which the final result is obtained. We tested the proposed ensemble clustering method on several commonly used datasets. The experimental results confirm that our method is effective and promising.

Keywords: Data Clustering, Ensemble Clustering, Dempster-Shafer (DS) Evidence Theory, Gaussian Mixture Modeling (GMM).

1 Introduction

At present, there is no single clustering method can achieve robust results in all situations. It is promising to integrate various clustering methods for obtaining the better performance. This solution is usually called ensemble clustering, which has attracted more and more attentions in recent years. The existing ensemble clustering methods can be classified into two main categories: voting based and hyper-graph based [1].

- Voting based methods firstly solve the label correspondence problem and then find out the consensus partition. Tumer and Agogino [2] introduced the criterion of Average Normalized Mutual Information (ANMI) to measure the ensemble results. They assigned the data points to different clusters dynamically by a voting procedure for achieving the best ANMI. Dimitriadou et al. [3] presented a voting scheme for integrating fuzzy clustering algorithms. The main steps include creating a mapping between candidate clusterings, calculating the highest percentage of common points, and assigning the points to the common clusters. Wang et al. [4] proposed a soft-voting method to integrate the candidate soft clustering results and achieved acceptable results.
- In hyper-graph based methods, the ensemble clustering problem is transformed into a hyper-graph partitioning problem. The data points are represented as edges in a hyper-graph and the clusters as undirected hyper-edges. Under this idea there

C.K. Loo et al. (Eds.): ICONIP 2014, Part II, LNCS 8835, pp. 1–8, 2014.

are three representative methods which are Cluster-based Similarity Partitioning Algorithm (CSPA) [5], Hyper-Graph Partitioning Algorithm (HGPA) [6] and Meta-Clustering Algorithm (MCLA) [7]. The CSPA creates a binary similarity matrix for each single clustering. The entry-wise average of single similarity matrices yields an overall similarity matrix. They re-cluster the overall matrix and get the ensemble results. In the HPGA, the ensemble problem is formulated as partitioning the hyper-graph by cutting a minimal number of hyper-edges. In the MCLA, the idea is to group and collapse related hyper-edges and assign each object to the collapsed hyper-edge. It provides better performance than HPGA and retains low computational complexity.

In this paper, we propose a new ensemble clustering approach based on Dempster-Shafer (DS) evidence theory and Gaussian Mixture Modeling (GMM) technique. Each group of candidate clustering results can be regarded as an evidence to determine the final clustering results. Thus the ensemble clustering problem can be solved by using DS theory to combine the evidences from involved clustering methods. Based on this core idea, we introduce the GMM technique to calculate the confidence of assigning a data point to a candidate cluster for each clustering method. Then the orthogonal sum of confidences from different clustering methods is computed and used to decide the final result under the DS theory. We evaluate the effectiveness of our proposed approach by conducting the experiments on commonly used data sets.

2 Single Clustering Methods

In this paper we use the proposed ensemble method to combine single clustering methods based on dense Gaussian distributions. In this type of single clustering methods, we use the Expectation-Maximization (EM) algorithm [8] to fit a GMM with a large number of Gaussians to the data set. The data subset corresponding to each generated Gaussian component is taken as a minimum unit of data. Then, the classical clustering methods can be performed on these units to complete the clustering. This means that all the operations will be processed on Gaussian distributions, instead of on data points. Furthermore, each cluster can be seen as a GMM composed by dense Gaussians.

For completing such clustering, we need a measure of similarity between Gaussians. In this paper, the Gaussian Quadratic Form Distance (GQFD) [9] is used, which is defined as

$$GQFD_{f_s}(g_a, g_b) = \sqrt{\left(\boldsymbol{\omega}^a \middle| -\boldsymbol{\omega}^b\right) \boldsymbol{A} \left(\boldsymbol{\omega}^a \middle| -\boldsymbol{\omega}^b\right)^T}\ , \tag{1}$$

where g_a and g_b be two Gaussian distributions, $\left(\boldsymbol{\omega}^a \middle| -\boldsymbol{\omega}^b\right)$ denotes the concatenation of weights from g_b and g_b, $\boldsymbol{A}$ is a matrix, each entry in which is the measure of similarity between two Gaussians. The GQFD has proved its effectiveness for modeling content-based similarity, for more details of which the reader is referred to Beecks et al. [9].

We summarize the algorithm framework of dense Gaussian distributions based single clustering method in Algorithm 1.

Algorithm 1. Dense Gaussian distributions based single clustering algorithm

Input: Data set
Output: Clustering Results
Steps:
Step1. Fit a GMM with a large number of components to the whole data set by using the EM algorithm.
Step2. Use a classical clustering method (such as *k*-means) configured with GQFD to cluster the generated Gaussians.
Step3. The data points are grouped according to the clusters of Gaussians.

3 Ensemble Method

The core task of our ensemble clustering approach is to combine the confidences for candidate results from single clustering methods. We introduce the DS evidence theory to complete this task. Suppose we have N single clustering algorithms, each of them organizes data points into K clusters. Let $\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_N$ be the corresponding clustering results from each single algorithm. Then our DS theory based ensemble clustering approach is explained as follows. For the details of DS evidence theory itself the reader is referred to Shafer [10].

3.1 Evidence from Single Clustering Method

Under the DS evidence theory, we need to compute the evidence corresponding with candidate clustering results from each method for further combining them. For each $\boldsymbol{R}_i \big|_{i=1}^{N}$, this evidence can be represented by the probabilities corresponding to each element in its power set. We calculate these probabilities based on the GMM.

Let $\boldsymbol{\Omega}_i = \left\{ \boldsymbol{\Omega}_i^1, \boldsymbol{\Omega}_i^2, \cdots, \boldsymbol{\Omega}_i^T \right\}$ be the power set of $\boldsymbol{R}_i$, where T be the number of elements in $\boldsymbol{\Omega}_i$. Obviously, $T = 2^K - 1$. The first K elements in $\boldsymbol{\Omega}_i$ are K clusters in $\boldsymbol{R}_i$, which are generated by the single clustering method. The rest elements in $\boldsymbol{R}_i$ are the possible clusters combined by these K elements. For example, if $\boldsymbol{R}_i$ is $\{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}$, where $\boldsymbol{r}_j \big|_{j=1}^{3}$ represent the three clusters from the *i*-th single clustering method, then $\boldsymbol{\Omega}_i = \{\{\boldsymbol{r}_1\}, \{\boldsymbol{r}_2\}, \{\boldsymbol{r}_3\}, \{\boldsymbol{r}_1, \boldsymbol{r}_2\}, \{\boldsymbol{r}_1, \boldsymbol{r}_3\}, \{\boldsymbol{r}_2, \boldsymbol{r}_3\}, \{\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3\}\}$.

For a large K, T will become a huge value and the following computation will become unfeasible if all the combinations in the power set $\boldsymbol{\Omega}_i$ are considered. In such cases, we just consider the clusters combined by at most 2 elements of $\boldsymbol{R}_i$. In other words, the combinations with 3 or more elements are neglected for a large K. Denoeux and Masson [11] have proved that this strategy is suitable for big power sets.

We introduce the GMM to model each element in $\boldsymbol{\Omega}_i$. As explained before, the first K elements in $\boldsymbol{\Omega}_i$ correspond to the GMMs generated by our single clustering method. The rest elements are composed by these generated GMMs. Let $\boldsymbol{x}$ be an arbitrary data point, $p\left(\boldsymbol{x}\middle|\boldsymbol{\Omega}_i^j\right)$ be the resultant GMM for $\boldsymbol{\Omega}_i^j$, then we have

$$p\left(x\middle|\boldsymbol{\Omega}_i^j\right)=\sum_{k=1}^{m}\omega_k\left(2\pi\right)^{-\frac{d}{2}}\left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}_k\right)\right) \tag{2}$$

where m be the number of Gaussian components, ω_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ be the weight, the mean vector, and the covariance matrix of the k-th Gaussian component, respectively.

Based on the GMM, we use Bayesian rule to calculate the posterior probability of assigning $\boldsymbol{x}$ to the cluster $\boldsymbol{\Omega}_i^j$:

$$P\left(\boldsymbol{\Omega}_i^j\middle|\boldsymbol{x}\right)=p\left(\boldsymbol{x}\middle|\boldsymbol{\Omega}_i^j\right)P\left(\boldsymbol{\Omega}_i^j\right)\Big/\sum\nolimits_{k=1}^{T}p\left(\boldsymbol{x}\middle|\boldsymbol{\Omega}_i^k\right)P\left(\boldsymbol{\Omega}_i^k\right), \tag{3}$$

where the prior probability $P\left(\boldsymbol{\Omega}_i^j\right)$ is assumed to be the same for each element.

3.2 Combining Evidences

The posterior probability $P\left(\boldsymbol{\Omega}_i^j\middle|\boldsymbol{x}\right)$ can be seen as the confidence of the i-th clustering algorithm for associating $\boldsymbol{x}$ with the j-th cluster. We combine this kind of confidences from all the involved algorithms by using the orthogonal sum method. Let $\boldsymbol{\Omega}^j$ be the combined result of the j-th cluster, then we have

$$P\left(\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right)=\left(\sum_{\cap\boldsymbol{\Omega}_i^j=\boldsymbol{\Omega}^j}\prod_{i=1}^{N}P\left(\boldsymbol{\Omega}_i^j\middle|\boldsymbol{x}\right)\right)\Bigg/\left(\sum_{\cap\boldsymbol{\Omega}_i^j\neq\phi}\prod_{i=1}^{N}P\left(\boldsymbol{\Omega}_i^j\middle|\boldsymbol{x}\right)\right). \tag{4}$$

As described above, $\boldsymbol{\Omega}^j\Big|_{j=1}^{K}$ is corresponding with K single clusters. To determine the final ensemble clustering results, we calculate the belief and plausibility values for these first K elements by

$$Bel\left(\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right)=\sum_{\boldsymbol{A}\subseteq\boldsymbol{\Omega}^j}P\left(\boldsymbol{A}\middle|\boldsymbol{x}\right),\quad j=1,2,\cdots,K\ , \tag{5}$$

and

$$Pl\left(\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right)=1-Bel\left(\neg\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right),\quad j=1,2,\cdots,K\ , \tag{6}$$

respectively. Based on the belief and plausibility functions, the final confidence of assigning $\boldsymbol{x}$ to the j-th cluster can be computed as the class probability. Let $\left|\boldsymbol{\Omega}^j\right|$

and $\left|\boldsymbol{\Omega}_i\right|$ denote the number of elements in $\boldsymbol{\Omega}^j$ and $\boldsymbol{\Omega}_i$, respectively. Then the class probability is

$$f\left(\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right)=Bel\left(\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right)+\frac{\left|\boldsymbol{\Omega}^j\right|}{\left|\boldsymbol{\Omega}_i\right|}\left[Pl\left(\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right)-Bel\left(\boldsymbol{\Omega}^j\middle|\boldsymbol{x}\right)\right],\quad j=1,2,\cdots,K, \tag{7}$$

Finally, the cluster with the maximum class probability is selected for $\boldsymbol{x}$.

Algorithm 2 summarizes our ensemble clustering approach described above, the meaning of symbols used there are same with the counterparts above.

Algorithm 2. Ensemble clustering based on DS evidence theory

Input: $\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_N$

Output: Ensemble clustering results

Steps:

Step 1. For each $\boldsymbol{R}_i\big|_{i=1}^N$,

Step 1.1 Generate the power set of $\boldsymbol{R}_i$, i.e., $\boldsymbol{\Omega}_i$

Step 2. For each data point $\boldsymbol{x}$,

Step 2.1 Calculate the posterior probabilities $P\left(\boldsymbol{\Omega}_i^j\middle|\boldsymbol{x}\right)$ for each pair of i and j by using Eq. 3.

Step 2.2 Calculate the orthogonal sum $P\left(\boldsymbol{\Omega}^j\middle|x\right)$ for each j by using Eq. 4.

Step 2.3 Calculate the belief and plausibility values by using Eq. 5 and 6.

Step 2.4 Compute the class probabilities for each $\boldsymbol{\Omega}^j$ using Eq. 7 and assign $\boldsymbol{x}$ to the cluster with the maximum class probability.

4 Experiments

4.1 Experimental Setup

In the experiments, we select 4 commonly used data sets for testing our clustering method. The first two are 2-D data vector (http://cs.joensuu.fi/sipu/datasets/), including Flame (240 vectors with 2 clusters) and Jain (373 vectors with 2 clusters). The clustering results over them can be visualized intuitively. The third one is the Iris in UCI Machine Learning repository (http://arch-ive.ics.uci.edu/ml/). The Iris contains 3 classes of 50 instances each. Each instance has 4 attributes and the class label. So the clustering accuracy can be measured exactly. The last data set is KDD Cup 04Bio (http://www.sigkdd.org). It provides 145751 data points with 74 attributes. The number of clusters is given as 2000, but there is no class labeling. It can be used to test the clustering performance over large size and high dimensional data.

We implemented two single clustering algorithms by embedding *k*-means or spectral clustering method into dense Gaussian distributions based clustering method described in Section 2, i.e., using them in Step 2 of Algorithm 1, respectively. The two resultant methods are called *k*-means_G and spectal_G for short, respectively. The clustering results from them are combined by our ensemble approach and the

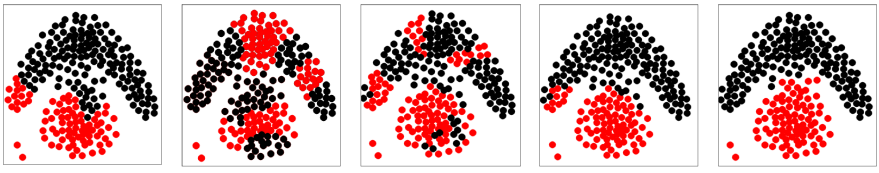
voting based counterpart of Weingessel et al. [12], respectively. We call the method of Weingessel et al. as WDH voting for the convenience of descriptions.

We use the following criteria to evaluate the accuracy and efficiency of clustering approaches: Accuracy Rate (AR), Internal Quality (IQ) and External Quality (EQ) of clusters, Execution Time (ET) and Cost of Memory (CoM).

- The AR can reflect the accuracy of clustering on the data sets with class labeling. Let a_i be the number of correctly classified instances of the i-th cluster, n be the number of all instances in the data set, then $\mathrm{AR} = \sum_{i=1}^{K} a_i \Big/ n$.
- The IQ and EQ of clusters [12] are widely used for data sets without labels. Let $\boldsymbol{C}_i$ be the i-th cluster, $\boldsymbol{\mu}_i$ be its mean vector, $d(\boldsymbol{x}, \boldsymbol{\mu}_i)$ be the Euclidean distance between a data point $\boldsymbol{x}$ and $\boldsymbol{\mu}_i$, then $\mathrm{IQ} = \sum_{i=1}^{K} \sum_{x \in \boldsymbol{C}_i} d(\boldsymbol{x}, \boldsymbol{\mu}_i)$, $\mathrm{EQ} = \sum_{1 \le j \le i \le K} d(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$. The EQ is proportional to the degree of closeness between different clusters, so the bigger EQ is, the better the clustering quality is. While the IQ is inversely proportional to the degree of data closeness within a cluster, so the less IQ is preferred.
- The ET and CoM are useful for measuring the computational complexity. Since the topic of this paper is ensemble clustering, we only consider the ensemble procedure in the calculation of ET and CoM and neglect the cost consumed for performing each single clustering method. Notice that all the following experiments are performed on a computer with 3.4GHz CPU and 10GB inner memory

4.2 Experimental Results

The results of 4 algorithms on Flame and Jain distributions are shown in Fig. 1-2, respectively. In both of two figures, sub-figures (a) to (d) show the clustering results from 4 algorithms and (e) the true distribution. The ARs for each algorithm are given in the title of each figure. It can be discovered intuitively that our DS based ensemble results are close to the true distributions. On the Flame dataset, it performed better than both single clustering methods. On the Jain dataset, it behaved better than *k*-means_G but a little worse than spectral_G. These results demonstrate that our ensemble method cannot guarantee to achieve better results than the best single result, but it does improve the worse single results obviously. Thus the combination of results is effective. Furthermore, our method behaved better than the WDH voting. Compared with it, our DS based method brought 12.7% and 11.7% increase in the AR over two data sets, respectively.



(a) (b) (c) (d) (e)

Fig. 1. The results on the Flame distribution: (a) *k*-means_G (0.85); (b) spectral_G (0.71); (c) WDH voting (0.79); (d) our method (0.89); (e) true distribution

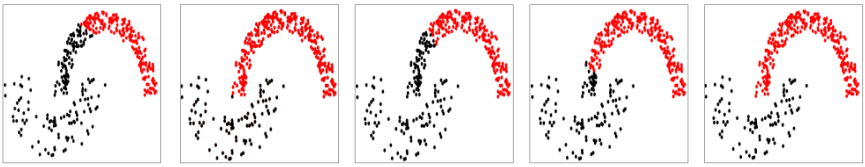


Fig. 2. The results on the Jain distribution: (a) *k*-means_G (0.83); (b) spectral_G (0.99); (c) WDH voting (0.85); (d) our method (0.94); (e) true distribution

For Iris data set, the AR from each algorithm are listed in Table 1. As shown in Table 1, both of two ensemble clustering algorithms can improve the worse single results, but our method still behaved better. Compared with the WDH voting, our method brought 10.8% increase in the AR. As for the efficiency, the ET of our method and the WDH voting on Iris data set are similar, which are 2.71s and 3.02s, respectively. The CoM is ignored since the data size is too small.

Table 1. Comparisons of ARs on Iris Data Set

k-means_G	Spectral_G	WDH Voting	Our method
84.3%	73.6%	75.1%	83.0%

For the clustering on KDD CUP 04Bio data set, the IQ and EQ are used to evaluate the effectiveness of the algorithms. The results are shown in Fig. 3. We can see that our method achieved the best IQ and the second best EQ on this data set. And the EQ from our method is very close to the best one from spectral_G. Compared with *k*-means_G, spectral_G and the WDH voting, the increase rates of IQ bought by our method are 10.3%, 23.8% and 19.1%, respectively. As for the EQ, the increase rates are 23.2%, 15.3% and -3.28% for *k*-means_G, WDH voting and spectral_G, respectively. Furthermore, the ET of our method on KDD CUP 04Bio is 13.22 minutes, which is a little better than 15.35 minutes of WDH voting. But the CoMs of ours are worse than that of WDH voting. The two values are 4.02MB and 2.88MB, respectively.

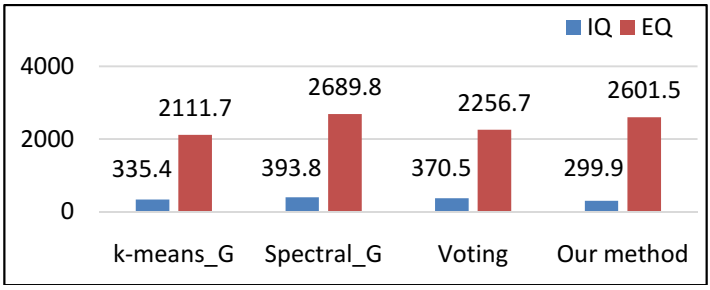


Fig. 3. IQ and EQ on KDD CUP 04Bio Data Set

5 Conclusions

In this paper, we have proposed a new ensemble clustering method based on Dempster-Shafer (DS) evidence theory and tested it for combining dense Gaussian distributions based clustering methods. The main contributions of this paper are:

(1) The Gaussian Mixture Modeling technique is introduced to compute the confidences of assigning each data point to candidate clusters, which reflects the evidences supplied by single clustering methods.

(2) The DS evidence theory is employed to combine evidences from various single clustering methods, based on which the final clustering result are obtained.

We tested the proposed approach on 4 commonly used data sets, including 2-D distributions (Flame and Jain) with intuitive visualization, Iris with exact class labeling and KDD Cup 04Bio with large size and high dimensional data. The results confirm that the proposed ensemble clustering method is effective and promising.

References

1. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence 25(3), 337–372 (2011)
2. Tumer, K., Agogino, A.K.: Ensemble clustering with voting active clusters. Pattern Recognition Letters 29(14), 1947–1953 (2008)
3. Dimitriadou, E., Weingessel, A., Hornik, K.: A combination scheme for fuzzy clustering. International Journal of Pattern Recognition and Artificial Intelligence 16(7), 901–912 (2002)
4. Wang, H., Yang, Y., Wang, H., Chen, D.: Soft-Voting Clustering Ensemble. In: Zhou, Z.-H., Roli, F., Kittler, J. (eds.) MCS 2013. LNCS, vol. 7872, pp. 307–318. Springer, Heidelberg (2013)
5. Strehl, A.: Relationship-based clustering and cluster ensembles for high-dimensional data mining. Ph.D dissertation, The University of Texas at Austin (2002)
6. Karypis, G., Kumar, V.: A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM Journal on Scientific Computing 20(1), 359–392 (1998)
7. Strehl, A., Ghosh, J.: Cluster ensembles-a knowledge reuse framework for combining partitioning. In: Proc. of 11th National Conf. on Artificial Intelligence, pp. 93–98 (2002)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39(1), 1–38 (1977)
9. Beecks, C., Ivanescu, A.M., Kirchhoff, S., Seidl, T.: Modeling image similarity by Gaussian mixture models and the signature quadratic form distance. In: Proc. of 2011 IEEE International Conference on Computer Vision (ICCV 2011), pp. 1754–1761 (2011)
10. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
11. Denœux, T., Masson, M.-H.: Dempster-Shafer Reasoning in Large Partially Ordered Sets: Applications in Machine Learning. In: Huynh, V.-N., Nakamori, Y., Lawry, J., Inuiguchi, M. (eds.) Integrated Uncertainty Management and Applications. AISC, vol. 68, pp. 39–54. Springer, Heidelberg (2010)
12. Weingessel, A., Dimitriadou, E., Hornik, K.: An ensemble method for clustering. In: Proc. of the 3rd International Workshop on Distributed Statistical Computing (2003)
13. Rokach, L.: A survey of clustering algorithms. In: Data Mining and Knowledge Discovery Handbook, pp. 269–298. Springer US (2010)