Fuzzy clustering ensemble based on mutual information

YAN GAO[1] SHIWEN GU[1] LIMING XIA[1] ZHINING NIAO[2]
[1]Faculty of Information Science and Engineering, Central South University
410075, Hunan, P.R.China China
[2]Department of Computer Science, Loughborough University
Leics, UK, LE113TU
{gaoyan,swgu,xlm}@csu.edu.cn
{

Abstract: -Clustering ensemble is a new topic in machine learning. It can find a combined clustering with better quality from multiple partitions. But how to find the combined clustering is a difficult problem. In this paper, we extend the object function proposed by Strehl & Ghosh which is based on mutual information and we present a new algorithm similar to information bottleneck to solve the object function. This algorithm can combine "soft" partitions and need not establish label correspondence between different partitions. We conducted experiments on four real-world data sets to compare our algorithm with other five ensemble algorithms, including CSPA, HGPA, MCLA, QMI. The results indicate that our algorithm provides solutions of improved quality.

Key-Words: - Clustering Ensemble, Mutual Information, soft partitions

1 Introduction

Clustering is the problem of partitioning a finite set of points in a multi-dimensional space into classes (called clusters) so that (i) the points belonging to the same class are similar and (ii) the points belonging to different classes are dissimilar. Clustering has been extensively studied in machine learning, databases, and statistics from various perspectives. Many applications of clustering have been discussed and many clustering techniques have been developed.

Ensemble learning refers to a collection of methods that learn a target function by training a number of individual learners and combining their predictions. In ensemble learning, a more reliable result can be obtained by combining the output of multiple "experts", and a complex problem can be decomposed into multiple sub-problems that are easier to understand and solve (divide-and-conquer approach). Ensemble learning is a hot topic in machine learning, and is regarded as one of four main directions in machine learning [1].

Most famous ensemble learning methods is designed for supervised learning, such as boosting [2], bagging [3]. Recently, people focus on using ensemble learning to improve performance of clustering. Fred & Jain [4], Fern & Brodley [5], Monti et al [6] established the co-association matrix based on similarities between different clustering solutions, and then use agglomerative hierarchical clustering. Topchy et al. [7][8] proposed a mixture model in order to obtain a consensus function. The basic idea is to consider the labels of the individual partitions as features characterizing the objects and the consensus partition is obtained by grouping this data set. They also established a quadratic mutual information criterion for clustering ensemble and the approximate results for this criterion can be obtained by running k-means [8]. W. Tang and Z.H.Zhou [9] proposed bagging-based selective cluster ensemble algorithm in which the mutual information between the clustering result and other results can be regard as the weight of clustering result in bagging. Frossyniotis [10] applied boosting to clustering ensemble. Strehl & Ghosh[11] proposed three different approaches to generating consensus functions, most of them based on hypergraph partitioning. They also pointed out that clustering ensemble can be regarded as the optimal problem based on mutual information, but not point out how to solve it and it is only to combine "hard" partition.

In this paper, we extend the object function based on mutual information introduced by Strehl & Ghosh, and propose a new ensemble algorithm to combine "soft" partitions. This algorithm need not consider the problem of label correspondence between different clustering results.

2 Clustering Ensemble

Definition (Clustering ensemble): Given a data set of n instances $X = \{X_1, X_2, \ldots, X_n\}$, a set of partitions produced by base clustering algorithm on this data set can be represented by $\Pi=\{\pi^1,\ldots,\pi^r\}$, where $\pi^i=\{c_1^i,\ldots c_k^i\}$, $U_k c_k^i=X$. Clustering ensemble is to deduce the final partition from this set of partitions Π.

For clustering ensemble, there are two important components: ensemble constructor and consensus function. Given a dataset, ensemble constructor generates a set of diverse partitions. Diversity guarantees that all the individual learners do not make the same errors. Consensus function is a good method to combine the different partitions and produce a single better partition on the data set. So choosing the appropriate consensus function is the key problem for clustering ensemble. In this paper, we proposed a new consensus function: ensemble algorithm based on mutual information.

3 The object function based on mutual information for clustering ensemble

In [11], Strehl and Ghosh thought that combined clustering should share the most information with the original clusterings: $\Pi=\{\pi^1, \ldots. \pi^i\}$. But how do we measure shared information between clusterings? Strehl and Ghosh used mutual information to measure it. In information theory, mutual information is a symmetric measure to quantify the statistical information shared between two distributions.

Suppose there are two partitions: π^a and π^b. Let n^h is the number of instance which label is C_l in partition π^b, n^l is the number of instance which label is C_h in partition π^a, n_l^h is the number of instance which label not only is C_l in partition π^b but also is C_h in partition π^a. n is the total number of instances in data set. In [11], Strehl & Ghosh used the normalized mutual information. So the object function is:

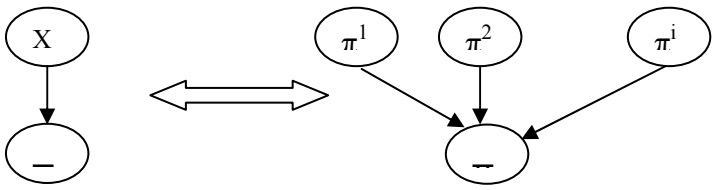
$$\phi^{NMI}(\pi^a,\pi^b)=\frac{2}{n}\sum_{k=l}^{K}\sum_{k=l}^{K}n_l^h\log_{k^2}\left(\frac{n_l^h n}{n^h n^l}\right)$$

$$\pi^{opt}=\arg\max_{\pi}\sum_{q=1}^{r}\phi^{NMI}(\pi,\pi^q) \qquad (1)$$

Although Strehl & Ghosh proposed the object function based on mutual information for ensemble clustering, they also pointed out that for finite populations, the trivial solution is to exhaustively search through all possible clusterings with k labels (approximately $k^n/k!$ for $n>> k$) for the one with the maximum ANMI which is computationally prohibitive. And this object function is only applied to combine the "hard" partitions.

We also thought that clustering ensemble should extract the combined clustering sharing the most information with the original clusterings. But in this paper, we modified the object function to cluster "soft" partitions.

First, we think that extracting clusters structure from the data can be viewed as data compression. So besides preserving more information of original clusterings, the data should be compressed as much as possible. In the information theory [12], compressing data means minimizing the mutual information between the clustering and data. Figure 1 depicts our ensemble model.



a. The compressed information b. The preserved information

Figure 1. The compressed information and the preserved information for clustering ensemble

So the new object function can be available by subtracting an item from original object function (1): $\eta\, I(X,\pi)$. For computing this function conveniently, we replaced the normalized mutual information in (1) with standard mutual information.

$$\pi^{opt}=\arg\max_{\pi}\sum_{q=1}^{r}I(\pi,\pi^q)-\eta I(x;\pi) \qquad (2)$$

$$I(\pi^a,\pi^b)=\sum_{i=1}^{k}\sum_{j=1}^{k}p(c_i^a,c_j^b)\log\frac{p(c_i^a,c_j^b)}{p(c_i^a)p(c_j^b)} \qquad (3)$$

Second, we extended the object function to combining "soft" partitions. Suppose there are a set of partitions on data set: $\Pi=\{\pi^1,\ldots\pi^r\}$, where $\pi^i=\{c_1^i,\ldots c_k^i\}$. For any two partition π^a, π^b, the information of the cluster c_i^a in partition π^a should be transferred to cluster c_j^b in partition π^b through x (c_j^b->x->c_i^a). So the $p(c_i^a,c_j^b)$ in (3) can be defined as:

$$p(c_i^a,c_j^b)=\sum_x p(c_i^a\mid x)p(c_j^b\mid x)p(x) \qquad (4)$$

$$p(c_i^a)=\sum_x p(c_i^a\mid x)p(x)$$

$$p(c_i^b) = \sum_x p(c_i^b \mid x)p(x) \qquad (5)$$

4 The solution for object function

Supposing that the labels of original clustering are used to the new feature for object, labels in every partition π^i can form the new feature space. So every object can be represented by r feature spaces. If we use y^i to represent the label in partition π^i and use c to represent label in the combined clustering. The Lagrangian function for (2) is:

$$\ell(p(c \mid x)) = \sum_{q=1}^{r} I(\pi, \pi^q) - \eta I(x;\pi) - \sum_x \lambda(x) \sum_c p(c \mid x) \qquad (6)$$

Proposition: The "local optimal" solution of (3) can be obtained by iterating the following equations:

$$\begin{cases} p_t(c \mid x) = \dfrac{p_{t-1}(c)}{Z} \exp(-\beta \sum_{i=1}^{r} \sum_{y^i} D_{KL}[p(y^i \mid x) \,||\, p_{t-1}(y^i \mid c)]) \\ p_t(c) = \sum_x p(x) p_t(c \mid x) \\ p_t(y^i \mid c) = \dfrac{1}{p_t(c)} \sum_{x \in X} p(y^i \mid x) p_t(c \mid x) p(x) \end{cases} \qquad (7)$$

where:

$$Z = \sum_c p_{t-1}(c) \exp(-\beta \sum_{i=1}^{q} \sum_{y^i} D_{KL}[p(y^i \mid x) \,||\, p_{t-1}(y^i \mid c)])$$

, $\beta = \dfrac{1}{\eta}$, D_{KL} is relative entropy of p with respect to q, it is also called Kullback Leiblers distance, $D_{KL}(p \,||\, q) = \sum_x p(x) \log \dfrac{p(x)}{q(x)}$.

Proof: Supposing that the labels in partition π^i are presented by y^i, the label in combined partition π is represented by c. The mutual information between π^i and π is:

$$I(\pi;\pi^i) = \sum_{y_i \in \pi^i} \sum_{c \in \pi} p(y^i, c) \log \frac{p(y^i, c)}{p(y^i) p(c)}$$

Considering the information about y^i must transfer to c through x , a Markov chain can consist of y^i , x and c: y^i ->x->c. So:

$$p(y^i, c) = \sum_x p(y^i \mid x) p(c \mid x) p(x)$$

For $I(\pi ; \pi^i)$, the partial derivative with respect to the variable p(c|x) is:

$$\begin{aligned} &\frac{\partial}{\partial p(c \mid x)} I(\pi;\pi^i) \\ &= \frac{\partial}{\partial p(t \mid x)} \sum_c \sum_{y^i} \sum_x p(y^i \mid x) p(c \mid x) p(x) \log \frac{p(y^i \mid c)}{p(y^i)} \\ &= p(x) \sum_{y^i} p(y^i \mid x) \log \frac{p(y^i \mid c) p(y^i \mid x)}{p(y^i \mid x) p(y^i)} \\ &= -p(x) \sum_{y^i} D_{KL}[p(y^i \mid x) \,|\, p(y^i \mid c)] \\ &+ p(x) \sum_{y^i} D_{KL}[p(y^i \mid x) \,|\, p(y^i)] \end{aligned} \qquad (8)$$

$$I(X;\pi) = \sum_x \sum_{c \in \pi} p(c \mid x) p(x) \log \frac{p(c \mid x)}{p(c)}$$

So the partial derivative of $I(X; \pi)$ with respect to the variable p(c|x) is:

$$\begin{aligned} &\frac{\partial}{\partial p(c \mid x)} I(X;\pi) \\ &= \frac{\partial}{\partial p(c \mid x)} \sum_x \sum_{c \in \pi} p(c \mid x) p(x) \log \frac{p(c \mid x)}{p(c)} \\ &= p(x) \log \frac{p(c \mid x)}{p(c)} \end{aligned} \qquad (9)$$

So according to (8), (9), the partial derivative of the function (6) with respect to the variable p(c|x) is:

$$\begin{aligned} \frac{\partial \ell}{\partial p(c \mid x)} &= p(x) \lambda(x, y^1, \ldots .. y^r) - p(x) \log \frac{p(c \mid x)}{p(c)} \\ &- \sum_{i=1}^{r} \sum_{y^i} \beta \{p(x) D_{KL}[p(y^i \mid x)] \,||\, p(y^i \mid c)]\} \end{aligned} \qquad (10)$$

where:

$$\lambda(x, y^1, \ldots .., y^r) = \frac{\lambda(x)}{p(x)} + \sum_{i=1}^{r} \sum_{y^i} \beta D_{KL}[p(y^i \mid x) \,||\, p(y^i)]$$

.

When (10) is equal to 0, the iterating equations (7) can be obtained, and function (2) has local optimal solution.

5 Algorithm implementation

According to the above proposition, we knew that given the initial distribution, the local optimal solution of (2) can be obtained by computing (7) iteratively. When the value of η is very small, the solution of (2) is the approximate solution of (1). So in our ensemble algorithm, η is 0.000001.

Observing (2) and (7), we found that when r were equal to 1, (2) is equal to the object function of

information bottleneck (IB) [13] and (7) became the iterating equations for IB. So the implementation of our algorithms is similar to IB.

There are many implementation algorithms for IB, such as sequence IB (sIB), agglomerative IB (aIB), iterative IB (iIB). Slonim [13] compared several IB algorithms, concluding that best hard clustering results are obtained with a sequential method (sIB), in which elements are first assigned to a fixed number of clusters and then individually moved from cluster to cluster while calculating a 1-step lookahead score, until the score converges. Here we adopted sIB. But we must make a little change in sIB. For clustering ensemble, we change the object function and iterating equations in sIB according to (2), (7).

6 Experiments

Table 1 The detail of four data sets

Data set	Num. of features	Num. of classes	Num. of objects	Fuzzy-kmeans (error rate)
Wine	13	3	178	0.314
Glass	9	6	241	0.509
Ionosp-here	34	2	351	0.291
Spamb-ase	57	2	4601	0.357

We experiment on four data sets from UCI [14]. The attributes in four data sets are numberical. The detail of four data sets is described in Table 1.

In experiments, we use two ensemble algorithms based on mutual information: f-MI and h-MI. f-MI is used to combine "soft" partitions and h-MI is used to combine "hard" partitions. We compared h-MI, f-MI with other ensemble algorithms, including MCLA, HGPA, CSPA and QMI. CSPA for partitioning of hypergraphs induced from the co-association values. HGPA establishes a hypergraph for cluster ensemble and partitions the hypergraph by cutting a minimal number of hyper-edges. MCLA modifies HGPA via extended set of hyperedge operations and additional heuristics, QMI is based on quadratic mutual information criterion and use k-means to obtain the approximately combined clustering. The CSPA, MCLA, HGPA code is available in [15].

In order to produce diverse partitions, we used random subspace method [16][17], where each base clustering is generated on a randomly selected subset of the original dimensions. Fuzzy k-means is used on new subspace to produce a "soft" partition. In order to obtaining the hard partitions, every object is assigned to the cluster in which its conditional probability is maximum. The maximum iterative time in fuzzy k-means is 100. The dimension of sub space is $\lceil d/4 \rceil$. The number of clusters for every data set: k, is the actual number of classes in data set.

Our algorithm is susceptible to the presence of local minima of the objective functions. To reduce the risk of convergence to a lower quality solution, we used a simple heuristic afforded by low computational complexities of these algorithms. The final partition was picked from the results of three runs (with random initializations) according to the value of objective function. The highest value of MI function (2) served as a criterion for our algorithm.

We randomly choose five value [10, 15, 20, 30, 40] for the size of cluster ensemble.

In experiments, the mean clustering error rate of 10 clustering ensemble procedures is used to measure the performance of clustering ensemble. Let C^{true} represent the true (given) clustering and C represent the ensemble clustering, Confusion represent the confusion matrix of two clusterings. Confusion(ktrue, k) = ($C_{ktrue}^{true} \cap C_k$) , i.e. number of points x that are cluster ktrue in true clustering and cluster k in the clustering produced, then

$$error\_rate = \frac{1}{n}\left(\sum_{ktrue}\sum_{k\neq ktrue} Confusion(ktrue, k)\right)/n \quad (11)$$

Where n is the total number of objects.

Clustering error rate is defined as the number of "misclassification". The low value of error indicates good quality of clustering.Figure1, 2, 3, 4 shows the mean error rate on four data sets.

Fig. 1 The mean error rate for "wine" data set

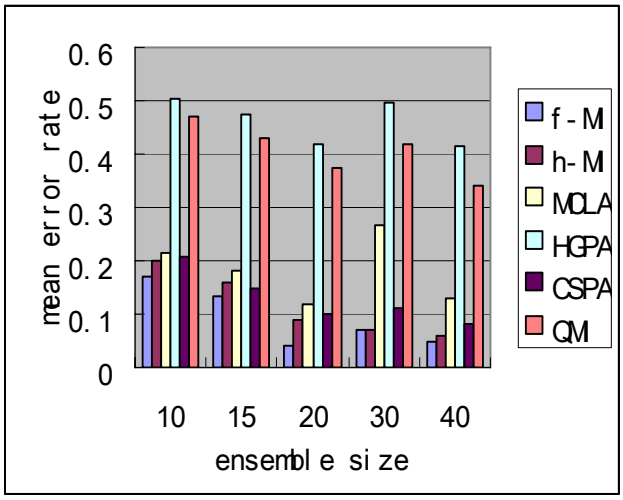


Fig. 2 The mean error rate for "glass" data set

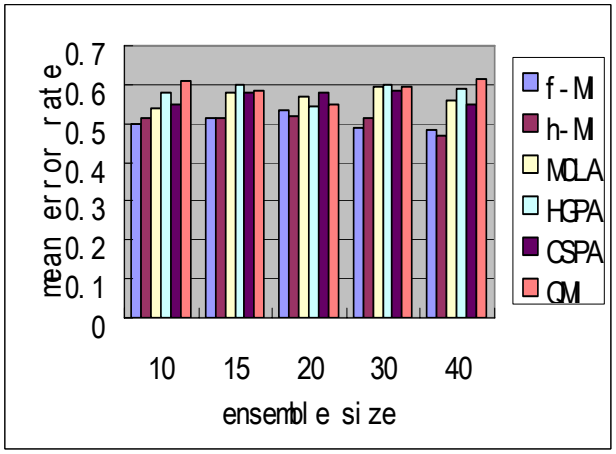


Fig. 3 The mean error rate for "spambase" data set

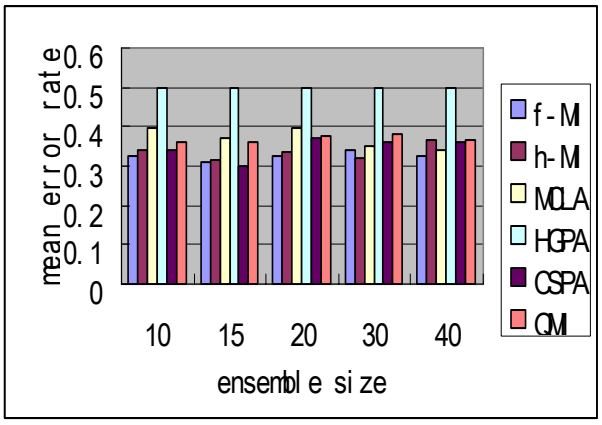
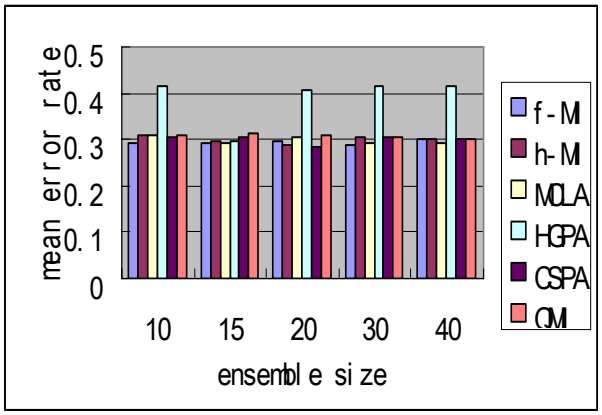


Fig. 4 The mean error rate for "ionosphere" data set



From Figure 1, 2, 3, 4, we find that among the four data sets, the performance of clustering ensemble in "wine" is best. The second is "spambase". When combining the partitions in "glass" or in "ionosphere", the performance of the five algorithms are all not good, the mean error rate is higher than that of fuzzy k-means on whole dataset.

From Figure 1,2,3 4, we also find that there is no algorithm which performance is best for all four data set with different ensemble's size. But on the whole, the performance of f-MI is best among six algorithms. Especially when the size of ensemble is small, the mean error rate of f-MI is lower than that of other algorithms, because original "soft" partitions contain much information than "hard" partitions. In experiments, although h-MI and QMI are both based on mutual information criterion for "hard" partition ensemble, h-MI provides clusters with better quality than QMI.

7 Conclusion

In this paper, we have proposed a clustering ensemble method to combine "soft" partitions. This method extends the object function based on mutual information proposed by Strehl & Ghosh. And the solution of the new object function can be obtained by using the algorithm which is similar to information bottleneck. This method is not only to combine "soft" and "hard" partitions, but also has an advantage that it need not establish label correspondence between different partitions. Experiments on 4 data sets from UCI indicate that using our algorithm to combine "soft" partitions can provide clusters with better quality than MCLA, HGPA, CSPA and QMI.

.

References:

[1] Dietterich TG, Machine learning research: Four current directions, AI Magazine, 18 (4), 1997, pp.97-136.

[2] Robert E. Shapire, A brief introduction to boosting, In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1999, pp. 1401-1406.

[3] Breiman, L, Bagging Predictors, Machine Learning, Vol. 24, No. 2, 1996, pp.123-140.

[4] A.L.N. Fred and A.K. Jain, Data Clustering Using Evidence Accumulation, In Proc. of the 16th International Conference on Pattern Recognition, ICPR 2002 , Quebec City, pp.276 – 280.

[5] Fern, X. Z., & Brodley, C. E. Random projection for high dimensional data clustering: A cluster ensemble approach, ICML 2003.

[6] Monti, S. Tamayo, P, Mesirov, J, & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, Machine Learning, vol. 52, no. 1-2, 2003, pp.91–118.

[7] A. Topchy, A. Jain, and W. Punch, A mixture model for clustering ensembles, In Proc. SIAM Data Mining, 2004, pp.379-390.

[8] A. Topchy, A. Jain, and W. Punch, Combining multiple weak clusterings, In Proc. Third IEEE International Conference on Data Mining (ICDM'03), November 2003.

[9] Wei Tang, ZhiHua Zhou, Bagging-Based Selective Clusterer Ensemble. Journal of software, Vol.16, No.4, 2005, pp.496-502

[10] D. Frossyniotis, A. Likas, A. Stafylopatis, A clustering method based on boosting, Pattern Recognition Letters 25 (2004), pp.641– 654.

[11] A. Strehl and J. Ghosh, Cluster ensembles-a knowledge reuse framework for combining partitionings, In Proc. Conference on Artificial Intelligence (AAAI 2002), Edmonton, pp. 93–98.

[12] T. M. Cover and J. A. Thomas, Elements of Information Theory, John Wiley&Sons,New York, 1991.

[13] Noam Slonim, The Information Bottleneck: Theory and Applications, PhD thesis, Hebrew University, Jerusalem, Israel, 2002.

[14] Blake C, Keogh E, Merz CJ, UCI Repository of machine learning databases, Irvine: Department of Information and Computer Science, University of California, 1998. http://www.ics. uci.edu /~mlearn/MLRepository.html

[15] CSPA, MCLA, HGPA code, http://strehl.com/

[16] 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 CBMS 2004, Bethesda, MD, National Library of Medicine/National Institutes of Health, IEEE CS Press, 2004, pp.576– 581.

[17] T. K. Ho,The random subspace method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8) ,1998, pp.832–844.