ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Consensus clustering by graph based approach

Haytham Elghazel[1], Khalid Benabdeslemi[1] and Fatma Hamdi[2]

1- University of Lyon 1, LIESP, EA4125, F-69622 Villeurbanne, Lyon, France;
{elghazel,kbenabde}@bat710.univ-lyon1.fr
2- University of Paris 13, LIPN, UMR CNRS 7030, 93430 Villetaneuse, France;
fatma.hamdi@lipn.univ-paris13.fr

Abstract. In this paper, we propose G-Cons, an extension of a graph minimal coloring paradigm for consensus clustering. Based on the co-association values between data, our approach is a graph partitioning one which yields a combined partition by maximizing an objective function given by the average mutual information between the consensus partition and all initial combined clusterings. It exhibits more important consensus clustering features (quality and computational complexity) and enables to build a combined partition by improving the stability and accuracy of clustering solutions. The proposed approach is evaluated against benchmark databases and promising results are obtained compared to other consensus clustering techniques.

1 Introduction

Consensus clustering [6], also called cluster ensemble, has received considerable attention in the statistics and machine learning communities.Different cluster ensemble approaches are considered in the literature, including graph partitioning, Voting approach, Mutual information algorithms and Co-association based functions. *Graph partitioning based methods* [1] summarize the cluster ensemble in a graph whose vertices correspond to the objects to be clustered and partition it to yield the final clustering. The *Voting Approaches* [2], also called relabeling approaches attempt to solve a correspondence problem between the labels of initial and derived clusters using a majority vote to determine the final consensus partition. *Mutual Information based approaches* [3] consider, as cluster ensemble objective function, the mutual information between the empirical probability distribution of labels in the consensus partition and the labels in the ensemble. *Co-association based functions* compute the co-association values for every pair of objects, as the number of clusters shared by these objects in the initial partitions and feed them into any reasonable similarity based clustering algorithms, such as hierarchical clustering and graph partitioning [1].

In this paper, we present a new efficient method, called GCons to solve the cluster ensemble problem. From the co-association values between objects, GCons approachs the problem by first transforming the set of initial clusterings into a graph representation and then partition it using a minimal coloring mechanism. Unlike the traditional graph partitioning methods, the main advantage to adopt the minimal coloring paradigm is its ability to ensure a high cohesion within the generated clusters that we will show its strong relation to the cluster ensemble objective function.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

2 GCons : A New Graph Based Consensus Function

In this section, a *minimal coloring based consensus clustering method* is proposed. Given a data set $\boldsymbol{X} = \{x_1, \ldots, x_n\}$ and an ensemble of r clusterings (partitions) $\boldsymbol{\Pi} = \{\pi_1, \ldots, \pi_r\}$ with the *q-th* clustering π_q having k_q clusters, a *consensus function* Γ is defined as a function $\mathbb{N}^{n \times r} \longrightarrow \mathbb{N}^n$ mapping a set of clusterings to a combined (integrated) clustering λ (*i.e.* $\Gamma : \boldsymbol{\Pi} \longrightarrow \lambda$). Our main goal is to construct a consensus partition without the assistance of the original patterns in $\boldsymbol{X}$, but only from their cluster labels. As showed in [1], the optimal consensus partition should *share* as much *information* as possible with the given original r clusterings. Therefore, the optimal combined clustering λ_{opt} will be defined as the one that has ***maximal*** average mutual information with all individual clusterings π_q. As given in [1], using the definition of normalised mutual information estimate(NMI) between two clusterings (*c.f.* eq.(1)), our objective function can be written as the average of pair-wise NMI between the combined partition and initial clusterings. One can easily compute its value for a candidate partition solution λ and the ensemble of r clusterings $\boldsymbol{\Pi}$ as in equation 2.

$$\phi_{NMI}(\pi_a, \pi_b) = \frac{\sum_{h=1}^{k_a} \sum_{l=1}^{k_b} n_{h,l} \log\left(\frac{n \cdot n_{h,l}}{n_h^a \cdot n_l^b}\right)}{\sqrt{\left(\sum_{h=1}^{k_a} n_h^a \cdot \log \frac{n_h^a}{n}\right)\left(\sum_{l=1}^{k_b} n_h^b \cdot \log \frac{n_h^b}{n}\right)}} \tag{1}$$

where n_h^a is the number of objects in cluster C_h according to the partition π_a, $n_{h,l}$ denote the number of objects that are in cluster C_h according to π_a as well as in group C_l according to π_b.

$$\phi(\boldsymbol{\Pi}, \lambda) = \frac{1}{r} \sum_{q=1}^{r} \phi_{NMI}(\lambda, \pi_q) \tag{2}$$

In the remainder of this section, we present an elegant solution to the consensus problem by developping a consensus function based on graph minimal coloring algorithm. Our function approaches the problem by first transforming the set of clusterings into a graph representation. However this function needs to a definition of dissimilarity level between objects. Given r component clusterings, the overall dissimilarity matrix D for objects is just the complement of the *co-association matrix* [1], with entry $D(i, j)$ denoting the fraction of components in the ensemble in which the two objects i and j are not assigned together. Based on D, the objects set $\{x_1, x_2, \ldots, x_n\}$ can be conceived as a weighted linkage graph $G = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{v_1, v_2, \ldots, v_n\}$ is the vertex set which corresponds to the objects (v_i for x_i), and $\boldsymbol{E} = \boldsymbol{V} \times \boldsymbol{V}$ is the edge set which corresponds to a pair of vertices (v_i, v_j) weighted by the dissimilaritiy $D(i, j)$.

As said before, the optimal combined clustering should share the most information with the original clusterings. Under the dissimilarity definition, the maximization of the underlying objective function ϕ (*c.f.* eq.(2)) can be related to the minimization of the *total intracluster dissimilarity criterion* of the combined partition. Essentially, if two objects are grouped together in the combined

partition, the fraction of components that not assign them together should be small (denoting that they are considered to be *fully similar*). Therefore, an adopted definition of optimal consensus clustering is a partitioning that *minimizes* dissimilarities within clusters. This condition amount to saying that edges between two vertices within one cluster should be small weighted. The partitioning problem can be formulated as a *graph minimal coloring problem*.

In [4], Hansen and Delattre showed that the partitioning problem into k classes with a minimal *diameter* (The *diameter* of one cluster is the largest dissimilarity between two objects belonging to the same cluster.), an equivalent criterion to the *total intracluster dissimilarity* one, can be reduced to the minimal coloring problem of a *superior threshold graph* in which vertices correspond to objects and edges correspond to dissimilarities between two elements which is higher than a given threshold value θ chosen among the dissimilarity matrix D. In other words, $G_{>\theta}$ is given by $\boldsymbol{V} = \{v_1, v_2, \ldots, v_n\}$ as vertex set and $\boldsymbol{E_{>\theta}} = \{(v_i, v_j) | D(i,j) > \theta\}$ as edge set. The goal is to divide the vertex set $\boldsymbol{V}$ into a combined partition $\lambda = \{C_1, C_2, \ldots, C_k\}$ (when k is not predefined).

Despite the fact that the r clustering components are considered to be obtained from diverse clustering strategies (different clustering approaches or views of the data), they can share common informations. Indeed, a reasonnable number of objects can be clustered together in all r components of the ensemble and having then a pairwise dissimilarity of 0. Therefore, it can be analytically shown that the dissimilarity matrix D is generally a sparse matrix. Under this assumption, we propose a pre-treatment step which concern the construction of the superior threshold graph that will be presented to the minimal coloring algorithm. For that, we need to introduce the following definition:

Definition 1 A composite vertex v' is a subset of objects such that all pairs among these objects appear together in the r initial clusterings.

The superior threshold graph $G_{>\theta}$ will be transformed to $G^{'}_{>\theta} = (\boldsymbol{V'}, \boldsymbol{E}^{'}_{>\theta})$ given by the following instructions:

- Using the previous definition 1, find the overall composite vertex set $\boldsymbol{V_1^{'}} = \{v_1^{'}, v_2^{'}, \ldots, v_{n_1}^{'}\}$ from the original vertex set $\boldsymbol{V}$. The composite vertices in $V_1^{'}$ are pairwise disjointed. In the other hand, the remaining vertices $(\boldsymbol{V} \setminus \bigcup_{i=1}^{n_1} v_i^{'})$ which are not involved in any composite vertex are affected each one to a proper composite vertex in the set $V_2^{'} = \{v_{n_1+1}^{'}, ..., v_m^{'}\}$. Finally, $V_1^{'}$ and $V_2^{'}$ are combined into $\boldsymbol{V'} = \{v_1^{'}, ..., v_m^{'}\}$ where $m < n$.

- The dissimilarity matrix D wil be reduced to a new dissimilarity matrix D'. Since each pair of objects x_i and x_j from the same composite vertex $v_i^{'}$ are always grouped together in all r clusterings, $D(x_i, x_k) = D(x_j, x_k)\ \forall x_k \in \boldsymbol{X}$. Consequently, $D^{'}(i,j)$ between $v_i^{'}$ and another composite vertex $v_j^{'}$ is given by the dissimilarity between any two pair of objects from both vertices. Likewise the construction of $\boldsymbol{E_{>\theta}}$ the edge set $\boldsymbol{E^{'}_{>\theta}}$ is given by $\{(v_i^{'}, v_j^{'}) | D'(i,j) > \theta\}$ where θ is chosen among the reduced dissimilarity matrix D'.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

This pre-treatment step is very important to the consensus problem since (1) it allows to decrease the runtime of the partitioning algorithm (we are dealing with $m < n$ vertices) and (2) it offers the possibility to minimize the intracluster dissimilarity (and then to maximize our objective function ϕ) since the fully similar objects are pre-clustered together before performing any partitioning.

In such *superior threshold graph* $G'_{>\theta}$, the *minimal coloring* is NP-*complete* and consists to determine the minimum number of colors (clusters) needed to color the vertices of the graph such that no two adjacent vertices (dissimilar in the sense of threshold θ) have the same color (*proper coloring*). A variety of approximations and search algorithms have been developed to solve the minimal graph coloring problem in a reasonable amount of time. The simplest and well-known graph minimal coloring algorithm is the *Largest First* (LF) one developped by *Welsh and Powell* in [5]. This algorithm, easy to implement and fast, sorts the vertices by decreasing degree. The top vertex is put in color class number one. The remaining vertices are considered in order, and each is placed in the first color class for which it has no adjacencies with the vertices already assigned to the class. If no such class exists, then a new class is created. The main problem of LF algorithm is to find the appropriate vertex to color it when there is a choice between many vertices with the same degree. For an illustration purpose, suppose that we have two adjacent vertices v'_i and v'_j having the same degree and no neighbors in one color c. Therefore, if v'_i is selected for coloring, it can be assigned to color c which will not be possible after for v_j, and vice versa. We note the reliance of the coloring based partitioning result to the selection manner for such vertices. As a solution, GCons constrains this choice to ***maximize*** the *intracluster homogeneity* and then our *objective function* ϕ of the returned (combined) partition λ. We propose the following strategy: when one vertex v'_i with degree d is selected for coloring and the first color c different from those of its neighborhood is found, the vertices not yet colored, having the same degree d and without any neighbor in c, will be simultanously considered for coloring. So the vertex whose dissimilarity with c is minimal will be the first to color with c and the remaining vertices will be considered later. GCons's complexity is $O(n^2)$ which is reduced to $O(m^2)$ $(m < n)$ after pre-treatment step.

3 Experimental Results

In this section, we illustrate our algorithm's performance on several relevant benchmark data sets [7] (*c.f.* Table 1). For our experiments, four clustering approaches are used to generate the partitions for the combination: (1) *k-means*; (2) *Agglomerative Hierarchical Classification* (AHC) in the form of *Ward-based* approach; *Self-Organizing Map* clustered based on (3) k-means, and (4) AHC. The four clusterings are then integrated using our proposed GCons approach and three other graph based consensus functions : CSPA, MCLA and HGPA [1]. We note that GCons is iterative and performs multiple runs, each of them increasing the value of the dissimilarity threshold θ. Once all threshold values passed,

the algorithm provides the best combined partition λ (corresponding to one threshold value θ_o) with the highest objective function $\phi(\mathbf{\Pi}, \lambda)$ (*c.f.* eq.(2)). For an interesting assess of the results gained with the different consensus clustering approaches the following performance indices are used:

- The *objective function* $\phi(\mathbf{\Pi}, \lambda)$ (*c.f.* eq.(2)). It gives an idea about the dependency between the combined partition and the four clusterings in the ensemble.
- A *statistical-matching* scheme given by the *Normalized Mutual Information* $\phi_{NMI}(\lambda, L)$ (*c.f.* eq.(1)). In our case, the used UCI data sets include *class information* (*label*) for each data instance. These labels are available for evaluation purposes but not visible to the clustering algorithms. Indeed, evaluation is based on this scheme in order to assess the *degree of agreement* between the combined partition λ and the correct predefined one $L(labels)$. When comparing two consensus clustering algorithms, the one that produces the greater ϕ_{NMI} should be preferred since the partition correctly identifies the underlying classes in the data set.

Data sets	*instances*	*features*	*#labels*
Wdbc	569	30	2
Rings	1000	3	2
Image Segmentation	2310	19	7
Engytime	4096	2	2

Table 1: Characteristics of used data sets.

Data sets	*CSPA*	*HGPA*	*MCLA*	*GCons*
Wdbc	0.2892	0.0001	0.7471	0.8830
Rings	0.6383	0.0010	0.6260	0.6663
Image Segmentation	0.6656	0.4713	0.6593	0.7612
Engytime	0.7952	0.0001	0.8072	0.8191

Table 2: Comparison of consensus functions in terms of the objective function.

Table 2 provides the clustering results according to the *objective function*. The reported values indicate better consensus clustering for all partitions generated by the proposed GCons approach. The combined partitions given from GCons are thus *highly related* to all individuals clusterings and share the most information with them, compared to the other consensus functions. HGPA performs the worst in these experiments which is also highlighted in [1]. This confirms the pertinence of the *graph minimal coloring technique* to offer a consensus partition with minimal diamater and then reaching a larger objective function.

Table 3 provides the clustering results according to the *normalized mutual information* with original labels. The ranking of the consensus algorithms is

Data sets	*CSPA*	*HGPA*	*MCLA*	*GCons*	*AIA**
Wdbc	0.0973	0.0007	0.3985	0.4514	0.4242
Rings	0.0705	0.0001	0.1296	0.2971	0.1703
Image Segmentation	0.4553	0.3134	0.4801	0.5760	0.4874
Engytime	0.7197	0.0001	0.7228	0.7278	0.6994

Table 3: Comparison of consensus functions in terms of their normalized mutual information with original labels. *AIA: Average Individual Algorithms

the same using this measure, with GCons best, followed by the other consensus clustering approaches: MCLA, CSPA, and HGPA worst. This indicates that the objective function we used $\phi(\mathbf{\Pi}, \lambda)$ is a suitable choice in real applications where the labels are not available. Consequently, it is observed that GCons achieves the highest correspondance with the *correct predefined partition* even when compared to the average quality of all individual algorithms. In fact, the average GCons normalized mutual information based quality $\phi_{NMI}(\lambda, L)$ over all data sets is 26% higher than the average quality of all individual clustering algorithms.

4 Conclusion

In this work we have proposed GCons, a *graph minimal coloring based approach* for consensus clustering. Two problems have been considered: 1) why the *minimal coloring paradigm* is well adapted to the cluster ensemble problem; 2) how to adopt it as best as possible in order to yield a good consensus partition. GCons is evaluated against benchmark data sets and the results of this study demonstrate that the proposed cluster ensemble approach is able to combine individual partitions in a better way than a well known graph partitioning based consensus methods and indicate the effectiveness of *minimal coloring paradigm* to offer an elegant solution to the cluster ensemble problem.

References

[1] A. Strehl and J. Ghosh, Cluster ensembles - a knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research*, 3: 583-617, 2002.

[2] S. Dudoit, J. Fridlyand, Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9): 1090-1099, 2003.

[3] A. Topchy, A. K. Jain and W. Punch, Clustering ensembles: Models of consensus and weak partitions, *IEEE Trans on Patt Anal and Mach Intell*, 27(12): 1866-1881, 2005.

[4] P. Hansen and M. Delattre, Complete-link cluster analysis by graph coloring, *Journal of the American Statistical Association*, 73: 397-403, 1978.

[5] D. J. A. Welsh and M. B. Powell, An upper bound for the chromatic number of a graph and its application to timetabling problems, *Computer Journal*, 10(1): 85-87, 1967.

[6] R. Ghaemi, N. Sulaiman, H. Ibrahim, and N. Mustapha, A survey: Clustering ensembles techniques. In *World Academy Science, Engineering and Technology*, 38, 2009.

[7] C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998.