Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

CCE: A Coupled Framework of Clustering Ensembles

Zhong She, Can Wang and **Longbing Cao**
Advanced Analytics Institute, University of Technology, Sydney, NSW 2008, Australia
{zhong.she, can.wang}@student.uts.edu.au; longbing.cao@uts.edu.au

Abstract

Clustering ensemble mainly relies on the pairwise similarity to capture the consensus function. However, it usually considers each base clustering independently, and treats the similarity measure roughly with either 0 or 1. To address these two issues, we propose a coupled framework of clustering ensembles *CCE*, and exemplify it with the coupled version *CCSPA* for *CSPA*. Experiments demonstrate the superiority of *CCSPA* over baseline approaches in terms of the clustering accuracy.

Introduction

Clustering ensemble (Christou 2011) has emerged as a hot research topic in recent years. By combining various clustering results, it enhances the clustering accuracy, robustness, and parallelism (Strehl and Ghosh 2002), etc. The whole process of clustering ensemble can be divided into two parts: building base clusterings and aggregating base clusterings. Various heuristics have been proposed to build the ensemble members, e.g., random initializations (Christou 2011). Meanwhile, how to combine the results of the base clusterings can be constructed by three kinds of methods: the consensus functions (Strehl and Ghosh 2002), the categorical clusterings such as *LIMBO* (Gionis, Mannila, and Tsaparas 2007), and the direct optimizations (Christou 2011).

Here, we mainly focus on the branch of consensus functions for the aggregation part, which include the direct best matching, hyper-graph mappings, and pairwise approaches, etc. With respect to the pairwise-based comparisons, the existing methods, such as *CSPA* (Strehl and Ghosh 2002) and *Correlation Mapping* (Gionis, Mannila, and Tsaparas 2007), have two shortcomings: (1) The similarity between data objects is binary, namely either 0 or 1. Such binary measure is rather rough in terms of capturing the relationships between data objects. (2) The consensus among base clusterings is considered independent of the couplings between these clustering results. However, the base clustering outcomes are expected to have some relationships with each other since they are induced from the same data set. No work on clustering ensemble that systematically takes into account the couplings between data objects and between base clusterings has been reported in the literature.

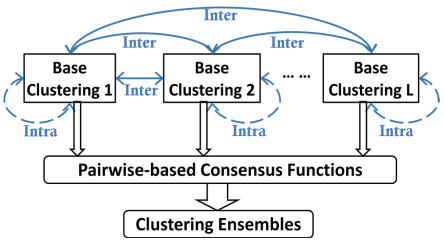

Figure 1: A coupled framework of clustering ensembles (*CCE*), where ⇠----⇢ indicates the intra-coupling within one base clustering result and ↔ refers to the inter-coupling between different base clustering results.

In this paper, we propose a coupled framework of clustering ensembles (*CCE*) by considering both the relationships within each base clustering and the interactions between distinct base clusterings, which also leads to a more accurate similarity ($\in [0, 1]$) between data objects. As indicated by Gionis, Mannila, and Tsaparas (2007), clustering ensemble can be converted to the problem of clustering categorical data by viewing each attribute as a way of producing a base clustering of the data. Thus, the coupled nominal similarity measure *COS* presented by Wang et al. (2011) can be used here to explicitly specify the couplings within and between different base clustering results. Our proposed framework *CCE* is shown in Figure 1. Below, we will illustrate this framework with an instance of the coupled version for *CSPA* (Strehl and Ghosh 2002), which is a classic and popular heuristic ensemble algorithm.

Problem Statement

We construct an information table S, all the objects for clustering consist of $U = \{u_1, \cdots, u_m\}$, and the outcomes of the L base clusterings are mapped into a set of attributes $A = \{a_1, \cdots, a_L\}$. Accordingly, the attribute value x_{ij} indicates the label of a cluster to which the object u_i belongs in the jth base clustering.

Strehl and Ghosh (2002) proposed a pairwise-based approach *CSPA*, which induces a similarity measure from the base clusterings and then reclusters the objects. The entry of the induced similarity matrix is the weighted average sum of each associated pairwise similarity between objects for

every base clustering, however, the pairwise similarity measure is rather rough since only 1 (when $x_{i_1 j} = x_{i_2 j}$, i.e., two objects have the same label), and otherwise 0, are considered. Besides, neither relationship within nor between base clusterings (i.e., attributes a_{j_1}, a_{j_2}) is explicated.

Therefore, based on the *COS* (Wang et al. 2011), we introduce a coupled version of *CSPA*, i.e., *CCSPA*, to solve the above two issues. Here, we define the coupled similarity:

Definition 1 *Given an information table S composed of a finite set of base clustering results, then the* ***Coupled Similarity*** *between data objects u_{i_1}, u_{i_2} in terms of the jth clustering result is defined as*

$$\delta_j^A(u_{i_1}, u_{i_2}) = \delta_j^{Ia}(u_{i_1}, u_{i_2}) \cdot \delta_j^{Ie}(u_{i_1}, u_{i_2}), \qquad (1)$$

where $\delta_j^{Ia}(u_{i_1}, u_{i_2})$ refers to the ***Intra-coupled Base Clustering Similarity*** *which captures the base clustering label frequency distribution, and $\delta_j^{Ie}(u_{i_1}, u_{i_2})$ denotes the* ***Inter-coupled Base Clustering Similarity*** *which characterizes the co-occurrence between base clusterings by data objects.*

Analysis and Algorithm

In *CCSPA*, we have considered the couplings both within and between base clusterings, which is a reasonable and sound alternative for the clustering ensemble. Besides, according to (Wang et al. 2011), we have the coupled similarity $\delta_j^A(u_{i_1}, u_{i_2}) \in [0, m/m+2]$ rather than $\delta_j^A(u_{i_1}, u_{i_2}) \in \{0, 1\}$ in *CSPA*, which indicates a more accurate measure. Below, Algorithm 1 describes the main process of *CCSPA*.

Algorithm 1: Coupled version of CSPA: $CCSPA()$

Data: Data set T with m objects, the number of clusters k and base clusterings L, and weight $\alpha = (\alpha_j)_{1\times L} \in [0, 1]$.
Result: The final result P^* of the clustering ensemble.
begin
  Generate base clusterings $C = \{P_1^{(k)}, \cdots, P_L^{(k)}\}$ for T
  Create an information table $S_{m\times L}$ according to C
  for *base clustering* $j = 1 : L$ **do**
    for *every object pair* $(i_1, i_2 \in [1, m])$ **do**
      $\delta_j^A(i_1, i_2) = COS\_IRSI(P_j^{(k)}(i_1), P_j^{(k)}(i_2))$
  Establish a consensus matrix with the (i_1, i_2) entry:
  $M(i_1, i_2) = \sum_{j=1}^{L} \alpha_j \times \delta_j^A(i_1, i_2)$
  Use a specific clustering method $\mathscr{A}$ on M to get P^*
end

In the above algorithm, $P_j^{(k)}(i)$ is the label for object i in clustering $P_j^{(k)}$; α_j is the user predefined weight for each base clustering; function $COS\_IRSI(\cdot)$ is applied to compute the coupled object similarity as in (Wang et al. 2011); and the final clustering method is any clustering algorithm which operates directly upon a similarity matrix or graph, such as *Single Linkage* and *METIS* (Strehl and Ghosh 2002).

Experimental Results

We conduct experiments on ten UCI data sets to verify the superiority of *CCSPA*. Firstly, various effective base clusterings are generated by bagging algorithm. In detail, k-means

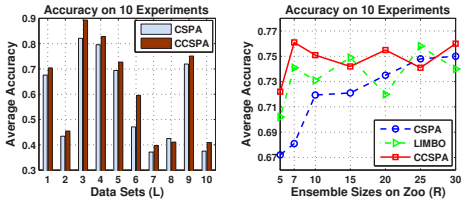

Figure 2: Accuracies on data sets as listed in Table 1

Table 1: Times of Accuracy (*CCSPA*) vs. Accuracy (*CSPA*)

Data set	High/Equal/Low	Data set	High/Equal/Low
Ionos	4/5/1	*Balance*	5/3/2
Hayes	5/5/0	*Tae*	5/4/1
Iris	6/4/0	*Vehicle*	5/4/1
Fisher	4/6/0	*Zoo*	3/7/0
Wine	4/6/0	*Segment*	5/5/0

is applied on a number of training samples picked from each data set to get N initial clustering results, these partitions whose mutual information values are the L largest are then selected as the base clusterings. Next, it is followed by combining the results of these base clusterings. For simplicity, we assume the weight $\alpha_j = 1/L$, $L = 10$, $\mathscr{A}$ is *METIS*, and experiments are conducted ten times on each data set.

Figure 2(L) shows that *CCSPA* outperforms *CSPA* in terms of the average accuracy on almost every data set by at least 2%. Moreover, Table 1 reveals that, for most times, the performance of *CCSPA* is better, at least not worse than that of *CSPA*. In addition, the impact of different ensemble sizes on *Zoo* is studied in Figure 2(R). It indicates that the average accuracies of *CCSPA* do not monotonically increase as *CSPA* does. When compared with the baseline approaches (i.e. *CSPA* and *LIMBO*), the average accuracy of *CCSPA* is much more stable and larger mostly only except few points.

Conclusion

We propose a new framework *CCE*, and exemplify it with *CCSPA* which considers both couplings within and between base clusterings in terms of the pairwise similarity between objects with impressive performance. We are currently applying this framework to the pairwise-agreement-based algorithm proposed by Gionis, Mannila, and Tsaparas (2007).

Acknowledgement

This work is sponsored in part by ARC Discovery Grants (DP1096218) and ARC Linkage Grant (LP100200774).

References

Christou, I. 2011. Coordination of cluster ensembles via exact methods. *IEEE TPAMI* 33(2):279–293.

Gionis, A.; Mannila, H.; and Tsaparas, P. 2007. Clustering aggregation. *ACM TKDD* 1(1):1–30.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles–a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3:583–617.

Wang, C.; Cao, L.; Wang, M.; Li, J.; Wei, W.; and Ou, Y. 2011. Coupled nominal similarity in unsupervised learning. In *CIKM 2011*, 973–978.