Adaptive Noise Immune Cluster Ensemble Using Affinity Propagation

(Extended Abstract)

Zhiwen Yu, Guoqiang Han
School of Computer Science and Engineering
South China University of Technology
Email: zhwyu@scut.edu.cn

Le Li
Department of Computer Science and Engineering
Chinese University of Hong Kong

Jiming Liu
Department of Computer Science
Hong Kong Baptist University

Jun Zhang
School of Advanced Computing
Sun Yat-Sen University

Cluster ensemble, as one of the important research directions in the ensemble learning area, is gaining more and more attention, due to its powerful capability to integrate multiple clustering solutions and provide a more accurate, stable and robust result [1]. Cluster ensemble has a lot of useful applications in a large number of areas. Although most of traditional cluster ensemble approaches obtain good results, few of them consider how to achieve good performance for noisy datasets. Some noisy datasets have a number of noisy attributes which may degrade the performance of conventional cluster ensemble approaches. Some noisy datasets which contain noisy samples will affect the final results. Other noisy datasets may be sensitive to distance functions.

To address the challenges posed by these noisy datasets, we propose a new noise immune cluster ensemble framework, named as the double affinity propagation based cluster ensemble (AP^2CE, [1]), which integrates the affinity propagation algorithm (AP) [2] and the normalized cut algorithm (Ncut) [3] into the cluster ensemble framework. This is in view of the capability of AP to capture the relationship among the attributes, to find a set of representative attributes, and to remove noisy attributes.

Figure 1 provides an overview of the double affinity propagation based cluster ensemble framework(AP^2CE). Specifically, Given a dataset $\mathbf{S}$ with n samples and m attributes, AP^2CE first transforms $\mathbf{S}$ to a new $m \times n$ data matrix $\mathbf{S}'$. Then, the affinity propagation algorithm (AP) designed by Frey et al.[2] is adopted to identify d representative attributes $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_d\}$ from the dataset $\mathbf{S}'$ from the original set of m attributes $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m\}$ (where $d << m$), and discover the underlying relationships among them. AP iteratively refines the representative attributes by exchanging information between data points until a set of high quality representative attributes is obtained.

By adjusting different distance functions and different parameter settings of AP, AP^2CE generates a set of new data matrices $\{\mathbf{S}^{1'}, \mathbf{S}^{2'}, ..., \mathbf{S}^{B'}\}$ (where B is the number of new datasets). It further transforms the data matrix $\mathbf{S}^{b'}$ with $d_b \times n$ entries to the new data matrix $\mathbf{S}^b$ with $n \times d_b$ entries (where $b \in \{1, 2, ..., B\}$). As a result, AP^2CE obtains a set of new datasets $\{\mathbf{S}^1, \mathbf{S}^2, ..., \mathbf{S}^B\}$.

In the following, AP^2CE adopts AP to perform clustering on the set of new datasets, and generates a set of clustering solutions $\{I^1, I^2, ..., I^B\}$.

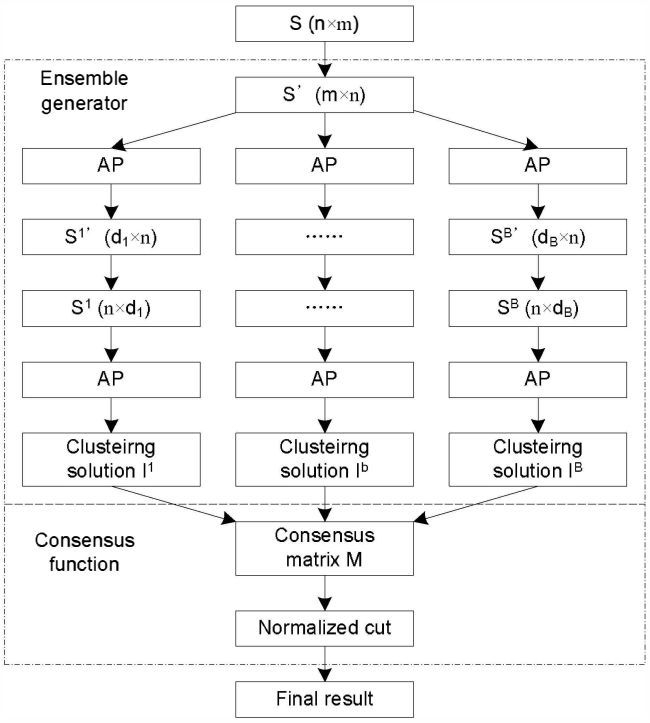

Fig. 1. Double affinity propagation based cluster ensemble framework AP^2CE

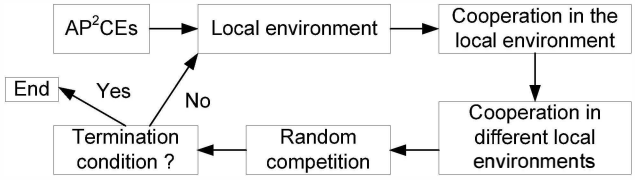

Fig. 2. An overview of the adaptive double affinity propagation based cluster ensemble (Adaptive AP^2CE)

Finally, AP^2CE constructs a graph $\Delta = (\mathbf{S}, \mathbf{\Psi})$ based on the original dataset S and the consensus matrix $\mathbf{\Psi}$, and adopts the normalized cut algorithm (Ncut) to partition the consensus matrix $\mathbf{\Psi}$ and summarize the clustering solutions obtained by AP in different runs, and obtains the final result. Instead of using Ncut as the consensus function, we also adopt AP as the consensus function and propose the triple affinity propagation

TABLE I. THE PERFORMANCE OF $M-AP$, AP^2CE AND ADAPTIVE AP^2CE ($A-AP^2CE$) WITH RESPECT TO THE AVERAGE VALUES AND THE CORRESPONDING STANDARD DEVIATIONS OF RI ON ALL THE DATASETS

Dataset	WDBC	Wine	Dermatology	Glass	DLBCL-A	Lung	Novartis	St. Jude
$AVG-AP$	0.7949 (0.0881)	0.7446 (0.0177)	0.8361 (0.0077)	0.7226 (0.0245)	0.7352 (0.0291)	0.5628 (0.0341)	0.9151 (0.0255)	0.8808 (0.0210)
AP^2CE	0.8768 (0.0131)	0.9328 (0.0295)	0.9187 (0.0096)	0.7088 (0.0216)	0.8603 (0.0105)	0.9399 (0.0039)	**0.9853** (0.0066)	0.9773 (0.0035)
$A-AP^2CE$	**0.8813** (0.0080)	**0.9428** (0.0114)	**0.9731** (0.0020)	**0.7245** (0.0141)	**0.9079** (0.0236)	**0.9566** (0.0185)	0.9811 (0.0014)	**0.9816** (0.0016)

based cluster ensemble framework AP^3CE.

We also perform a theoretical analysis of AP^2CE and AP^3CE on their time and space complexity. The computational cost of AP^2CE is approximately $O(n^3)$, while the memory consumption is $O(t^2)$ (where $t=\max(n,m)$), n is the number of the samples, m is the number of the attributes. In addition, the computational cost of AP^3CE is approximately $O(t^2)$, and the memory consumption of AP^3CE is $O(t^2)$.

The adaptive AP^2CE as illustrated in Figure 2 is further designed to improve the performance of AP^2CE, which adopts a newly proposed optimization process to search for the optimal AP^2CE. Specifically, adaptive AP^2CE first generates a set of AP^2CEs $\Upsilon=\{\Upsilon^1,\Upsilon^2,...,\Upsilon^E\}$ (where E is the number of AP^2CEs, and Υ^e denotes the e-th AP^2CEs, $e\in\{1,...,E\}$). Then, the local environment is constructed for the AP^2CEs. The objective function of adaptive AP^2CE is also formulated. In the following, adaptive AP^2CE defines three kinds of operators, which are the cooperation in the local environment, the cooperation in different local environments and the random competition. It recursively executes the above three operators. When the termination condition is satisfied, AP^2CE will terminate. The termination condition is specified by the user, which is that the maximum number of iterations reaches, or the algorithm keeps stable in several runs. In summary, there are E updated AP^2CE during the adaptive process. The optimal AP^2CE as $AP^2CE(O)$ is one of E updated AP^2CE with the minimum value of the objective function.

The performance of AP^2CE is evaluated using three synthetic datasets, four real datasets from the UCI machine learning repository, and four cancer gene expression profiles, with respect to the purity measure and the rand index.

In order to explore the effect of noisy attributes, we perform the experiments with AP^2CE on synthetic datasets with different levels of noisy attributes. (where m_n denotes the number of noisy attributes). Synthetic41 (S41), Synthetic42(S42), Synthetic43(S43) and Synthetic44(S44) are the same datasets which are injected by different levels of noisy attributes. The experimental results with respect to RI obtained by AP^2CE on synthetic datasets with different levels of noisy attributes are comparable. In general, AP^2CE is not sensitive to the datasets with noisy attributes.

We also compare $AVG-AP$, AP^2CE with adaptive AP^2CE ($A-AP^2CE$). $AVG-AP$ denotes the average value of the RI values for each clustering solution in the ensemble of AP^2CE, while adaptive AP^2CE incorporates an adaptive process for the cluster ensemble based on RI. As shown in Table I, $A-AP^2CE$ outperforms AP^2CE on 7 out of the 8 datasets, which attests to the effectiveness of adopting a judicious adaptation process for the cluster ensemble. When compared with AP^2CE and $A-AP^2CE$, the results obtained by $AVG-AP$ are less satisfactory. These results demonstrate the rationale of using ensemble of AP instead of single AP to deal with the clustering problem of noisy dataset.

In this paper, we investigate the problem of clustering datasets with noisy attributes. Our major contribution is a double affinity propagation based cluster ensemble framework AP^2CE which is applied to perform clustering on noisy datasets. We perform a thorough investigation of AP^2CE on both synthetic datasets and real datasets in our experiments, and obtain several conclusions. First, a suitable preference measure value will improve the performance of the proposed approach. Second, adopting multiple distance functions will achieve better performance. Third, if one of the AP stages in AP^2CE is missing, the performance of AP^2CE will be compromised. Fourth, AP^2CE is a better choice for most of the datasets when compared with other cluster ensemble approaches. We also compare AP^2CE with a number of single clustering algorithms and other cluster ensemble approaches on both synthetic datasets and real datasets. The results in the experiments show that (1) AP^2CE outperforms most of the state-of-the-art clustering approaches. (2) AP^2CE provides a more accurate, stable and robust results on most of the real datasets. (3) The adaptive process is able to improve the performance of AP^2CE further. In the future, we shall consider how to improve the efficiency of AP^2CE.

ACKNOWLEDGMENT

The work described in this paper was partially funded by the grant from the National High-Technology Research and Development Program (863 Program) of China No. 2013AA01A212, the grant from the NSFC for Distinguished Young Scholars 61125205, and the grants from the NSFC No. 61332002, No. 61300044, No. 61472145, No. 61572199, and No. 61502173, the grant from the Guangdong Natural Science Funds for Distinguished Young Scholars (No. S2013050014677), the Fundamental Research Funds for the Central Universities (No. 2014G0007 and 2015PT016), the grant from the key lab of cloud computing and big data in Guangzhou (No. SITGZ[2013]268-6), the grant from Science and Technology Planning Project of Guangdong Province, China (No. 2015A050502011).

REFERENCES

[1] Z. Yu, L. Li, J. Liu, J. Zhang, G. Han, "Adaptive Noise Immune Cluster Ensemble Using Affinity Propagation", *IEEE Transactions on Knowledge and Data Engineering*, 2015, DOI: 10.1109/TKDE.2015.2453162.

[2] B.J. Frey, D. Dueck, "Clustering by Passing Messages Between Data Points", *Science*, vol. 315, pp. 972-976, 2007.

[3] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888-905, 2000.