Spectral Aggregation for Clustering Ensemble

Xi Wang
Department of Automation
Tsinghua University
xwang02@mails.thu.edu.cn

Chunyu Yang
Department of Automation
Tsinghua University
yangchunyu@mails.thu.edu.cn

Jie Zhou
Department of Automation
Tsinghua University
jzhou@tsinghua.edu.cn

Abstract

Since a large number of clustering algorithms exist, aggregating different clustered partitions into a single consolidated one to obtain better results has become an important problem. We propose a new algorithm for clustering ensemble based on spectral clustering. We also propose a criteria along with this algorithm, for the detection of cluster numbers. Our algorithm can determine the number of clusters more accurately with less volatility, and therefore can deduce a better combined clustering result. Experimental results on both synthesis and real data-sets show the capability and robustness of our approach.

1. Introduction

Clustering is an important approach of unsupervised learning and large quantities of clustering algorithms exist [4, 7]. However, no single one is versatile to all kinds of data-sets, and they also suffer from some kinds of problems, such as dependence on initializations and volatility of results. Inspired by sensor fusion and classifier combination, clustering ensemble came into birth, for the sake of overcoming these drawbacks and improving the performance. This process can be simply described as below: given a variety of clusterings on a certain data-set, the manipulation will combine the inputs to a single consolidated partition.

Some work has been done in this new field [1, 2, 3, 5, 8, 9]. Generally speaking, the process of clustering ensemble consists of three steps: (1) *Clustering Representation*, establishing models to represent partition labels, (2) *Representation Combination*, combing all these representations by a certain method, and (3) *Combined Representation Repartition*, applying relative algorithms to partition the data-set on the combined representation for clustering ensemble.

Clustering representation is an important step, and 0-1 affinity matrix is a popular method in previous researches. For a clustering on a data-set with n samples, its corresponding affinity matrix is an $n \times n$ symmetrical binary matrix. The value 1 means that corresponding pair are clustered into the same cluster, while 0 denotes that they are divided into two different groups. Mean of all these affinity matrices is defined as *co-association matrix*. Partition algorithms, such as MST in *Evidence Accumulation* (***EA***) [1, 2, 3] and METIS in *Cluster-based Similarity Partitioning Algorithm* (***CSPA***) [8], are then conducted on the co-association matrix for the final combined partition.

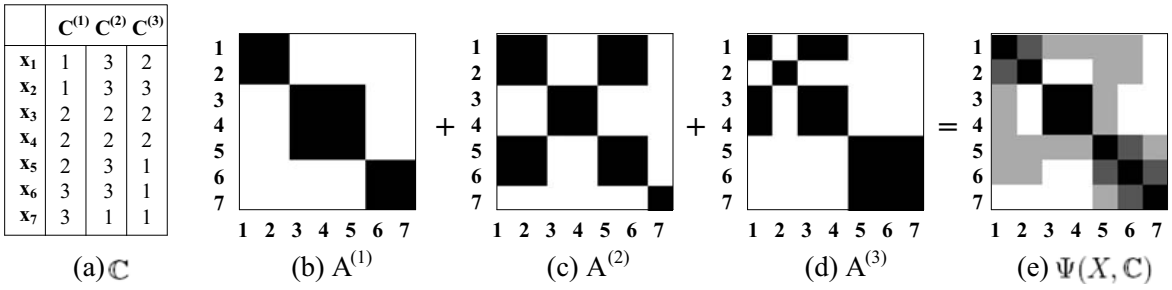
In this paper, we propose a new algorithm for clustering ensemble also based on these 0-1 affinity matrices. Our contributions lie in two aspects: (1) Referring to the characters of the 0-1 affinity matrix, we suggest a new criteria for determining the number of clusters in the combined partition. It works by finding out the dominating eigenvalues of the co-association matrix. (2) Since the co-association matrix measures pairwise correlations of the data-set, the repartition on it complies with the framework of spectral clustering. Thus, we apply spectral clustering to the step of repartition, with the number of clusters determined by new criteria as input. We call this algorithm ***Spectral Aggregation*** (***SA***).

The rest of this paper is structured as follows: In section 2, we describe the proposed *Spectral Aggregation* algorithm for clustering ensemble and the criteria for determining the numbers of clusters. In section 3, experimental results on both synthetic and real data-sets are shown and analyzed. Finally, section 4 gives a concise conclusion.

2 Spectral Aggregation

We begin our discussion about ***SA*** by introducing our notation: Let $X = \{x_1, x_2, \cdots, x_n\}$ be a data-set, with $|X| = n$. $C^{(p)} = \{c_1^{(p)}, c_2^{(p)}, \cdots, c_n^{(p)}\}$ denotes the partition label from the ***p***-th clustering result, where $p = 1, 2, \cdots, d$. There, $c_i^{(p)}$ means that in the ***p***-th

978-1-4244-2175-6/08/$25.00 ©2008 IEEE



	$C^{(1)}$	$C^{(2)}$	$C^{(3)}$
x_1	1	3	2
x_2	1	3	3
x_3	2	2	2
x_4	2	2	2
x_5	2	3	1
x_6	3	3	1
x_7	3	1	1

Figure 1. Example of Overall Procedure for Clustering Aggregation

clustering, the sample x_i is partitioned into the $c_i^{(p)}$-th cluster. As a result, the process of clustering ensemble can be described as follows: given a series of partition labels, i.e., $\mathbb{C} = \{C^{(p)}\}_{p=1}^{d}$, the combined clustering, $P = \{p_1, p_2, \cdots, p_n\}$, will be determined and output.

2.1 Algorithm Procedure

As we've talked previously, the 0-1 affinity matrix is a common approach for clustering representation. There, we set $A^{(p)}$ to denote the affinity matrix created from the clustering $C^{(p)}$. Then we have:

$$A^{(p)}(x_i, x_j) = \begin{cases} 1, & c_i^{(p)} = c_j^{(p)}, \quad (1) \\ 0, & c_i^{(p)} \neq c_j^{(p)}. \quad (1') \end{cases}$$

Mean of all $A^{(p)}$ is marked as $\Psi(X, \mathbb{C})$ and can be regarded as a combination of all these pairwise correlations provided by $\mathbb{C}$:

$$\Psi(X, \mathbb{C}) = \frac{1}{d} \sum_{p=1}^{d} A^{(p)}. \quad (2)$$

Ψ is actually also an affinity matrix, and all its elements are continuous values between 0 and 1. The closer $\Psi(x_i, x_j)$ to 1, the more original clusterings put them into the same cluster, and the stronger bond this pair have, and vice versa. This procedure is demonstrated in Figure 1 by a simple example.

Now, we have $\Psi(X, \mathbb{C})$ describing pairwise correlation of the whole data-set. By repartitioning X on Ψ we can get a new clustering, which can be acknowledged as the combination of all given clusterings. In actual fact, the process of partitioning samples by its affinity matrix conforms to the framework of spectral clustering, and therefore we suggest relevant approaches be used here for clustering ensemble. After getting $\Psi(X, \mathbb{C})$, the subsequent procedure can be carried out as introduced by [6].

2.2 Determination of Cluster Number

There, we set $\{\psi_1, \psi_2, \cdots, \psi_n\}$ to denote all eigenvalues of $\Psi(X, \mathbb{C})$. They are already sorted in descending, i.e., $\psi_1 > \psi_2 > \cdots > \psi_n$. Before running the procedure above to partition $\Psi(X, \mathbb{C})$, the number of clusters to divide X into, i.e. k, should be determined firstly. By analyzing the 0-1 affinity matrix, we can find following characters:

Observation 1: For any $C^{(p)}$ and its corresponding $A^{(p)}$, the number of clusters determined by $C^{(p)}$ is equal to the rank of $A^{(p)}$.

Observation 2: Numbers of samples clusters in $C^{(p)}$ contain respectively equal to eigenvalues of $A^{(p)}$.

Take $C^{(1)}$ and $A^{(1)}$ in Figure 1 for example. The data-set is partitioned into three clusters by $C^{(1)}$, two in cluster 1, three in cluster 2, and other two in cluster 3. Meanwhile, the rank of $A^{(1)}$ is just 3, and three eigenvalues are 2, 3 and 2. These two characters are easy to prove in mathematics. We neglect the proof for short.

These observations reflect that eigenvalues of individual 0-1 affinity matrix contain clustering information, and Ψ is a combination of all these information. We believe that each original clustering contains both real information and noise. By combining them, the dominating information are strengthened while noise is partly offset. As a result, eigenvalues of Ψ is also a joint of all the original clustering information, and can be used to support the clustering ensemble.

In order to analyze the distributions of Ψ's eigenvalues, we pick three data-sets for test: *Iris*, *Image*, and *Optical-Digit*. They are all downloaded from **UCI** Repository, and respectively contain 3, 7, and 10 natural clusters. For each data-set, we run K-means 50 times with the number of clusters, k, randomly picked in the range $[max\{K_{real} - 5, 2\}, K_{real} + 5]$. Then, we combine all these 50 clusterings for Ψ, get its largest 12

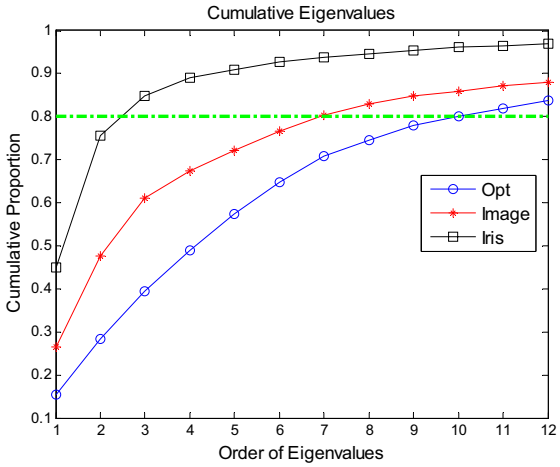


Figure 2. Eigenvalues of Core Matrix

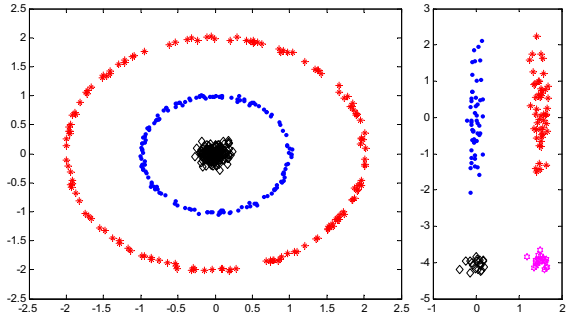


Figure 3. Data-sets for Experiments

eigenvalues, and sort them in descending. Additionally, we divide them by the trace of Ψ for standardization, and calculate the cumulative proportions they dominate orderly. Results are plotted in Figure 2.

The green horizontal line represents the proportion of 0.8, and we can easily find that this threshold can help to find out the largest few eigenvalues corresponding to the natural number of clusters. As demonstrated previously, the largest eigenvalues contain the clustering information while others are just noise. By distinguishing them, the number of clusters can be determined. For this task, we propose a new criteria, which defines the dominating eigenvalues by finding the number of eigenvalues that firstly makes the cumulative proportions surpass a given threshold. This criteria is formulated as follows:

$$k = \min_{s}(\sum_{i=1}^{s}\psi_i > \varepsilon \cdot \sum_{i=1}^{n}\psi_i) \qquad (3)$$

In consequence, the overall procedure of ***SA*** can be summarized as follows: (1) Transfer all given clusterings $C^{(p)}$ into corresponding affinity matrices $A^{(p)}$ by equation 1. (2) Calculate the combined representation $\Psi(X, \mathbb{C})$ by equation 2. (3) Determine the number of clusters by the proposed criteria and apply an algorithm of spectral clustering to $\Psi(X, \mathbb{C})$.

3 Experimental Results

We have conducted extensive experiments to test the quality of ***SA*** on both synthetic and real data-sets. Among all clustering ensemble algorithms, ***EA*** is reported to be the best by the experimental results shown in [3]. In our experiments, we compared the proposed algorithm with ***EA*** used in [3] to illustrate its effectiveness.

In our experiments, we select nine data-sets. *3rings* and *Cigar* are 2D data-sets and shown in Figure 3, while others are from **UCI**. There, the *Ptrain* and *Ptest* are composed of the first 100 samples respectively from the training and testing sets of *Pen-digit*. Dimensions of all these data-sets range between 2 and 77, and real numbers of clusters range between 2 and 10. Besides, their distributions vary in different modes and they are suitable to be used to test and compare the algorithms' capabilities. These data-sets are also similar to those used in previous studies [3].

We design the experiment as follows: for each data-set, we firstly run k-means ten times with given numbers of clusters, K, as initialization. These ten partition labels are then transported to the process of clustering ensemble as inputs. We run ***EA*** with both Single Link (***SL***) and Average Link (***AL***), as well as ***SA*** to aggregate clusterings. This process is repeated for 50 times, and we record their results. The parameter K in K-means is randomly picked in the range $[max\{K_{real}-5, 2\}, K_{real}+5]$, where K_{real} represents the natural number of clusters in the data-set. Besides, the threshold ε for the cluster number determination criteria is set as 0.8.

For each run of clustering ensemble on every data-set, we record the numbers of clusters detected by the three approaches. After 50 cycles, we calculate their means as well as standard deviation, which are shown in Table 1. By referring to the real number of clusters also revealed in this table, we can easily find that our ***SA*** can detect a much closer number to the real number than both ***EA*** with ***SL*** and ***AL***. In addition, the numbers

Table 1. Statistic Results of 50 Runs

Data	Real Num.	Average Numbers			Standard Deviation			Average Error Rates (%)			
		EA-S	EA-A	SA	EA-S	EA-A	SA	K-means	EA-S	EA-A	SA
3rings	3	2.30	3.02	3.80	1.074	1.407	0.535	50.18	39.83	47.69	38.66
Cigar	4	2.88	2.08	3.94	1.118	0.444	0.956	38.48	40.29	48.81	36.06
Iris	3	2.40	2.04	3.00	0.535	0.283	0.700	31.76	32.00	33.13	31.19
Wbc	2	2.78	2.02	2.28	1.266	0.141	0.497	15.83	9.13	3.97	8.27
Opt	10	2.64	7.14	9.70	1.998	2.330	0.953	34.69	73.14	35.20	31.74
Wine	3	2.86	2.02	3.32	1.325	0.141	0.794	38.89	34.85	32.90	31.33
Image	7	5.92	3.40	7.40	3.779	1.654	1.010	51.59	55.14	61.98	49.13
Ptrain	10	3.76	7.66	9.52	3.061	2.752	0.886	36.81	66.78	36.78	36.62
Ptest	10	3.26	6.74	10.12	1.925	2.834	1.062	42.33	63.44	46.60	40.48
Avg.	–	–	–	–	1.787	1.332	0.821	37.84	46.07	38.56	33.72

output by ***SA*** also have averagely smaller standard deviations than other two algorithms, which means that ***SA*** performs more stably.

In addition, since we have the ground labels of all these data-sets, we can calculate the error rates of each combined clustering with the error measurement used in [3]. There, we also calculate means of all 50 runs on each data-set. Besides, we also record the errors of all original clusterings generated by K-means and calculate their means. All these average error rates are shown in Table 1 as well. We can find that ***SA*** outperform other ensemble algorithms, as well as K-means. Statistic results also reveal that, our algorithm performs especially preeminently on data-sets containing more clusters, such as *Opt*, *Image*, *Ptrain* and *Ptest*. Much more accurate numbers of clusters, smaller standard deviations, and much smaller average error rates are shown in corresponding rows in Table 1.

4 Conclusion

In this paper, we have proposed a new algorithm for clustering ensemble called ***Spectral Aggregation***. This algorithm is based on the 0-1 affinity matrix for clustering representation, and then applies spectral clustering to the mean of all these affinity matrices. Besides, in order to determine the number of clusters, we also propose a new criteria by finding this average matrix's dominating eigenvalues on basis of the cumulative proportions they contribute to. Experimental results show that the new criteria can detect the number of clusters more accurately and stably, and our algorithm can obtain better combined clustering results.

References

[1] A. Fred and A. Jain. Data clustering using evidence accumulation. *Proceedings of the 16th International Conference on Pattern Recognition, (ICPR 2002)*, pages 276–280, 2002.

[2] A. Fred and A. Jain. Evidence accumulation clustering based on the k-means algorithm. *Proceedings of the InternationalWorkshop s on Structural and Syntactic Pattern Recognition (SSPR 2002)*, pages 442–451, 2002.

[3] A. Fred and A. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):835–850, June 2005.

[4] A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, September 1999.

[5] L. Kuncheva, S. Hadjitodorov, and L. Todorova. Experimental comparison of cluster ensemble methods. *International Conference on Information Fusion*, 2006.

[6] W. Y. Ng A, Jordan M. On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS)*, pages 849–856, 2001.

[7] R. O.Duda, P. E.Hart, and D. G.Stock. Pattern classification. 2nd ed., New York, America: Jonh Wiley & Sons, 2001.

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

[9] A. Topchy, A. Jain, and W. Punch. A mixture model of clustering ensembles. *Proc. SIAM Intl. Conf. on Data Mining*, 2004.