Ensemble Re-clustering: Refinement of Hard Clustering by Three-Way Strategy

Pingxin Wang[1,4(✉)], Qiang Liu[2], Xibei Yang[2,4], and Fasheng Xu[3,4]

[1] School of Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China
pingxin_wang@hotmail.com
[2] School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212003, China
[3] School of Mathematical Sciences, University of Jinan, Jinan 250022, China
[4] Department of Computer Science, University of Regina, Regina, SK S4S 0A2, Canada

Abstract. In this paper, we propose a three-way ensemble re-clustering method based on ideas of cluster ensemble and three-way decision. In the proposed method, we use hard clustering methods to produce different clustering results and cluster labels matching to align each clustering results to a given order. The intersection of the clusters with same labels are regarded as the core region and the difference between the union and the intersection of the clusters with same labels are regarded as the fringe region of the specific cluster. Therefore, a three-way result of the cluster is naturally formed. The results on UCI data sets show that such strategy is effective in improving the structure of clustering results and F_1 values.

Keywords: Three-way decisions · Three-way clustering · Cluster ensemble · Label matching

1 Introduction

Clustering is one of the most significant unsupervised learning problems which has been used in diverse areas like machine vision and pattern recognition as well as in medical applications. The fundamental objective of data clustering is to group similar objects in one cluster and divide dissimilar objects into different clusters. Research on clustering algorithm has received much attention and a number of clustering methods have been developed over the past decades.

The various methods for clustering can be divided into two categories: hard clustering and soft clustering. Hard clustering methods, such as C-means [11], spectral clustering [4], are based on an assumption that a cluster is represented by a set with a crisp boundary. That is, a data point is either in or not in a specific cluster. The requirement of a sharp boundary leads to easy analytical results, but may not adequately show the fact that a cluster may not have a well-defined cluster boundary. Furthermore, It is not the best way to absolutely

Y. Sun et al. (Eds.): IScIDE 2017, LNCS 10559, pp. 423–430, 2017.
DOI: 10.1007/978-3-319-67777-4_37

divide the boundary objects into one cluster. In such cases, an object in the boundary should belong to more than one cluster.

In order to relax the constraint of hard clustering methods, many soft clustering methods were proposed for different application backgrounds. Fuzzy sets are a well known generalization of crisp sets, first introduced by Zadeh [23]. Incorporating fuzzy sets into C-means clustering, Bezdek [2] proposed Fuzzy C-Means (FCM), which is assumed that a cluster is represented by a fuzzy set that models a gradually changing boundary. Another effective tool for uncertain data analysis is rough set theory [15], which use a pair of exact concepts, called the lower and upper approximations, to approximate a rough (imprecise) concept. Based on the rough set theory, Lingras and West [10] introduced Rough C-Means (RCM) clustering, which describes each cluster not only by a center, but also with a pair of lower and upper bounds. Incorporating membership in the RCM framework, Mitra et al. [13] put forward a Rough-Fuzzy C-Means (RFCM) clustering method. Shadowed set, proposed by Pedrycz [16], provides an alternate mechanism for handling uncertainty. As a conceptual and algorithmic bridge between rough sets and fuzzy sets, shadowed set has been successfully used for clustering analysis, resulting in Shadowed C-Means (SCM) [14]. A brief summary of existing clustering methods can be shown in Fig. 1.

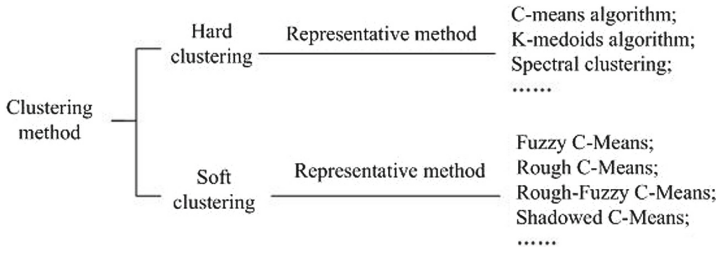

Fig. 1. Classification diagram of existing clustering methods

Although there are a lot of clustering methods, the performance of many clustering algorithms is critically dependent on the characteristics of the data set and the input parameters. Improper input parameters may lead to clusters that deviate from those in the data set. There is not one clustering method that can identify any form of data structure distribution. In order to solve this problem, Strehl and Ghosh [17] proposed cluster Ensemble algorithm, which combines multiple clusterings of a set of objects into one clustering result without accessing the original features of the objects. It has been shown that cluster ensembles are useful in many applications, such as knowledge-reuse [6], multi-view clustering [8], distributed computing [12] and in improving the quality and robustness of clustering results [3,5,7,9].

Recently, three-way decisions for problem solving was proposed by Yao [18, 19], which is an extension of the commonly used binary-decision model by adding

a third option. The approach of three-way decisions divides the universe into the Positive, Negative and Boundary regions which denote the regions of acceptance, rejection and non-commitment for ternary classifications. Specifically, for the objects partially satisfy the classification criteria, it is difficult to directly identify them without uncertainty. Instead of making a binary decision, we use thresholds on the degrees of satisfiability to make one of three decisions: accept, reject, non-commitment. The third option may also be referred to as a deferment decision that requires further judgments. Three-way decisions have been proved to build on solid cognitive foundations and are a class of effective ways commonly used in human problem solving and information processing [20]. Many soft computing models for leaning uncertain concepts, such as interval sets, rough sets, fuzzy sets and shadowed sets, have the tri-partitioning properties and can be reinvestigated within the framework of three-way decisions [19].

Motivated by the three-way strategy, Yu [21,22] proposed a new soft clustering framework, three-way clustering, which uses two regions to represent a cluster, i.e., core region (Co) and fringe region (Fr) rather than one set. The core region is an area where the elements are highly concentrated of a cluster and fringe region is an area where the elements are loosely concentrated. There are maybe common elements in the fringe region among different clusters.

This paper aims at presenting a three-way clustering method by using cluster ensemble and three-way decisions based on the results of hard clustering. In the proposed method, hard clustering methods are used to produce different clustering results and cluster labels matching are used to align each clustering results to a given order. The three-way ensemble re-clustering results are obtained by the following strategy. The intersection of the clusters with same order are regarded as the core region and the difference between the union and the intersection of the clusters with same order are regarded as the fringe region of the specific cluster.

The study is organized into five sections. We start with a briefly introduction of the background knowledge in Sect. 2 and in Sect. 3 we present the process of Ensemble Re-clustering by two main steps. Experiment results are reported in Sect. 4.

2 A Three-Way Ensemble Re-clustering

By following ideas of cluster ensemble and three-way decisions, we present a three-way ensemble re-clustering algorithm. In this section, we assume that the universal has been divided into k disjoint sets m times by existing hard clustering algorithms. We discuss how to design a valid consensus function to obtain a three-way clustering based on the hard clustering results.

We begin our discussion by introducing some notations. We suppose that $V = \{v_1, \cdots, v_n\}$ is a set of n objects and $\mathbb{C}_i, (i = 1, \cdots m)$, denotes i-th clustering of V, where $\mathbb{C}_i = \{C_{i1}, \cdots, C_{ik}\}$ is a hard clustering results of V. Although we have obtained the the clustering results of V, $\mathbb{C}_i$ can not be directly used for the conclusion of the next stage due to the lack of priori category information.

As an example, consider the dataset $V = (v_1, v_2, v_3, v_4, v_5, v_6)$ that consists of six objects, and let $\mathbb{C}_1, \mathbb{C}_2$ and $\mathbb{C}_3$ be three clusterings of V which are shown in Table 1. Although they are expressed in different orders, they represent the same clustering result, so in order to combine the clustering results, the cluster labels must be matched to establish the correspondence between each other.

Table 1. Different ways of the same clustering results

	$\mathbb{C}_1$	$\mathbb{C}_2$	$\mathbb{C}_3$
v_1	1	2	3
v_2	1	2	3
v_3	2	3	2
v_4	2	3	2
v_5	3	1	1
v_6	3	1	1

In general, the number of identical objects covered by the corresponding cluster labels should be the largest, so the cluster labels can be registered based on this heuristic. Assuming that there are two clustering results $\mathbb{C}_1$ and $\mathbb{C}_2$. Each divides the dataset into k clusters, respectively, denoted by $\{C_{11}, \cdots, C_{1k}\}$ and $\{C_{21}, \cdots, C_{2k}\}$. First, the numbers of identical objects covered by each pair of cluster labels C_{1i} and C_{2j} in the two clusters are recorded in the overlap matrix of $k \times k$. And then select the cluster label that covers the largest number of identical objects to establish the correspondence and remove the result from the overlap matrix. Repeat the above process until all the cluster labels have established the corresponding relationship.

When there are $m(m > 2)$ clustering results, we can randomly select one as the matching criterion and match the other clustering results with the selected results. The matching algorithm only needs to check the $m-1$ clustering results and store the overlap matrix with the storage space of $(m-1) \times k^2$. The whole matching process is fast and efficient.

After all clustering labels match, all objects of V can be divided into three types for a given label j based on the results of labels matching:

$$\begin{aligned}
\text{Type I} &= \{v \mid \forall i = 1, \cdots, m, v \in C_{ij}\}, \\
\text{Type II} &= \{v \mid \exists i \neq p, i, p = 1, \cdots, m, v \in C_{ij} \wedge v \notin C_{pj}\}, \\
\text{Type III} &= \{v \mid \forall i = 1, \cdots, m, v \notin C_{ij}\},
\end{aligned}$$

From the above classifications, it can be seen that the objects in Type I are assigned to j-th cluster in all clustering results. The objects of Type II are assigned to j-th cluster in part of clustering results. The objects in Type III have not intersection with j-th in each clustering results. Based on the ideas of three-way decisions and three-way clustering, The elements in Type I are clearly

attributable to the j-th cluster. And should be assigned to core region of j-th cluster. The elements in Type II should be assigned to fringe region of j-th cluster and all the elements in Type III should be assigned to trivial region of j-th cluster. From the above discussion, we get the following strategy to obtain a three-way clustering by cluster ensemble.

$$\mathrm{Co}(C_j) = \{v \mid \forall i = 1, \cdots, m, v \in C_{ij}\} = \bigcap_{i=1}^{m} C_{ij},$$

$$\mathrm{Fr}(C_j) = \{v \mid \exists i \neq p, i, p = 1, \cdots, m, v \in C_{ij} \wedge v \notin C_{pj}\} = \bigcup_{i=1}^{m} C_{ij} - \bigcap_{i=1}^{m} C_{ij}$$

The above clustering method are called a three-way ensemble re-clustering. The procedure of three-way ensemble re-clustering consists mainly of three steps.

1. Obtain a group of hard clustering results $\mathbb{C}_i, (i = 1, \cdots m)$ by using existing methods.
2. Randomly select one clustering result in step 1 as the matching criterion and match the other clustering results with the selected results
3. Compute the intersection of the clusters with same labels and the difference between the union and the intersection of the clusters with same labels.

The above procedure can be depicted by Fig. 2. Finally, we present Algorithm 1, which describes the proposed three-way ensemble re-clustering based on hard clustering results. In Algorithm 1, we choose the first clustering results $\mathbb{C}_1$ as matching criterion and match the other clustering results with $\mathbb{C}_1$ during labels matching.

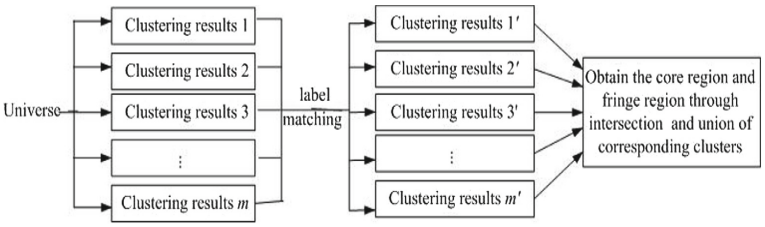

Fig. 2. Procedure diagram of three-way ensemble re-clustering

3 Experimental Illustration

To illustrate the effectiveness of Algorithm 1, some experiments on UCI [1] data sets are employed in this section. Before running Algorithm 1 on date sets, we need to obtain m clustering results. We use NJW spectral clustering with different scale parameter to get 10 clustering results in our experiments. All codes

Algorithm 1. Three-way ensemble re-clustering

Input: m clustering results $\mathbb{C}_1, \cdots, \mathbb{C}_m$
Output: Three-way ensemble re-clustering result
$\mathbb{C} = \{(\mathrm{Co}(C_1), \mathrm{Fr}(C_1)), (\mathrm{Co}(C_2), \mathrm{Fr}(C_2)), \cdots, (\mathrm{Co}(C_k), \mathrm{Fr}(C_k))\}$
for each $\mathbb{C}_i$ in $\{\mathbb{C}_i\}, i = 2, \cdots, m$ **do**
  for $j = 1$ to k, $p = 1$ to k **do**
    overlap (j, p)=Count (C_{ij}, C_{1p});
    //overlap is a $k \times k$ matrix;
    //Count (C_{ij}, C_{1p}) counts the number of same elements of C_{ij} and C_{1p}
  end for
  $\Gamma = \phi$
  while $\Gamma \neq \{C_{i1}, C_{i2}, \cdots, C_{ik}\}$ **do**
    (u, v)=argmax(overlap(j, p)) //(u, v) is the biggest element
    $C_{iu} = C_{iv}$ //align C_{iu} to C_{1v}
    Delete overlap$(u, *)$
    Delete overlap$(*, v)$
    $\Gamma = \Gamma \cup \{C_{iu}\}$
  end while
end for
for $j = 1$ to k **do**
  Calculate $\mathrm{Co}(C_j) = \bigcap_{i=1}^{m} C_{ij}$
  Calculate $\mathrm{Fr}(C_j) = \bigcup_{i=1}^{m} C_{ij} - \bigcap_{i=1}^{m} C_{ij}$
end for
return $\mathbb{C} = \{(\mathrm{Co}(C_1), \mathrm{Fr}(C_1)), (\mathrm{Co}(C_2), \mathrm{Fr}(C_2)), \cdots, (\mathrm{Co}(C_k), \mathrm{Fr}(C_k))\}$

Table 2. A description of 5 data sets

ID	Data sets	Samples	Attributes	Classes
1	Banknote authentication	1372	4	2
2	Congressional voting	435	16	2
3	Hill valley	1212	100	2
4	Ionosphere	351	34	2
5	Vertebral column	310	6	2

are run in Matlab R2013b on a personal computer. The details of these data are shown in Table 2.

In order to measure the tests accuracy, we use Macro F_1 values and Micro F_1 values [24] of cluster results as the evaluation indicator, which are two commonly used methods of testing the effect of classification. The results of above 5 UCI data sets by spectral clustering and three-way ensemble re-clustering are computed respectively. We list the experimental results in Table 3.

With a deep investigation of Table 3, it is not difficult to observe that both the Macro F_1 values and the Micro F_1 values of three-way ensemble re-clustering are

Table 3. Comparisons of clustering results

Data sets	Spectral Micro F_1	Clustering Macro F_1	Ensemble Micro F_1	Re-clustering Macro F_1
Banknote authentication	0.6150	0.6131	**0.6346**	**0.638**
Congressional voting	0.8639	0.8608	**0.9640**	**0.9633**
Hill valley	0.6359	0.6358	**0.6390**	**0.6389**
Ionosphere	0.7012	0.6927	**0.7280**	**0.7550**
Vertebral column	0.6785	0.6690	**0.7487**	**0.7425**

higher than those based on spectral clustering. From Table 3, we can conclude that three-way ensemble re-clustering can significantly improve the structure of classification results by comparing with the traditional spectral clustering algorithm.

4 Concluding Remarks

In this paper, we developed a three-way ensemble re-clustering method by employing the ideas of three-way decisions and cluster ensemble. Hard clustering methods are used to produce different clustering results and cluster labels matching is used to align each clustering results to a given order. The intersection of the clusters with same labels are regarded as the core region and the difference between the union and the intersection of the clusters with same labels are regarded as the fringe region of the specific cluster. Based on the above strategy, a three-way explanation of the cluster is naturally formed and experimental results demonstrate that the new algorithm can significantly improve the structure of classification results by comparing with the traditional clustering algorithm. The present study is the first step for the research of three-way clustering. How to determine the number of clusters is an interesting topic to be addressed for further research.

Acknowledgements. This work was supported in part by National Natural Science Foundation of China (Nos. 61503160 and 61572242), and Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 15KJB110004).

References

1. UCI Machine Learning Repository (2005). http://www.ics.uci.edu/mlearn/MLRepository.html
2. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
3. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional clustering: A cluster ensemble approach. In: Proceedings of the Twentieth International Conference on Machine Learning (2003)

4. Fiedler, M.: Algebraic connectivity of graphs. ICzechoslovak Math. J. **23**, 298–305 (1973)
5. Fred, A., Jain, A.K.: Data clustering using evidence accumulation. In: Proceedings of the Sixteenth International Conference on Pattern Recognition (ICPR), pp. 276–280 (2002)
6. Ghosh, J., Strehl, A., Merugu, S.: A consensus framework for integrating distributed clusterings under limited knowledge sharing. In. In Proceedings of NSF Workshop on Next Generation Data Mining, pp. 99–108 (2002)
7. Hadjitodorov, S., Kuncheva, L., Todorova, L.: Moderate diversity for better cluster ensembles. Inf. Fusion **7**, 264–275 (2006)
8. Kreiger, A.M., Green, P.: A generalized rand-index method for consensus clustering of separate partitions of the same data base. J. Classif. **16**, 63–89 (1999)
9. Kuncheva, L., Hadjitodorov, S.: Using diversity in cluster ensembles. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp. 1214–1219 (2004)
10. Lingras, P., West, C.: Interval set clustering of web users with rough k-means. J. Intell. Inf. Syst. **23**, 5–16 (2004)
11. Macqueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California Press, Berkeley (1967)
12. Merugu, S., Ghosh, J.: Privacy-preserving distributed clustering using generative models. In: Proceedings of The Third IEEE International Conference on Data Mining (ICDM), pp. 211–218 (2003)
13. Mitra, S., Banka, H., Pedrycz, W.: Rough-fuzzy collaborative clustering. IEEE Trans. Syst. Man Cybern. (Part B) **36**, 795–805 (2006)
14. Mitra, S., Pedrycz, W., Barman, B.: Shadowed c-means: integrating fuzzy and rough clustering. Pattern Recogn. **43**, 1282–1291 (2010)
15. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**, 314–356 (1982)
16. Pedrycz, W.: Shadowed sets: representing and processing fuzzy sets. IEEE Trans. Syst. Man Cybern. (Part B) **28**, 103–109 (1998)
17. Strehl, A., Ghosh, J.: Cluster ensembles – a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2002)
18. Yao, Y.: Three-way decisions with probabilistic rough sets. Inf. Sci. **180**, 341–353 (2010)
19. Yao, Y.: An outline of a theory of three-way decisions. In: Yao, J.T., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) RSCTC 2012. LNCS (LNAI), vol. 7413, pp. 1–17. Springer, Heidelberg (2012). doi:10.1007/978-3-642-32115-3_1
20. Yao, Y.: Three-way decisions and cognitive computing. Cogn. Comput. **8**, 543–554 (2016)
21. Yu, H., Jiao, P., Yao, Y., Wang, G.: Detecting and refining overlapping regions in complex networks with three-way decisions. Inf. Sci. **373**, 21–41 (2016)
22. Yu, H., Liu, Z., Wang, G.: A tree-based incremental overlapping clustering method using the three-way decision theory. Knowl.-Based Syst. **91**, 189–203 (2016)
23. Zadeh, L.: Fuzzy sets. Inform. Control **8**, 338–353 (1965)
24. Zhou, Z.: Machine Learning. Tsinghua University Press, Beijing (2016)