Link-based Cluster Ensemble Method for Improved Meta-Clustering Algorithm

Changlong Shao[1], Shifei Ding[1,2,*]

[1](School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)
[2](Mine Digitization Engineering Research Center of Ministry of Education of the People's Republic of China, Xuzhou 221116, China)
Corresponding author: Shifei Ding, E-mail: dingsf@cumt.edu.cn
Tel. +86 051683885189

Abstract: Ensemble clustering has become a hot research field in intelligent information processing and machine learning. Although significant progress has been made in recent years, there are still two challenging issues in the current ensemble clustering research. First of all, most ensemble clustering algorithms tend to explore similarity at the level of object but lack the ability to explore information at the level of cluster. Secondly, many ensemble clustering algorithms only focus on the direct relationship, while ignoring the indirect relationship between clusters. In order to solve these two problems, a link-based meta-clustering algorithm (L-MCLA) have been proposed in this paper. A series of experiment results demonstrate that the proposed algorithm not only produces better clustering effect but is also less influenced by different ensemble sizes.

Keywords: Inter-cluster Similarity, Ensemble Clustering, Clustering, Connected Triple, Meta-clustering Algorithm (MCLA)

1. Introduction

In the field of intelligent information processing and machine learning, clustering analysis is an important learning tool for unlabeled data. Generally speaking, clustering is to classify a given dataset into clusters, so that the data objects within the cluster have larger similarity, while the data objects between clusters are quite different from each other[1]. Clustering has been used in various fields, such as image processing[2], cognitive computing[3], time series analysis[20] and medical diagnosis[17]. In the past few decades, a large number of clustering algorithms have been developed, among which the most representative ones are partitional clustering[18], hierarchical clustering[19], spectral clustering[45], density clustering[67], adaptive clustering[89] and semi-supervised clustering[121]. Nevertheless, there are still some problems in the current clustering algorithm. For instance, the clustering result largely depends on parameters and initialization without which the clustering result is not robust enough. In order to solve these problems, ensemble clustering was proposed by researchers.

Unlike the traditional method of using an algorithm to generate a single clustering result, ensemble clustering is a process of ensemble multiple different clustering results to generate better clustering result. Due to the effectiveness of ensemble clustering algorithm, more researchers have been attracted and proposed many related algorithms. Despite significant advances in ensemble clustering research, most algorithms only focus on direct connection, while ignoring indirect connection between clusters. As shown in Fig 1, two objects appear in the same cluster and thus we say that they are directly connected. However, like (b) and (c), two objects are in two different clusters but we cannot conclude that there is no connection between them because they are likely to be related to each other indirectly. Such indirect connection information may affect the consensus result. In order to explore indirect connection information, we propose a link-based meta-clustering Algorithm (L-MCLA) in this paper.

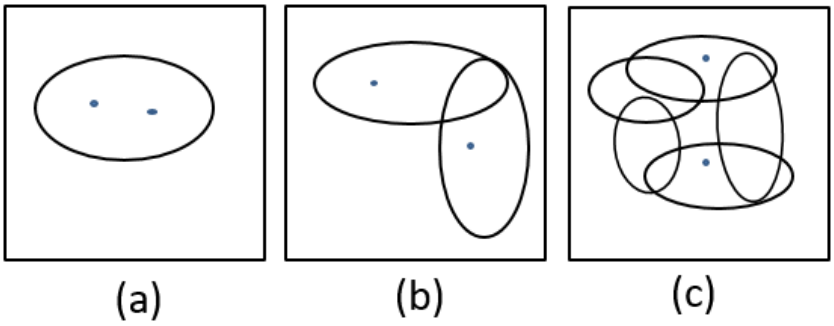

Fig1. Relationship between two points. (a) in the same cluster. (b) belong to two clusters with common parts. (c) belong to two unrelated clusters, but both of them are related to the third cluster.

The remainder of paper is organized as follows. Section 2 reviews the background of this study. Section 3 details the proposed method in this paper. Section 4 shows the experiment results. Section 5 concludes this paper.

2. Background

2.1 Ensemble Clustering

Ensemble clustering is an algorithm to improve the clustering effect by ensemble multiple base clusterings, which can be generally expressed as follows:

Let $X=\{x_1,x_2,...,x_n\}$ denotes a dataset with n objects. We use clustering algorithms to obtain m clustering results $P=\{p_1,p_2,...p_m\}$ and call them as base clusterings. Each base clustering contains several clusters, which is written as $p_i=\{C_i^1,C_i^2,...C_i^j\}$, where j is the number of clusters in the base clustering p_i. Ensemble clustering is to merge the set P through the consensus function T to obtain the final clustering result P*. The specific process of ensemble clustering is shown in Fig 2.

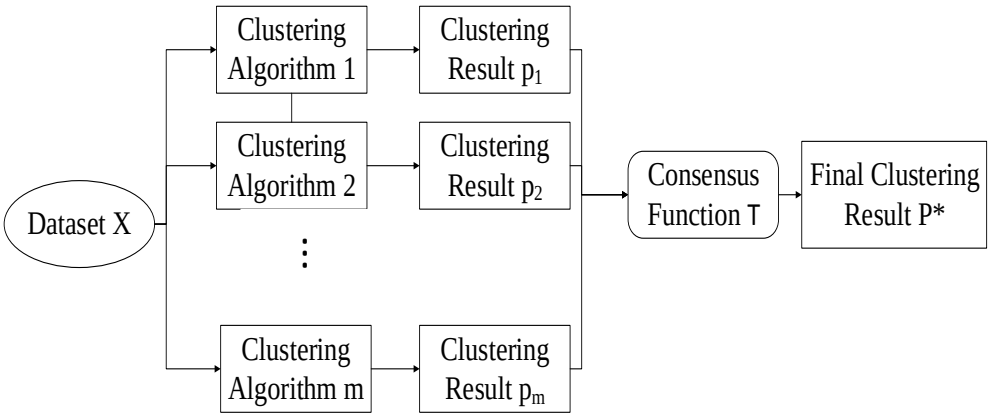

Fig 2. Ensemble clustering process diagram

2.2 Meta-Clustering Algorithm

Meta-clustering algorithm (MCLA) is proposed by Strehl and Ghosh which is an ensemble clustering algorithm working on the level of cluster. Jaccard coefficient is used to calculate similarity between clusters. The jaccard coefficient between cluster C_i and C_j can be calculated as follows:

$$J(C_i,C_j)=\frac{|C_i \bigcap C_j|}{|C_i \bigcup C_j|} \quad (1)$$

Where $\bigcap$ denotes the intersection of two sets, $\bigcup$ denotes the union of two sets, and |*| denotes the number of objects in a set.

Specifically, the meta-clustering algorithm consists of the following four steps:

1) Construct a similarity matrix by calculating jaccard coefficient between clusters contained in base clusterings.

2) Regard the similarity matrix of the previous step as an undirected graph, which is called meta-graph.

3) Use graph partitioning package METIS[16] to divide the meta-graph of the previous step to obtain the meta-cluster and each meta-cluster contains several clusters.

4) Assign each object to the corresponding meta-cluster to get the final clustering result.

3. Link-Based Meta-Clustering Algorithm

3.1 Construct similarity matrix

The meta-clustering algorithm is superior, but it still has a shortcoming.The similarity matrix constructed by Jaccard coefficient can only reflect the direct relationship between clusters while lacking capability to find the indirect relationship. In 2011, the concept of weighted connected-triple (WCT) was proposed by lam-on et al[12], which makes it possible to explore the hidden indirect relationship between clusters.

In this section, connected triple is used to construct a refined cluster similarity matrix. The connected triple is shown in Fig 3.

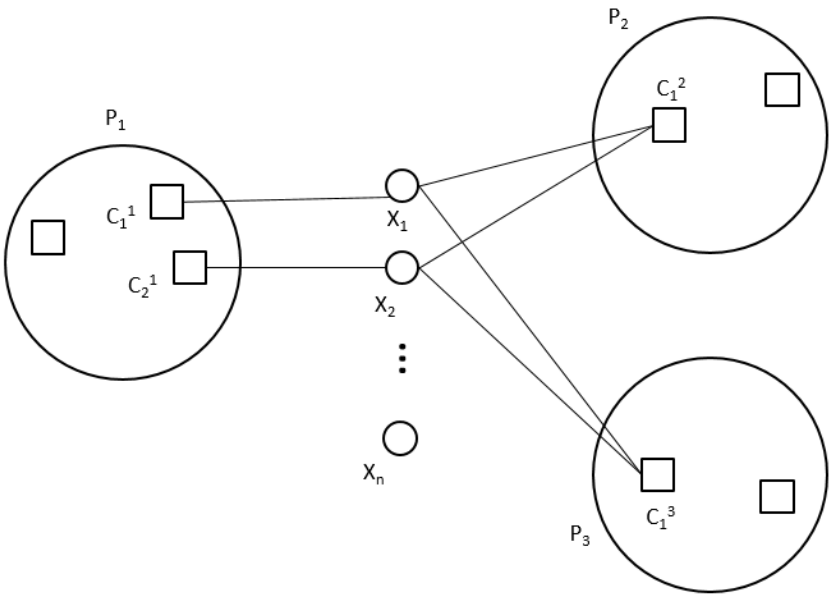

Fig 3. Connected triple diagram

P_1, P_2, P_3 are three base clusterings. C_1^1 and C_2^1 are unrelated clusters (ie, have no public part) and in common sense should have no similarity. C_1^1 and C_1^2 have a common point x_1, and C_2^1 and C_1^2 have a common point x_2. Therefore, C_1^1 and C_1^2 are similar and C_2^1 and C_1^2 are similar.Because both C_1^1 and C_2^1 have a similar third-cluster C_1^2, C_1^1 and C_2^1 are indirectly connected to each other. In the same way, C_1^1 and C_1^3 have a common point x_1 while C_2^1 and C_1^3 have a common point x_2. Accordingly, C_1^1 and C_2^1 have indirectly connection to each other. It can be seen that connected triple can help find more connection between clusters, which is beneficial for reaching a consensus result later.

Firstly, similarity matrix Z is constructed by jaccard coefficient.

$$Z(i,j)=\begin{cases} 1, & i=j \\ J(C_i,C_j) & \text{otherwise} \end{cases} \tag{2}$$

Let C_k have similarity with C_i and C_j, then the weighted connected triple between C_i and C_j is defined as follow:

$$WCT_{ij}^{k} = \min(J(C_i,C_k),J(C_j,C_k)) \tag{3}$$

Then the indirectly connection between cluster C_i and C_j is calculated as follows:

$$Sim^{WCT}(i,j)=\frac{\sum_{C_k \in C} WCT_{ij}^{k}}{\mathrm{WCT}_{\max}} \tag{4}$$

For any two clusters C_i and C_j, their indirectly similarity is defined as:

$$Sim(i,j) = Sim^{WCT}(i,j)\times DC \tag{5}$$

Where DC is a constant decay factor. That is the confidence level of accepting two nonidentical clusters as being similar.

The refine similarity matrix S is constructed as:

$$S(i,j) = \begin{cases} 1, & i=j \\ Sim(i,j)+Z(i,j) & \text{otherwise} \end{cases} \tag{6}$$

3.2 graph division and object allocation

We regard the refined similarity matrix S as the adjacency matrix of graph G. Graph segmentation algorithm is our consensus function. In the selection of graph segmentation algorithm, since the normalized cut (Ncut) is effective and robust, we select it in this study[13]. Normalized cut is a kind of spectral clustering. The basic idea is to define a cut criterion, which considers the total dissimilarity between different clusters and the total similarity within the cluster.

By normalized cut, K meta-cluster groups can be obtained, that is:

$$MC = \{MC_1, MC_2, ..., MC_K\} \tag{7}$$

Here we use the voting method to assign objects. For given object x_i, x_i belong to zero or more clusters in MC_j. Specifically, the voting score of x_i for the meta-cluster MC_j can be defined as follow:

$$\begin{aligned} Score(x_i, MC_j) &= \frac{1}{|MC_j|}\sum_{C_h \in MC_j} l \\ l &= \begin{cases} 1, & if\ x_i \in C_h \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{8}$$

Where $|MC_j|$ denotes the number of clusters in MC_j.

We assign the point x_i to the meta-cluster with the highest score. The final clustering result can be obtained by this way.

For clarity, the algorithm of L-MCLA is described in Algorithm 1.

Algorithm 1: Link-Based Meta-Clustering Algorithm.

Input: Dataset X, number of clusters K

1) Using clustering algorithm to generate m base clusterings $P = \{p_1, p_2, \ldots p_m\}$ for dataset X.
2) The inter-cluster similarity matrix Z is constructed by jaccard coefficients, which can be calculated by the equation (2).
3) For similarity matrix Z, equation (3)-(6) is used to obtain the refined similarity matrix S.
4) The similarity matrix S is regarded as a graph G. K meta-clusters are obtained by using Ncut algorithm to segment this graph as equation (7).
5) The clustering result Label is obtained by allocating object to corresponding meta-cluster by equation (8).

Output: Final cluster result Label

4. Experiments

In this section, we conduct experiments on multiple real-world datasets and compare results with several existing ensemble clustering algorithms to evaluate the performance of the algorithm proposed in this paper. Moreover, the robustness of the algorithm is evidenced by experiment on different ensemble sizes.

4.1 Datasets and evaluation measures

In our experiments, nine datasets in the UCI (University of California Irvine) machine learning database are used as experimental datasets[22]. Table 1 lists the detail of each dataset:

Table1. Description of UCI datasets

Datasets	Object	Dimension	Class
Aggregation	788	2	7
Cardiotocograph(CTG)	2126	21	10
Diabetes	768	8	2
Ecoli	336	8	8
Ionosphere	351	34	2
Segmentation	2130	19	7
Soybean	47	35	4
Thyroid	215	5	3
Yeast	1484	8	10

In our experiments, adjusted Rand index (ARI) and normalized mutual information (NMI) are selected to evaluate the performance of the clustering results. The two evaluation are described as follows:

ARI is a clustering evaluation index that measures the similarity between two clustering results by calculating the number of sample point pairs in the same cluster and different clusters. The equation is as follow:

$$\mathrm{ARI} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \tag{9}$$

Where a denotes the number of point pairs that belong to the same cluster in both real and experimental, b denotes the number of point pairs that belong to the same cluster in real label and different clusters in experimental result, and c denotes the number of point pairs that belong to the same cluster in the experimental result and different clusters in the real label, and d represents the number of point pairs that belong to different clusters in both real and experimental. Its value range [-1,1]. The larger the value is, the more consistent it is with the real result, namely the better clustering effect.

NMI is a common external evaluation index of clustering. It evaluates the similarity of two clustering results from the perspective of communication theory. Let the experimental result be X and the real label be Y, then the equation is as follows:

$$\mathrm{NMI}(\mathrm{X},Y) = \frac{\mathrm{I}(\mathrm{X},\mathrm{Y})}{\sqrt{H(X)H(Y)}} \tag{10}$$

Where I(X, Y) represents the mutual information between X and Y, and H(X) and H(Y) represent the entropy of X and Y. Its value range [0,1]. The larger value indicates more shared information with the real label, that is, the better clustering result.

4.2 Comparative methods and experimental settings

In our experiments, seven ensemble clustering algorithms are compared with L-MCLA algorithm. The seven comparison algorithms are as follows:

1) Evidence accumulation clustering(EAC)[10]

2) Hybrid bipartite graph formulation(HBGF)[14]

3) Weighted evidence accumulation clustering(WEAC)[15]

4) Graph partitioning with multi-granularity link analysis(GP-MGLA)[15]

5) Cluster-based similarity partitioning algorithm(CSPA)[11]

6) Hypergraph partitioning algorithm(HGPA)[11]

7) Meta-clustering algorithm(MCLA)[11]

The experiments are implemented in MATLAB R2016a. The PC configuration is as follows: Windows7 64-bit, Intel i5 1.7GHz CPU, 8G RAM.

In our experiments, k-means is used to generate base clusterings with the parameter k randomly selected in the range $\left[2,\sqrt{\mathrm{N}}\right]$. For parameter DC, high DC values (i.e., 0.7 to 0.9) bring about a data partition of exceptionally good quality[12], so we set DC=0.9 in our experiment. We call the number of base clusterings m as ensemble size and set ensemble size m=50 to compare the L-MCLA algorithm with other ensemble clustering algorithms. Furthermore, we change the ensemble size to test the robustness of L-MCLA algorithm.

4.3 Comparison with other ensemble clustering methods

This section make a comparison experiment of our algorithm. Each ensemble clustering algorithm runs 20 times on each dataset and each run randomly generates base clustering according to section 4.2. The average scores and standard deviation of ARI and NMI are recorded. The experimental results are shown in Table 2 and Table 3, with the highest score shown in bold.

Table 2. Average ARI scores by different ensemble clustering methods. The highest score in each comparison is in bold

	EAC	HBGF	WEAC	GP-MGLA	CSPA	HGPA	MCLA	L-MCLA
Aggregation	0.804(±0.044）	0.805(±0.050)	0.806(±0.045)	0.860(±0.082)	0.549(±0.006)	0.621(±0.022)	0.612(±0.027)	**0.925(±0.057)**
CTG	0.133(±0.007)	0.130(±0.008)	0.129(±0.008)	0.137(±0.004)	0.115(±0.003)	0.117(±0.010)	0.120(±0.004)	**0.140(±0.005)**
Diabetes	0.051(±0.027)	0.044(±0.028)	0.018(±0.013)	0.007(±0.011)	-0.001(±0.001)	-0.001(±0.001)	-0.001(±0.001)	**0.061(±0.032)**
Ecoli	0.495(±0.	0.416(±0	0.468(±0	0.471(±0	0.300(±0.	0.333(±0	0.367(±0	**0.542(±0**

	069)	.015)	.059)	.049)	013)	.026)	.030)	**.053)**
Ionosphere	0.150(±0.011)	0.167(±0.010)	0.148(±0.010)	0.162(±0.013)	0.124(±0.000)	0.020(±0.023)	0.163(±0.014)	**0.169(±0.004)**
Segmentation	0.347(±0.075)	0.425(±0.054)	0.404(±0.059)	0.442(±0.031)	0.384(±0.034)	0.255(±0.036)	0.436(±0.074)	**0.447(±0.039)**
Soybean	0.551(±0.006)	0.547(±0.005)	0.553(±0.009)	0.547(±0.004)	0.483(±0.029)	0.543(±0.013)	0.552(±0.008)	**0.568(±0.029)**
Thyroid	0.370(±0.182)	0.378(±0.178)	0.497(±0.053)	0.558(±0.031)	0.129(±0.037)	0.104(±0.025)	0.271(±0.070)	**0.587(±0.036)**
Yeast	0.173(±0.018)	0.148(±0.005)	0.173(±0.015)	0.167(±0.004)	0.109(±0.013)	0.093(±0.016)	0.125(±0.006)	**0.182(±0.007)**

Table 3. Average NMI scores by different ensemble clustering methods. The highest score in each comparison is in bold

	EAC	HBGF	WEAC	GP-MGLA	CSPA	HGPA	MCLA	L-MCLA
Aggregation	0.868(±0.027)	0.867(±0.031)	0.868(±0.027)	0.900(±0.047)	0.688(±0.008)	0.748(±0.014)	0.746(±0.015)	**0.932(±0.037)**
CTG	0.267(±0.007)	0.273(±0.011)	0.264(±0.008)	**0.274(±0.006)**	0.244(±0.005)	0.246(±0.013)	0.253(±0.005)	0.274(±0.006)
Diabetes	0.018(±0.010)	0.017(±0.012)	0.007(±0.004)	0.003(±0.004)	0.001(±0.000)	0.001(±0.000)	0.001(±0.000)	**0.025(±0.013)**
Ecoli	0.579(±0.027)	0.537(±0.012)	0.568(±0.025)	0.567(±0.025)	0.442(±0.010)	0.465(±0.024)	0.504(±0.020)	**0.608(±0.021)**
Ionosphere	0.115(±0.007)	**0.126(±0.005)**	0.114(±0.006)	0.123(±0.007)	0.102(±0.000)	0.018(±0.018)	0.120(±0.009)	0.123(±0.004)
Segmentation	0.480(±0.066)	0.529(±0.052)	0.523(±0.047)	**0.562(±0.024)**	0.502(±0.026)	0.374(±0.040)	0.545(±0.055)	0.545(±0.029)
Soybean	0.714(±0.003)	0.712(±0.002)	0.714(±0.005)	0.711(±0.002)	0.619(±0.023)	0.706(±0.017)	0.714(±0.004)	**0.721(±0.010)**
Thyroid	0.316(±0.075)	0.325(±0.104)	0.369(±0.026)	0.411(±0.036)	0.164(±0.022)	0.153(±0.011)	0.213(±0.031)	**0.484(±0.035)**
Yeast	0.264(±0.012)	0.257(±0.006)	0.267(±0.011)	0.273(±0.005)	0.206(±0.015)	0.191(±0.020)	0.226(±0.005)	**0.290(±0.007)**

As shown in Table 2, the ARI score of L-MCLA algorithm on 9 datasets are all the highest. As can be seen from Table 3, L-MCLA algorithm on six datasets has the highest NMI value, which is only slightly inferior on the CTG, Ionosphere and Segmentation, but the difference is not significant. To summarize, the L-MCLA method exhibits overall better performance (with respect to ARI, NMI) than the other methods.

4.4 Robustness to ensemble size

In this section, we evaluate the performance of L-MCLA algorithm under different ensemble size on nine datasets. Ensemble size is in the range of [10,100], increasing by 10. The generation settings for base clustering are same as section 4.2. Then we record the average score of ARI and NMI. The change of score is shown in Fig 4 and Fig 5.

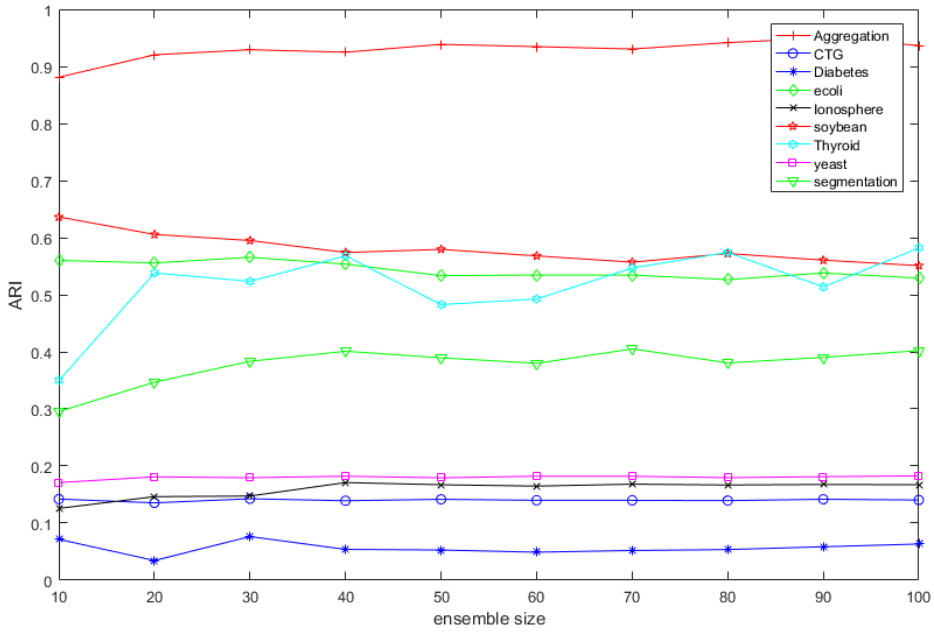

Fig 4. Average ARI scores of L-MCLA under different ensemble size

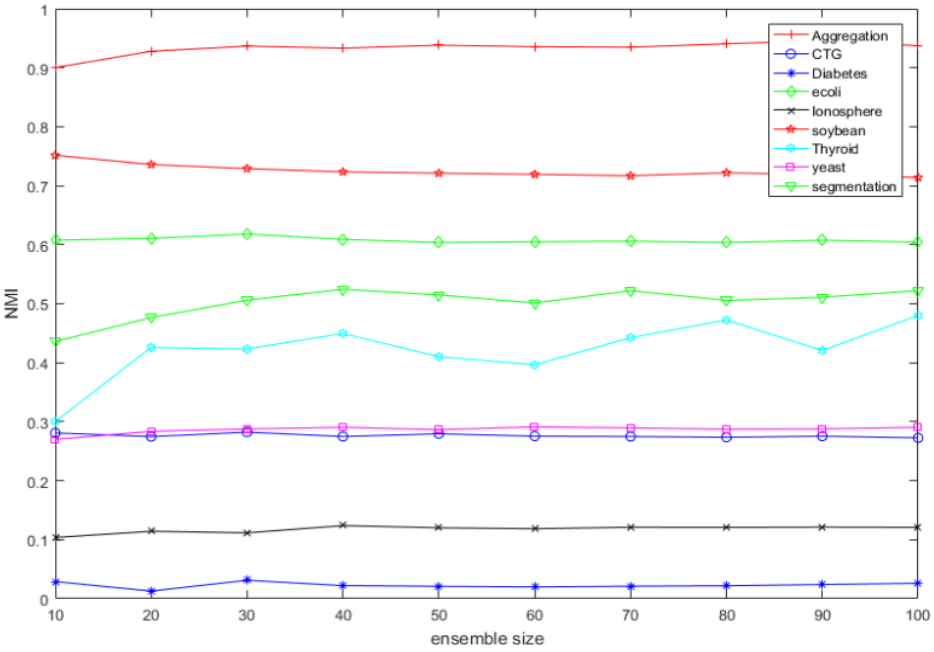

Fig 5. Average NMI scores of L-MCLA under different ensemble size

Fig 4 shows the ARI index values of the L-MCLA algorithm in 9 datasets under different ensemble sizes. It can be seen from the observation that there is only a slight fluctuation in the L-MCLA algorithm on most datasets and the fluctuation gradually decreases after the integration scale reaches 40. In addition to the Thyroid dataset and the Soybean dataset, the ARI value of the Thyroid data set increases sharply from 10 to 20, and then shows a state of fluctuating increase with the increase of ensemble sizes. With the increase of base clusterings, the ARI value of the Soybean dataset shows a slow decline but the range is small.

Fig 5 shows the NMI index values of the L-MCLA algorithm on 9 datasets under different ensemble sizes. It can be seen from observation that the NMI values of most datasets tend to be stable except for the Thyroid datasets. However, the NMI value of the Thyroid dataset shows a significant upward trend

when the base clusterings are about 10-40, but slightly decreases when the base clusterings is 40-50, and then shows an upward trend and gradually stabilizes.

According to the above experimental analysis, the ensemble size has little influence on L-MCLA algorithm. On most datasets, L-MCLA algorithm relys on fewer base clusterings to obtain more robust results.

5. Conclusion

Ensemble clustering is the use of multiple clustering results to generate better clustering results. However, the existing ensemble clustering algorithms often only pay attention to the direct inter-cluster connection and ignore the indirect connection. In this paper, we propose a link-based meta-clustering algorithm which uses connected triple to explore indirect connection. Link-based method is used to enrich similarity matrix for generating better results. Our algorithm has the following advantages: 1. This algorithm considers the information from the cluster level and the object level. 2. It use the link-based method to explore the indirect connection between clusters. A series of experiments proved the advantages of our algorithm. Our future work is to further explore the hidden information in the base clustering and improve the clustering results in this way.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No.61672522 and No. 61976216.

References

1. Ding S, Jia H, Du M, et al. A semi-supervised approximate spectral clustering algorithm based on HMRF model. Information Sciences, 2018, 429: 215-228.
2. Cong L, Ding S, Wang L, et al. Image segmentation algorithm based on superpixel clustering. IET Image Processing, 2018, 12(11):2030-2035.
3. Saini N, Saha S, Bhattacharyya P. Automatic Scientific Document Clustering Using Self-organized Multi-objective Differential Evolution. Cognitive Computation, 2018: 1-23. (DOI: 10.1007/s12559-018-9611-8)
4. Ding S, Cong L, Hu Q, et al. A multiway p-spectral clustering algorithm. Knowledge-Based Systems, 2019, 164: 371-377.
5. Løkse S, Bianchi F M, Salberg A B, et al. Spectral clustering using pckid–a probabilistic cluster kernel for incomplete data//Scandinavian Conference on Image Analysis. Springer, Cham, 2017: 431-442.
6. Liu R, Wang H, Yu X. Shared-nearest-neighbor-based clustering by fast search and find of density peaks. INFORMATION SCIENCES, 2018. (DOI: 10.1016/j.ins.2018.03.031)
7. Du M, Ding S, Xue Y, et al. A novel density peaks clustering with sensitivity of local density and density-adaptive metric. Knowledge and Information Systems, 2019, 59(2): 285-309.
8. Fan S, Ding S, Xue Y. Self-adaptive kernel K-means algorithm based on the shuffled frog leaping algorithm. Soft Computing, 2018, 22(3): 861-872.
9. Ding S, Xu X, Fan S, et al. Locally adaptive multiple kernel k-means algorithm based on shared nearest neighbors. Soft Computing, 2018, 22(14): 4573-4583.
10. Fred A L N, Jain A K. Combining Multiple Clusterings Using Evidence Accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6):835-850.
11. Strehl A, Ghosh J. Cluster ensembles---a knowledge reuse framework for combining multiple partitions. Journal of machine learning research, 2003, 3(Dec): 583-617.
12. Iam-On N, Boongoen T, Garrett S M, et al. A Link-Based Approach to the Cluster Ensemble Problem. IEEE Transactions on Software Engineering, 2011, 33(12):2396-2409.
13. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.
14. Fern X Z, Brodley C E. Solving cluster ensemble problems by bipartite graph partitioning//Proceedings of the twenty-first international conference on Machine learning. ACM, 2004: 36.(DOI:10.1145/1015330.1015414)
15. Huang D, Lai J H, Wang C D. Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis. Neurocomputing, 2015, 170: 240-250.
16. Karypis G, Kumar V. A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs. SIAM Journal on Scientific Computing, 1998, 20(1):359-0.
17. Thanh N D, Ali M. A novel clustering algorithm in a neutrosophic recommender system for medical diagnosis. Cognitive Computation, 2017, 9(4): 526-544.
18. Nguyen B, De Baets B. Kernel-Based Distance Metric Learning for Supervised k-Means Clustering. IEEE Transactions on Neural Networks and Learning Systems, 2019:1-12. (DOI:10.1109/TNNLS.2018.2890021)

19. Cohen-Addad V, Kanade V, Mallmann-Trenn F, et al. Hierarchical clustering: Objective functions and algorithms. Journal of the ACM (JACM), 2019, 66(4): 26.
20. Mikalsen K Ø, Bianchi F M, Soguero-Ruiz C, et al. Time series cluster kernel for learning similarities between multivariate time series with missing data. Pattern Recognition, 2018, 76: 569-581.
21. Zhang H, Lu J. SCTWC: An online semi-supervised clustering approach to topical web crawlers. Applied Soft Computing, 2010, 10(2):490-495.
22. A. Asuncion, D.J. Newman, Uci Machine Learning Repository, http://www.ics. uci.edu/mlearn/MLRepository.html, 2007.