Link-based Pairwise Similarity Matrix for Fuzzy C-means Clustering Ensemble

Pan Su, Changjing Shang, Qiang Shen

Department of Computer Science
Aberystwyth University
Aberystwyth, Wales, UK, SY23 3DB
E-mail: {pas23, cns, qqs}@aber.ac.uk

***Abstract*—Cluster ensemble is an effective way which aggregates multiple clustering results in order to improving robustness and stability. This technique can also help improving accuracy by combing clustering results with different parameters (such as number of clusters) together, rather than carefully selecting the values of them in a single clustering process. Since founded, many topics within cluster ensemble research have been proposed and promising results are gained. These include the generation of ensemble members, consensus of ensemble members and so on. In this paper, link-based consensus methods for the ensemble of fuzzy c-means are proposed. In different with traditional clustering techniques, the clusters which are generated by fuzzy c-means are fuzzy sets, and hence the proposed methods employ a fuzzy graph $< \{\widetilde{C}_1, ..., \widetilde{C}_n\}, \widetilde{L} >$ to represent the relations between base-clusters and generate the final ensembled clustering results based on it. With various benchmark datasets, the proposed methods are tested against the traditional methods and the experimental results show that the proposed fuzzy-link-based clustering ensemble methods are better then its counterparts in terms of accuracy.**

***Index Terms*—Clustering Ensemble, Fuzzy C-means, Link-based Similarity Matrix**

I. Introduction

Clustering is one of the important approaches within the framework of unsupervised learning which is used for finding the hidden structure of unlabelled data sets. In general, the task of clustering is to assign objects to groups (namely clusters) such that objects in the same group are similar to each other, and dissimilar to those in the other clusters [1]. It has been successfully applied to several important problem domains of computational intelligence such as machine learning, pattern recognition [2], [3], bioinformatics [4], [5] and so on. A number of clustering algorithms and their successful applications have been proposed in the literature. However, each algorithm has its own properties and limitations, and there is no single clustering algorithm is able to discover all types of cluster structures for all data sets [6]. For a given set of data, different algorithms, or even the same algorithm with different parameters (such as the number of clusters), usually provide distinct solutions [7], and hence an inexperienced user runs the risk of picking an inappropriate clustering method. Also, in unsupervised learning, there is usually no ground truth against which the result can be matched. Therefore it is extremely difficult for users to decide which algorithm would be the better one [8].

To overcome these limitations and improve the robustness as well as the quality of using single clustering, cluster ensemble methods have emerged as effective solutions which combine results of various clustering algorithms in different ways. One of the main objectives of the combination is to achieve accuracy superior to those of individual clustering [7]. Although combining multiple partitions of a set of objects into a single consolidated clustering is not necessarily dependent on accessing the features or algorithms that determined these partitions [9], it has empirically verified that the performance of cluster ensembles depends on both the quality and the diversity of ensemble members [6], [8]. Therefore, two essential steps are involved in cluster ensemble: the generation of base-clustering members and the consensus of them.

A number of papers have been published that have helped to develop these fields. In order to generate base-clustering members with dissimilar results, different parameter configurations of one clustering algorithm are tested in [10], [11]; re-sampling techniques are also applied to diverse base clusters [12], [13], [14]. For the consensus methods, there are also several methods such as: feature-based approach where each base-clustering member provides cluster labels as new features describing data points, which is utilised to formulate the final solution [15], [16]; pairwise similarity approach which creates a matrix, containing the pairwise similarity among data points, then any similarity-based clustering algorithm (such as the hierarchical clustering) can be applied [10]; graph-based approach which makes use of the graph representation to solve the cluster ensemble problem [9], [17].

Although much effect has been made in the development of cluster ensemble, modelling a mechanism that is effective for integrating multiple data partitions in a cluster ensemble is far from trivial, and the practice of cluster ensembles is still in its early stage [18]. Most of the cluster ensemble methods found in the literature are based on crisp base clusterings, except in [19] where the problem of aggregate "soft" base-clustering members is defined. In this paper, link-based consensus methods for the ensemble of fuzzy c-means are proposed. In different with ensembles of crisp clusters, the proposed method is able to handling fuzzy base-clusters. In different with the link-

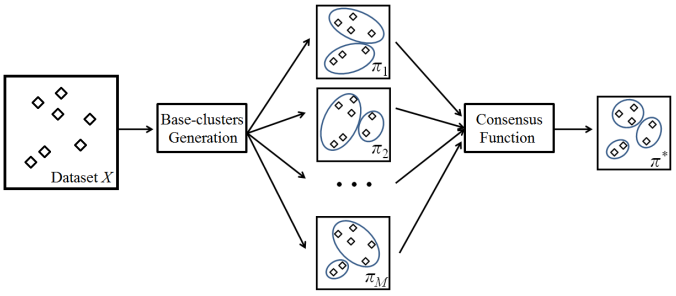


Fig. 1. Cluster Ensemble

based crisp cluster ensemble [7], [20], the proposed method employs a fuzzy graph $< \{\widetilde{C}_1, ..., \widetilde{C}_n\}, \widetilde{L} >$ to represent the relations between base-clusters and refine the pairwise similarity matrix for ensemble. With several UCI benchmark datasets [21], the proposed methods are tested against its crisp counterpart and the fuzzy co-association matrix without link-based refinement. The experimental results show that the proposed fuzzy link-based clustering ensemble methods are better then its counterparts in terms of accuracy.

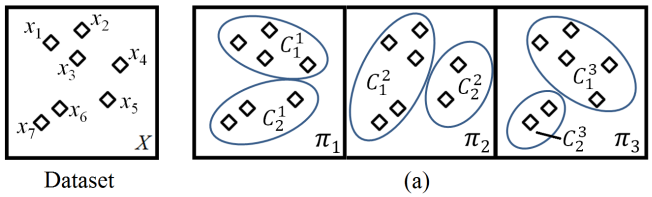
The remainder of this paper is organised as follows. Section II introduces the basics of cluster ensemble. Section III describes the definition of fuzzy co-association matrix and link-based pairwise similarity matrices and their applications to agglomerative clustering to fulfill ensemble of fuzzy clusters. Section IV presents the experimental evaluation of the proposed approach, along with a discussion of the results. Finally, Section V concludes the paper with some suggestions for further development.

II. Preliminaries

A. Cluster Ensemble Problem

Formally, the cluster ensemble problem can be described as follows: let $X = \{x_1, \cdots, x_N\}$ be a set of N data points and let $\Pi = \{\pi_1, \cdots, \pi_m, \cdots, \pi_M\}$ be M base-clustering results (or base-clustering members). Each base-clustering result returns a set of clusters $\pi_m = \{C_1^m, \cdots, C_k^m, \cdots, C_{K_m}^m\}$ which satisfies that $\bigcup_{k=1}^{K_m} C_k^m = X$, where K_m is the number of clusters in the m-th clustering result and C_k^m is a base-cluster. For each $x_i \in X$ and each base-clustering result $\pi_m \in \Pi$, $C^m(x_i) \in \pi_m$ denotes the cluster label to which the object x_i belongs in π_m. The task of cluster ensemble is to find a new clustering result π^* of the data set X that summarises the information from the cluster ensemble Π. Two key procedures are involved in the clustering ensemble. Base clustering results are firstly generated by artificial diversifying methods such as: different parameters and re-sampling. After that, a consensus function is then applied on those base clustering results to generate the final clustering result. The procedure of cluster ensemble is illustrated in Figure 1.

A consensus function is a map which is from a set of base-clustering results to one final partition of the original data $f : \Pi \rightarrow \pi$. Once the base-clusters are generated from the



	π_1	π_2	π_3
x_1	C_1^1	C_1^2	C_1^3
x_2	C_1^1	C_1^2	C_1^3
x_3	C_1^1	C_1^2	C_1^3
x_4	C_1^1	C_2^2	C_1^3
x_5	C_2^1	C_2^2	C_1^3
x_6	C_2^1	C_1^2	C_2^3
x_7	C_2^1	C_1^2	C_2^3

(b)

	C_1^1	C_2^1	C_1^2	C_2^2	C_1^3	C_2^3
x_1	1	0	1	0	1	0
x_2	1	0	1	0	1	0
x_3	1	0	1	0	1	0
x_4	1	0	0	1	1	0
x_5	0	1	0	1	1	0
x_6	0	1	1	0	0	1
x_7	0	1	1	0	0	1

(c)

Fig. 2. Examples of Ensemble-information Matrices

data, a variety of consensus functions are available to derive the final data partition. Most of the consensus functions utilise an ensemble-information matrix which summaries the base-clustering members. Given the ensemble of Fig. 2(a), two types of such a matrix: the label-assignment matrix and the binary cluster-association matrix are illustrated in Figures 2(a) and (b) respectively. The binary cluster-association matrix provides a cluster-specific view of the original label-assignment matrix. If crisp clustering algorithm such as k-means is used in base-clusters generation, the association degree of a data point belonging to a specific cluster is either 1 or 0. Usually, a categorical data clustering algorithm is further applied to this type of ensemble-information matrix to achieve the final partition of the original data. Another type of techniques represents an ensemble as a graph, where the nodes are base-clusters or data points and links between them represents the relations defined between them. Graph partition methods are then applied on the graph to get clustering ensemble output [17].

B. Pairwise Similarity Matrix for Cluster Ensemble

Besides the consensus functions mentioned above, pairwise similarity matrix is another type of consensus methods which transfer the relation between data point and base-cluster to relation amongst data points. Take the co-association (CO) matrix [10] as an example: each base-clustering result $\pi_m \in \Pi$ is able to transfer into a $N \times N$ similarity matrix by using Eqn. (1), denoted as $S_m, m = 1, \cdots, M$:

$$S_m(x_i, x_j) = \begin{cases} 1, & \text{if } C^m(x_i) = C^m(x_j) \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Having obtained all the M similarity matrices of base-clustering results $S_m, m = 1, \cdots, M$, they are merged to form the co-association matrix by using Equation (2). The CO matrix entries represent the similarity between data points x_i and x_j:

$$CO(x_i, x_j) = \frac{1}{M}\sum_{m=1}^{M} S_m(x_i, x_j). \quad (2)$$

Many pairwise similarity based clustering algorithms can be applied to the CO matrix. The agglomerative clustering is employed to derive the final partitions in [10]. The main drawback of crisp CO matrix is that many entries of it are zeros, which means that the according two data points are assigned to different clusters by all base-clustering members. Investigations revealed that the zero-similarity values can be as much as 75% in some UCI datasets [20]. Intuitively, this particular characteristic is commonly encountered with the ensemble of crisp clustering results, and it may limit the quality of a data partition generated by any consensus function [7]. In order to refine the sparse ensemble-information matrices, link-based refining methods are proposed for crisp cluster ensemble problems. In this paper, the fuzzy c-means are employed to generate base-clustering results and the CO matrix for fuzzy c-means ensemble, the FCO, is proposed. To further improve the quality of the proposed FCO matrix, two link-based methods $FLink$ and $FCTS$ are also designed for its refinement.

III. Pairwise Similarity Matrices for Fuzzy C-means Clustering Ensemble

A. FCO: Co-association Matrix for Fuzzy C-means Ensemble

Fuzzy c-means is an effective method to generate a fuzzy partition of a given data. Each cluster in a partition $\widetilde{\pi}_m$ is a fuzzy set $\widetilde{C}_k^m, k = 1, \cdots, K_m$ where $\widetilde{C}_k^m(x_i) \in [0, 1]$ represents the degree of a data point $x_i \in X$ belongs to the according fuzzy cluster. Usually, this degree is normalised with all the clusters in a partition to satisfy that $\sum_{k=1}^{K_m} \widetilde{C}_k^m(x_i) = 1$. In order to keep the consistence with crisp cluster ensemble, the similarity measure of two objects $x_i, x_j \in X$ within each base-clustering result, $S_{\widetilde{m}}(x_i, x_j)$, and the according FCO matrix are defined in Eqn. (3) and Eqn. (4) respectively:

$$S_{\widetilde{m}}(x_i, x_j) = \sum_{k=1}^{K_m}(\widetilde{C}_k^m(x_i) \wedge \widetilde{C}_k^m(x_j)) \quad (3)$$

$$FCO(x_i, x_j) = \frac{1}{M}\sum_{m=1}^{M} S_{\widetilde{m}}(x_i, x_j). \quad (4)$$

Since $\sum_{k=1}^{K_m} \widetilde{C}_k^m(x_i)$ is normalised to 1, so that $S_{\widetilde{m}}(x_i, x_j) \in [0, 1]$ and $FCO(x_i, x_j)) \in [0, 1]$. It worth notice that Eqn. (3) is a general version of Equation (1). If the degree of a data point belongs to a crisp cluster is represented as $\widetilde{C}_k^m(x_i) \in \{0, 1\}$, then Eqn. (3) can also be applied to crisp cluster ensemble equivalently.

One of the advantages of the fuzzy c-means is that most of the data points have non-zero memberships to all clusters. This feature is very helpful for cluster ensemble to keep more details of base-clustering members in the pairwise similarity matrix. Even two data points which are not assigned in same cluster in crisp clustering can also have non-zero values in the FCO matrix defined in Equation (4).

B. $FLink$: Link-based Pairwise Similarity Matrix for Fuzzy C-means Ensemble

Except of certain re-sampling methods, base-clustering members are usually generated from the same dataset, and hence the resulting base-clusters in a cluster ensemble may share common data points. Those shared data points build the linkage among base-clusters and it is possible to estimate the similarity of base-cluster pair by using the underlying link information. A graph based on a set of base-clusters and a set of weighted links between them are also defined in [20]. Given a cluster ensemble defined in section II-A, a graph $< V, L >$ can be constructed where $V = \bigcup_{m=1}^{M} \pi_m = \{C_1, \cdots, C_n\}, n = \sum_{m=1}^{M} K_m$ is the set of vertices each representing a base-cluster, and L is a set of weighted links between clusters. The weighted links between base-clusters C_i, C_j is defined as:

$$w(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \quad (5)$$

where $|*|$ indicates the number of members in a set. However, in crisp cluster ensemble, base-clusters within same base-clustering member do not have common data points with each other, e.g., $\forall C_k^m, C_l^m \in \pi^m$, if $k \neq l$ then $C_k^m \cap C_l^m = \emptyset$. Therefore, the weights of links between those clusters within the same base-clustering member are of zero-values, and hence extra refinement is necessary to them before they are used in ensemble. In order to maintain more information from base-clustering results and refine the FCO matrix for fuzzy c-means ensemble, a fuzzy graph of fuzzy c-means ensemble is proposed.

Formally, given a set of fuzzy base-clusters $C = \{\widetilde{C}_1, \cdots, \widetilde{C}_n\}$ on a dataset $\{x_1, \cdots, x_N\}$, a fuzzy graph $< C, \widetilde{L} >$ is defined on the set of fuzzy base-clusters where $\widetilde{L}$ is a fuzzy set of links defined on $C \times C$. The membership of a link $(\widetilde{C}_i, \widetilde{C}_j), i, j = 1, \cdots, n$ to the fuzzy set $\widetilde{L}$ is defined as:

$$\widetilde{L}(\widetilde{C}_i, \widetilde{C}_j) = \frac{\sum_{t=1}^{N}(\widetilde{C}_i(x_t) \wedge \widetilde{C}_j(x_t))}{\sum_{t=1}^{N}(\widetilde{C}_i(x_t) \vee \widetilde{C}_j(x_t))} \quad (6)$$

where $\widetilde{C}_i(x_t)$ indicates the the degree of a data point x_t belongs to a fuzzy cluster $\widetilde{C}_i$, and it is obvious that $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_j) \in [0, 1]$, $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_i) = 1$ and $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_j) = \widetilde{L}(\widetilde{C}_j, \widetilde{C}_i)$. The degree assigned to the link connecting fuzzy clusters $\widetilde{C}_i$ and $\widetilde{C}_j$ is defined in accordance with the proportion of their overlapping degree on all data points in X. One of the advantages of the proposed fuzzy graph is that even for two fuzzy base-clusters within the same base-clustering member, the degree of the link between them is possible to be an non-zero value. Therefore each base-cluster can has a link to all the other base-clusters, and the fuzzy degree of the link represents the similarity/weight between the according two base-clusters.

Having obtained this fuzzy graph, the link-based pairwise similarity matrix of data points can be defined by using the fuzzy links between them. Specifically, for the a clustering member $\widetilde{\pi}_m$, the link-based similarity of data points x_i and x_j is estimated by:

$$LS_{\widetilde{m}}(x_i, x_j) = \begin{cases} 1, \text{if } i = j \\ L(\arg \widetilde{C}^m_{\max}(x_i), \arg \widetilde{C}^m_{\max}(x_j)) \times \\ (\widetilde{C}^m_{\max}(x_i) \wedge \widetilde{C}^m_{\max}(x_j)), \text{otherwise} \end{cases} \quad (7)$$

where $\widetilde{C}^m_{\max}(x_i) = \bigvee_{k=1}^{K_m} \widetilde{C}^m_k(x_i)$ and $\arg \widetilde{C}^m_{\max}(x_i) \in \pi_m$ represents the fuzzy cluster in which x_i obtained its maximum membership. If a draw situation happens, a random one is picked. The similarity of two data points in the overall fuzzy c-means clustering members is defined as: $FLink(x_i, x_j) = \sum_{m=1}^{M} FS_{\widetilde{m}}(x_i, x_j)/M$.

In different with the FCO, the link-based based similarity defined in Eqn. (7) only associates a data point x_i to the cluster with which x_i has the maximum degree. If two data points obtained their maximum degrees in the same cluster, then their similarity values assigned by $LS_{\widetilde{m}}$ equals the smaller degree value of the two assigned with the cluster, since $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_i) = 1$. Otherwise, the link-based similarity of two data points x_i and x_j is defined as the the smaller value of their maximum degree times the degree of the link between those two base-clusters where x_i and x_j obtained their maximum degree values respectively. It worth noticing that, though the advantage of $\widetilde{L}$ is that non-zero links not only exists between base-clusters within a base-clustering member, e.g., $\exists \widetilde{L}(\widetilde{C}^m_k, \widetilde{C}^m_l) > 0$, but also exists between base-clusters cross base-clustering members, e.g., $\exists \widetilde{L}(\widetilde{C}^m_k, \widetilde{C}^n_l) > 0, m \neq n$. Since the definition of $LS_{\widetilde{m}}$ does not employ links cross base-clustering members, the computing time and memory space of $\widetilde{L}(\widetilde{C}^m_k, \widetilde{C}^n_l), m \neq n$ can be saved in the algorithm implementation. However, in crisp cluster ensemble, the links cross base-clustering members are employed to imply the similarity within base-clustering members by using link-base methods such as the connected-triple, and the quality of final ensemble result is improved after this type of process. In order to test whether the cross links can help to refine $FLink(x_i, x_j)$ further more and maintain the consistence with link-based crisp cluster ensemble, the connected-triple is also applied to $\widetilde{L}$ in the following subsection.

C. FCTS: Connected-triple-based Pairwise Similarity Matrix for Fuzzy C-means Ensemble

The connected-triple approach has been used in a bibliographic dataset which has rich links between data points [22]. It assumes that if two nodes are both connected to a third node then this is indicative of similarity between those two nodes. The connected-triple is also applied to the weighted crisp cluster ensemble graph $< V, L >$ described in Eqn. (5) to generate the similarity of nodes within clustering members [20]. Specifically, the weighted connected-triple regards the similarity of two base-clusters C_i and C_j as the sum their minimum weight to all common neighbour:

$$w'(C_i, C_j) = \sum_{t=1}^{n} (w(C_i, C_t) \wedge w(C_j, C_t)) \quad (8)$$

where $n = \sum_{m=1}^{M} K_m$ represents the total number of base-clusters of all base-clustering results. The $w'(C_i, C_j)$ is also normalised as: $n\_w'(C_i, C_j) = w'(C_i, C_j)/w'_{\max}$, where $w'_{\max}$ is the maximum $w'(C_i, C_j)$ value of any two base-clusters C_i and C_j. Having obtained this, the similarity of two data points x_i and x_j with base-clustering member C^m is defined as:

$$S'_m(x_i, x_j) = \begin{cases} 1, \text{ if } C^m(x_i) = C^m(x_j) \\ n\_w'(C^m(x_i), C^m(x_j)) \times DC, \text{otherwise} \end{cases} \quad (9)$$

where $DC \in [0, 1]$ is a constant decay factor. The connected-triple-based similarity matrix for base-clusters is defined the same as Eqn. (2): $CTS(x_i, x_j) = \sum_{m=1}^{M} S'_m(x_i, x_j)/M$.

Accordingly, the fuzzy version of CTS is also defined similarly, where $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_j)$ is refined by using connected-triple as $L'(\widetilde{C}_i, \widetilde{C}_j) = \sum_{t=1}^{n} \widetilde{L}(\widetilde{C}_i, \widetilde{C}_t) \wedge \widetilde{L}(\widetilde{C}_j, \widetilde{C}_t)$ and normalised to $n\_L'(\widetilde{C}_i, \widetilde{C}_j) = L'(\widetilde{C}_i, \widetilde{C}_j)/L'_{\max}$, where $L'_{\max}$ is the maximum $L'(\widetilde{C}_i, \widetilde{C}_j)$ value of any two fuzzy base-clusters $\widetilde{C}_i$ and $\widetilde{C}_j$. Therefore, the similarity of two data points x_i and x_j with base-clustering member $\widetilde{C}^m$ is modified as:

$$LS'_{\widetilde{m}}(x_i, x_j) = \begin{cases} 1, \text{if } i = j \\ L'(\arg \widetilde{C}^m_{\max}(x_i), \arg \widetilde{C}^m_{\max}(x_j)) \times \\ (\widetilde{C}^m_{\max}(x_i) \wedge \widetilde{C}^m_{\max}(x_j)), \text{otherwise} \end{cases} \quad (10)$$

where $\widetilde{C}^m_{\max}(x_i) = \bigvee_{k=1}^{K_m} \widetilde{C}^m_k(x_i)$ and $\arg \widetilde{C}^m_{\max}(x_i) \in \pi_m$ represents the fuzzy cluster of which x_i obtained its maximum membership. If a draw situation happens, a random one is picked. The similarity of two data points in the overall fuzzy c-means clustering members is defined as: $FCTS(x_i, x_j) = \sum_{m=1}^{M} FS'_{\widetilde{m}}(x_i, x_j)/M$.

D. Steps of Link-based Fuzzy C-means Ensemble

The complete process of using the proposed metrics in cluster ensembles is similar to that of other pairwise similarity matrices. Two main steps are needed, which are:

1) Fuzzy c-means are used on the dataset X for M times to generate fuzzy base-clusters. The diversity of base-clustering results can be gained by using re-sample of original datasets, different clusters number and different initial centroids for fuzzy c-means. Many other methods which used in crisp cluster ensemble can also be used in fuzzy c-means ensemble.
2) All the three proposed methods (FCO, $FLink$, $FCTS$) can be used to generate a pairwise similarity matrix of data points based on the information provided by base-clustering members. After that, a pairwise similarity based clustering algorithm such as hierarchical

TABLE I
A SUMMARY OF THE DATASETS USED

Datasets	Instances	Attributes	Classes
Iris	150	4	3
Wine	178	13	3
Parkinsons	195	22	2
Glass (Identification)	214	9	6
Ecoli	336	7	8
Ionosphere	351	34	2
(Pima Indians) Diabetes	768	8	2

clustering can be used to generate the final partition of the dataset as the output of cluster ensemble.

IV. EXPERIMENT AND EVALUATION

This section presents an experimental evaluation of the proposed work. It shows the set-up of the experiments carried out and also discusses the results obtained. One experiment is designed to test the trend of accuracy when the diversity of base-clustering members is changed, and the other experiment is designed to compare the performance of proposed methods.

A. Experimental Set-up

To evaluate the performance of proposed methods, they are experimentally tested over seven datasets obtained from UCI benchmark repository [21], where true labels of instances are known but are not explicitly used in the cluster ensemble process. The details of the used datasets are summarised in Table I. Since the labels are available for all the seven datasets, the final results of cluster ensemble are evaluated by accuracy.

The fuzzy c-means clustering algorithm is specifically used to generate the base clustering members. Thirty clustering-members are generated ($M = 30$) and the cluster centroids are randomly initialised in each run of ensemble. Two agglomerative clustering approaches (complete-linkage and average-linkage) are selected to achieve the consensus function. These consensus functions finally divide data points into clusters by using the underlying similarity matrix $FLink$, FCO, $FCTS$, or CTS. For comparison purpose, the numbers of final clusters on each dataset in set to the number of its true classes and the decay factor (DC) of CTS is set to 0.5 [20] in the following experiments. To make the qualities of base-clusters equally amongst different ensemble methods, the base-clustering results used in CTS are defuzzified from the base fuzzy c-means used in the other three fuzzy methods.

B. Results and Discussion

In the first experiment, the sensitivity of the proposed methods to the diversity of base-clustering members is tested. In order to change the diversity of base-clustering members, the maximum number of base-clusters $\max(K_m)$ in each test is set from 3 to 30 with the increment of 3. The number of base-clusters in each clustering-member K_m is randomly choose from $[3, \max(K_m)]$. Figure 3 shows the change of accuracies with the increase of diversity in base-clustering members when agglomerative clustering with average-linkage is used in consensus function. Each point in Fig. 3 is an average value of 50 runs.

Generally, the accuracies of three methods ($FLink$, $FCTS$ and CTS) are increased along with the increase of diversity in five of the seven datasets. This indicates that when using the link-based pairwise similarity metrics in fuzzy c-means ensemble, brings in more difference in base-clustering members will generate better results. The results of FCO seems more stable compared with the link-based methods, which indicates that the FCO is not sensitive to the number of clusters in each base-clustering member. An intuitive explanation is that in fuzzy c-means, each data points has gained the memberships to all the clusters, it helps the base-clustering members which have smaller number of clusters can keep as much information as the ones of lager cluster numbers. However, the accuracies of FCO are not as high as the other link-based methods in general, which shows that though fuzzy c-means can help FCO to keep more information for ensemble, the link-based refinements are still helpful to generate more efficient pairwise similarity metrics.

To further analyse the results achievable by the link-based methods, another experiment is carried out by using fixed number ($K_m = \lceil\sqrt{N}\rceil$) and random number ($K_m \in [3, \lceil\sqrt{N}\rceil]$) of clusters in each base-clustering member. The resultant accuracies and are shown in Table 3 and Table 4 respectively, where the best-2 results on each dataset is highlighted in boldface and each number in these tables is an average value based on 50 runs. To validate the significance of the experiment results, the paired-t tests are carried out between $FLink$ and the rest on each dataset. The sign "(-)" in these tables indicates that the corresponding result is significantly ($p < 0.05$) worse than that of $FLink$, while "(*)" indicates one is significantly better than that of $FLink$. In each "pair" of results, the generation of base-clustering members are based on the same number of clusters and same initialisation centroids.

The results show that for both fixed and random K_m, the link-based pairwise similarity matrix $FLink$ achieved the best average accuracy over the seven datasets when applied to the fuzzy c-means ensemble. This indicates that the proposed fuzzy graph proposed in Section III-B is helpful to build a better ensemble-information matrix than FCO. However, the performance of the $FCTS$ is not significantly better than $FLink$ in general. This implies that the connected-tripe method does not further refine the $FLink$ effectively. Note that both $FLink$ and $FCTS$ are achieved better accuracy than CTS on most of the datasets. Although the CTS employed connected-triple to infer the similarities amongst clusters within each base-clustering member, it seems that the inferred similarities are not as effective as those generated by the set of fuzzy links $\widetilde{L}$ in $FLink$ and $FCTS$. Particulary, the $FLink$ can use the fuzzy links $\widetilde{L}(\widetilde{C}_k^m, \widetilde{C}_l^m)$ where $k, l = 1, \cdots, K_m$ directly without infer them from $\widetilde{L}(\widetilde{C}_k^m, \widetilde{C}_l^n), m \neq n$, the time for running connected-triple or other similar refinement is saved. In conclusion, the $FLink$ achieved both higher accuracy and less time-consuming than the CTS.

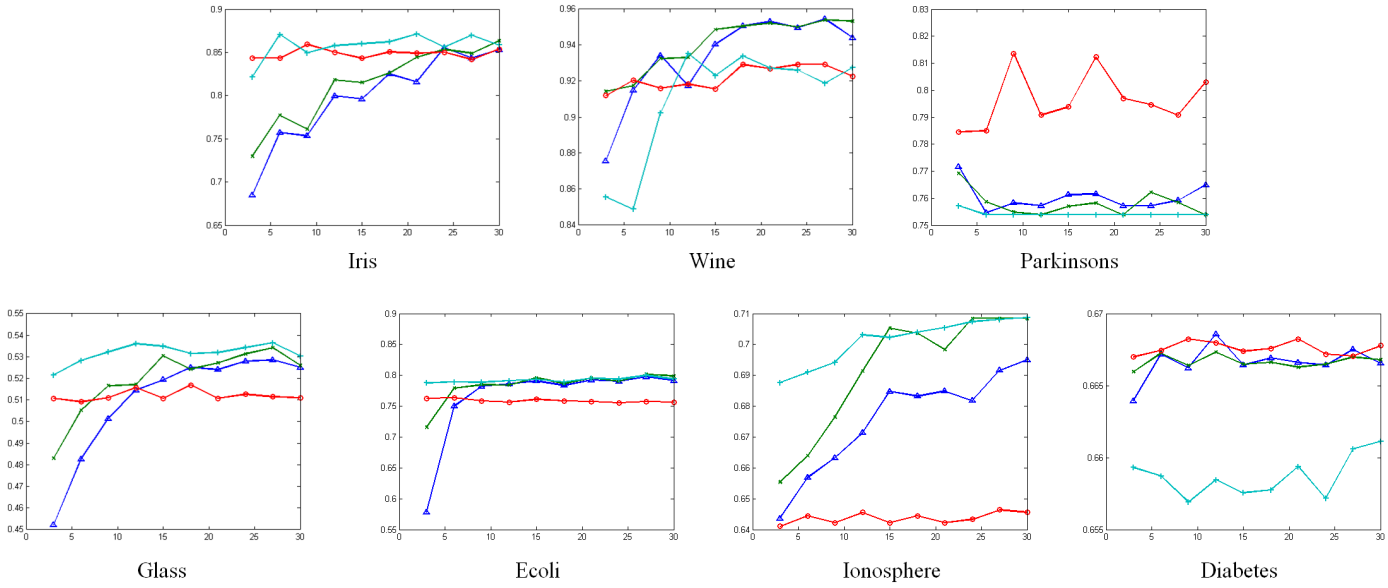


Fig. 3. Trend of Accuracy Change Against Diversity (FCO: –o–, $FLink$: –x–, $FCTS$: –△–, CTS: –+–)

TABLE II
COMPARISON OF ACCURACY - FIXED CLUSTER NUMBER

	Complete-link				Average-link			
	$FLink$	FCO	$FCTS$	CTS	$FLink$	FCO	$FCTS$	CTS
Iris	**86.36**	**87.60**(*)	80.97(-)	71.35(-)	**77.53**	**80.91**(*)	67.20(-)	76.80
Wine	**94.51**	91.58(-)	**94.45**	80.75(-)	**94.45**	90.31(-)	**94.44**	71.99(-)
Parkinsons	**81.92**	81.54(-)	**81.92**	76.18(-)	**81.92**	75.38(-)	**82.05**	80.58(-)
Glass	48.25	45.37(-)	**48.31**	**52.60**(*)	**51.31**	47.79(-)	49.28(-)	**57.34**(*)
Ecoli	**79.53**	76.15(-)	**79.90**(*)	75.29(-)	**82.86**	64.99(-)	**83.58**(*)	77.29(-)
Ionosphere	64.10	64.10	64.10	64.10	64.10	64.10	64.10	64.10
Diabetes	**66.64**	**66.87**(*)	66.63	65.82(-)	66.65	**66.91**(*)	**66.66**	65.65(-)
Means	**74.4728**	73.3157	**73.7547**	69.4409	**74.1176**	70.0552	**72.4729**	70.5357

TABLE III
COMPARISON OF ACCURACY - RANDOM CLUSTER NUMBER

	Complete-link				Average-link			
	$FLink$	FCO	$FCTS$	CTS	$FLink$	FCO	$FCTS$	CTS
Iris	**86.21**	85.52	**85.73**	85.51	**85.12**	85.03	84.81	**86.15**
Wine	**95.16**	91.40(-)	**95.13**	86.58(-)	**95.33**	93.02(-)	**95.38**	91.72(-)
Parkinsons	75.84	**76.27**	**76.28**(*)	75.38(-)	75.84	**79.57**(*)	**76.45**	75.38(-)
Glass	**52.83**	51.20(-)	52.53	**52.74**	**52.98**	51.33(-)	52.84	**53.53**(*)
Ecoli	**78.98**	77.11(-)	**79.33**	76.92(-)	**79.24**	75.89(-)	**78.95**	78.86
Ionosphere	**68.43**	68.17	67.35	**70.51**(*)	**70.83**	64.39(-)	69.38(-)	**70.79**
Diabetes	66.71	66.63	**66.73**	**66.92**(*)	66.68	**66.74**	**66.71**	65.94(-)
Means	**74.8800**	73.7571	**74.7257**	73.5086	**75.1457**	73.7100	**74.9314**	74.6243

V. CONCLUSION

This paper has presented the co-association matrix and two link-based pairwise similarity matrices for fuzzy c-means cluster ensemble. The proposed matrices takes the advantage of fuzzy c-means that each data point can has memberships to all clusters. A set of fuzzy links between base-clusters $\widetilde{L}$ is defined and a fuzzy graph $< C, \widetilde{L} >$ is employed to generate the link-based similarity matrices. Experimental results on seven UCI datasets indicate that the proposed method are generally better than the CTS. Furthermore, the link-based methods also help to build better pairwise similarity matrices

compared with the non-link based matrix FCO.

Whilst promising, the presented work also opens up an avenue for further investigation. For instance, many other base-clustering member generating methods such as re-sampling may also be applied. It would be useful to investigate the performance of the proposed fuzzy graph with different consensus functions. It is also interesting to use methods based on fuzzy graph theory rather than the connected-triple to combined with the proposed fuzzy graph, as they may be more suitable and efficient to handle fuzzy graphs.

REFERENCES

[1] A. Jain, M. Murty, and P. Flynn, "Data clustering: a review," *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[2] E. Diday, G. Govaert, Y. Lechevallier, and J. Sidi, "Clustering in pattern recognition," in *Digital Image Processing*. Springer, 1981, pp. 19–58.

[3] A. Baraldi and P. Blonda, "A survey of fuzzy clustering algorithms for pattern recognition. i," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 29, no. 6, pp. 778–785, 1999.

[4] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[5] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of computational biology*, vol. 6, no. 3-4, pp. 281–297, 1999.

[6] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," in *Systems, man and cybernetics, 2004 IEEE international conference on*, vol. 2. IEEE, 2004, pp. 1214–1219.

[7] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 12, pp. 2396–2409, 2011.

[8] S. T. Hadjitodorov, L. I. Kuncheva, and L. P. Todorova, "Moderate diversity for better cluster ensembles," *Information Fusion*, vol. 7, no. 3, pp. 264–275, 2006.

[9] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *The Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2003.

[10] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 6, pp. 835–850, 2005.

[11] A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 4, 2007.

[12] C. Domeniconi and M. Al-Razgan, "Weighted cluster ensembles: Methods and analysis," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, no. 4, p. 17, 2009.

[13] Z. Yu, H.-S. Wong, and H. Wang, "Graph-based consensus clustering for class discovery from gene expression data," *Bioinformatics*, vol. 23, no. 21, pp. 2888–2896, 2007.

[14] S. Y. Kim and J. W. Lee, "Ensemble clustering method based on the resampling similarity measure for gene expression data," *Statistical methods in medical research*, vol. 16, no. 6, pp. 539–564, 2007.

[15] N. Nguyen and R. Caruana, "Consensus clusterings," in *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*. IEEE, 2007, pp. 607–612.

[16] A. Topchy, A. K. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 12, pp. 1866–1881, 2005.

[17] X. Z. Fern and C. E. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 36.

[18] N. Iam-On and T. Boongoen, "Comparative study of matrix refinement approaches for ensemble clustering," *Machine Learning*, pp. 1–32, 2013.

[19] K. Punera and J. Ghosh, "Soft cluster ensembles," *Advances in fuzzy clustering and its applications*, pp. 69–90, 2007.

[20] N. Iam-On, T. Boongoen, and S. Garrett, "Refining pairwise similarity matrix for cluster ensemble problem with cluster relations," in *Discovery Science*. Springer, 2008, pp. 222–233.

[21] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[22] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley, "Analysing social networks within bibliographical data," in *Database and Expert Systems Applications*. Springer, 2006, pp. 234–243.