2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
July 6-11, 2014, Beijing, China

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

Pan Su, Changjing Shang, Qiang Shen
Department of Computer Science
Institute of Mathematics, Physics and Computer Science
Aberystwyth University, UK
E-mail: {pas23, cns, qqs}@aber.ac.uk

***Abstract*—Cluster ensemble offers an effective approach for aggregating multiple clustering results in order to improve the overall clustering robustness and stability. It also helps improve accuracy by combing clustering results from component methods that utilise different parameters (e.g., number of clusters), avoiding the need for carefully pre-setting the values of such parameters in a single clustering process. Since founded, many topics regarding cluster ensemble have been proposed and promising results gained. These include the generation of ensemble members and consensus of ensemble members. In this paper, link-based consensus methods for the ensemble of fuzzy c-means are proposed. Different from traditional clustering techniques, the clusters which are generated by fuzzy c-means are fuzzy sets. The proposed methods therefore employ a fuzzy graph to represent the relationships between component clusters upon which to derive the final ensemble clustering results. Using various benchmark datasets, the proposed methods are tested against typical traditional methods. The experimental results demonstrate that the proposed fuzzy-link-based clustering ensemble approach generally outperforms the others in terms of accuracy.**

I. Introduction

Clustering is one of the important approaches within the framework of unsupervised learning which is helpful 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]. A good number of clustering algorithms have been proposed in the literature, and successfully applied to a range of all data sets [2]. For a given problem, different algorithms, and indeed even the same algorithm with different parameter settings (e.g., the number of clusters assumed), typically lead to different solutions [3]. 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 evaulated. Therefore it is extremely difficult for users to decide on which algorithm to employ given their carefully selected problem domains [4].

To overcome the aforementioned limitations, improving the robustness as well as the accuracy of individual clustering methods, clustering ensemble has emerged as effective solutions. Similar to the classifier ensemble [5] and feature selection ensemble [6], cluster ensemble combines results of various clustering algorithms and may do so in different ways. One of the main objectives of the combination is to achieve accuracy superior to those of individual clustering [7]. By combining multiple partitions of a set of objects into a single consolidated clustering, the performance of cluster ensembles generally depends on both the quality and the diversity of ensemble components. This has been empirically verified [2], [4]. Consequently, two essential steps are identified that are commonly involved in the development of clustering ensemble: 1) the generation of clustering base members, and 2) the consensus of them.

A number of methods have been proposed that have helped to address these issues. For example, in order to ensure diversity of component clustering means, different parameter configurations of a given clustering algorithm have been tested [8], [9]; re-sampling techniques [10] have also been applied to diverse base clusters [11], [12], [13]. In particular, regarding the techniques for the issue of consensus, existing methods include: feature-based approach where each base-clustering member provides cluster labels as new features describing data points, which is then utilised to formulate the final solution [14], [15]; pairwise similarity-based approach which creates a matrix, containing the pairwise similarity measures amongst data points, then any similarity-based clustering algorithm (say, hierarchical clustering) can be applied [8]; graph-based approach which manipulates data partitions by exploiting graph representation [7], [16].

Although much effort has been made in the development of clustering ensemble, modelling a mechanism that is effective for integrating multiple data partitions in a cluster ensemble is far from trivial. The development and application of cluster ensembles are still at an early stage [17]. Most of the existing clustering ensemble methods are based on crisp base clusterings. However, interesting departures from the traditional work have recently been reported, such as that reported in [18] where the problem of aggregating "soft" base-clustering members is defined. Following this desirable trend, in this paper, a link-based consensus approach for building ensembles of fuzzy c-means is proposed. Different from ensembles of crisp clusters, the proposed method is able to handle fuzzy components. The work also differs from the link-based crisp clustering ensemble [3], [19], since it employs a fuzzy graph $< \{\widetilde{C}_1, ..., \widetilde{C}_n\}, \widetilde{L} >$ to represent the relationships between base-clusters and to refine the pairwise similarity matrix for generating the ensembles. With a number of benchmark datasets [20], the proposed methods are tested against their crisp counterparts and those that utilise a fuzzy co-association matrix without link-based refinement. The experi-

978-1-4799-2072-3/14/$31.00 ©2014 IEEE

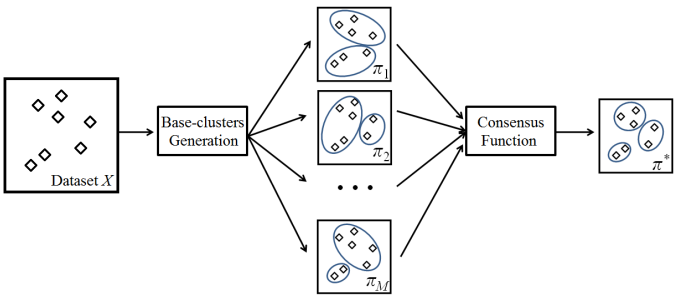

Fig. 1. Clustering Ensemble

mental results demonstrate that the fuzzy link-based clustering ensemble methods developed herein perform better than their counterparts in terms of accuracy.

The remainder of this paper is organised as follows. Section II introduces the basics of clustering ensemble. Section III defines fuzzy co-association matrix and link-based pairwise similarity matrices, and presents their applications to agglomerative clustering in an attempt to create ensembles of fuzzy clusters. Section IV reports on the experimental evaluation of the proposed approach and discuss the results. Finally, Section V concludes the paper with suggestions for further development.

II. Preliminaries

A. Clustering Ensemble

Formally, the clustering ensemble problem can be described as follows. Let $X = \{x_1, \cdots, x_N\}$ be a set of N data points and $\Pi = \{\pi_1, \cdots, \pi_m, \cdots, \pi_M\}$ be M base-clustering members. Each base-clustering member returns a set of clusters $\pi_m = \{C_1^m, \cdots, C_k^m, \cdots, C_{K_m}^m\}$ such that $\bigcup_{k=1}^{K_m} C_k^m = X$, where K_m is the number of clusters constructed by that member. For each $x_i \in X$ and each base-clustering member $\pi_m \in \Pi$, $C^m(x_i) \in \pi_m$ denotes the cluster label to which the object x_i belongs in π_m. The task of clustering ensemble is to find a new clustering result π^* given a data set X which summarises the information embedded in the whole cluster ensemble Π.

As indicated previously, two key procedures are involved in the development of a clustering ensemble technique. First, base clustering members are generated, typically by artificially diversifying methods for parameter settings and data resampling. Second, a consensus function is then applied on those base clustering members to generate the final clustering result. The procedure of clustering ensemble is illustrated in Figure 1.

A consensus function can be generally viewed as a map from a set of base-clustering members to one final partition of the original data $f : \Pi \rightarrow \pi$. Once the base-clusters are generated from the data, a variety of consensus functions that are readily available may be applied to derive the final data partition. Most of the consensus functions utilise an ensemble-information matrix which aggregates the base-clustering members. Given the ensemble of Fig. 2, two types of such a matrix: the label-assignment matrix and the binary cluster-association matrix are illustrated in Tables I and II, respectively.

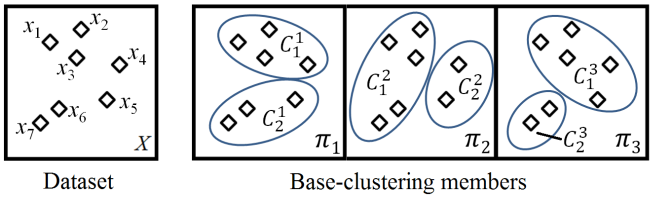

Fig. 2. Examples of Ensemble-information Matrices

TABLE I. Label-assignment Matrix

	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

TABLE II. Binary Cluster-association Matrix

	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

If a crisp clustering algorithm such as k-means is used in the generation of base-clusters, 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. Alternatively, an ensemble may be represented as a graph, where the nodes are base-clusters or data points and links between them define the relationships holding amongst the clusters and points. Graph partition methods can then be applied to the graph in order to obtain a clustering ensemble output [16].

B. Pairwise Similarity Matrix for Cluster Ensemble

Apart from the consensus functions described above, pairwise similarity matrices form another type of consensus methods. There have been various approaches for this. Take the co-association (CO) matrix [8] as an example: Given N data points, the functionality of each base-clustering member $\pi_m \in \Pi, m = 1, \cdots, M$ is equivalent to transferring the data into an $N \times N$ similarity matrix, using Eqn. (1) below:

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

Having obtained all the M similarity matrices regarding the base-clustering members, they are aggregated to form the so-called co-association matrix using Equation (2).

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

The entries in a CO matrix therefore capture the similarities between data points x_i and x_j, $i, j \in \{1, 2, ..., N\}$.

Many pairwise similarity based clustering algorithms can be applied to such a CO matrix. The agglomerative clustering is often employed to derive the final partitions [8]. The main drawback of using a crisp CO matrix is that many entries of it are zeros, which implies that two corresponding 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 [19]. Unfortunately, this characteristic is commonly encountered with the crisp clustering ensembles, thereby significantly limiting the quality of the final data partition that is to be generated by any given consensus function [3].

In order to modify such sparse-information ensemble matrices, link-based refining methods are herein proposed. In particular, the fuzzy c-means are employed to generate base-clustering members. This leads to the following CO matrix-based method for fuzzy c-means ensemble, nick-named FCO hereafter. To further improve the quality of FCO, two link-based methods (to be named as $FLink$ 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 set. 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 corresponding fuzzy cluster. Usually, this degree is normalised with all the clusters in a partition to satisfy that $\sum_{k=1}^{K_m} \widetilde{C}_k^m(x_i) = 1$.

Following the representational form used in crisp clustering ensemble (for notational consistency), the similarity measure of two objects $x_i, x_j \in X$ with respect to each base-clustering member, $S_{\widetilde{m}}(x_i, x_j)$ and subsequently, the FCO matrix are defined in Eqn. (3) and Eqn. (4) respectively:

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

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

Since $\sum_{k=1}^{K_m} \widetilde{C}_k^m(x_i)$ is normalised to 1, it follows that $S_{\widetilde{m}}(x_i, x_j) \in [0, 1]$ and $FCO(x_i, x_j)) \in [0, 1]$. Note that Eqn. (3) is a generalised version of Eqn. (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 properties of fuzzy c-means is that most of the data points have non-zero memberships to many or even all clusters. This feature is very helpful for clustering ensemble helping to retain more details of the base-clustering members in the pairwise similarity matrix. Even two data points which are not assigned in the same cluster in crisp clustering can also have non-zero values in the FCO matrix with regard to the definition Eqn. (4). This gives potentially finer discrimination of the data points.

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

In clustering ensemble, base-clustering members are usually generated from the same dataset. Hence, the resulting base-clusters in a cluster ensemble may share common data points. These shared data points create the linkage amongst base-clusters and therefore, it is possible to estimate the similarity of any base-cluster pair by exploring the underlying link information [21]. Note that the concept of a graph formulated from a set of base-clusters and a set of weighted links between them has been introduced previously, as of [19]. Given a cluster ensemble as 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 the clusters. The weighted links between base-clusters C_i and C_j is defined as:

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

where $|U|$ stands for the cardinality of the set U.

In crisp cluster ensemble, however, base-clusters within the same base-clustering member do not have common data points with each other, that is, $\forall C_k^m, C_l^m \in \pi^m$, if $k \neq l$ then $C_k^m \cap C_l^m = \emptyset$. The weights of those links between the clusters within the same base-clustering member are of a value of zero. Further refinement will therefore be necessary before they can be used in the emerging ensemble. In order to retain more information from base-clustering members and refine the FCO matrix for fuzzy c-means ensembles, 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 the 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 computed by

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

where $\widetilde{C}_i(x_t)$ indicates the the degree of a data point x_t belonging to a fuzzy cluster $\widetilde{C}_i$. Obviously, $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_j) \in [0, 1]$, $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_i) = 1$ 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 thus defined in accordance with the proportion of their overlapping degree on all data points in X. In so doing, even for two fuzzy base-clusters within the same base-clustering member, the weight of the link between them is possible to be of a non-zero value. As such, in general, each base-cluster may have a

link to all the other base-clusters, and the fuzzy degree of a given link represents the similarity between the corresponding two base-clusters.

Given a fuzzy graph, link-based pairwise similarity matrix of data points can be introduced using the fuzzy weights associated with the links. In particular, for the a clustering member $\widetilde{\pi}_m$, the link-based similarity of data points x_i and x_j can be estimated by

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

where $\widetilde{C}^m_{\max}(x_i) = \bigvee_{k=1}^{K_m} \widetilde{C}^m_k(x_i)$ and $\arg \widetilde{C}^m_{\max}(x_i) \in \pi_m$ representing the fuzzy cluster in which x_i has the maximum membership. In case of a draw, a random pick is made amongst those even clusters. From this, it has a natural appeal to define the similarity of two data points in the overall fuzzy c-means clustering ensemble as: $FLink(x_i, x_j) = \sum_{m=1}^{M} FS_{\widetilde{m}}(x_i, x_j)/M$.

Different from FCO, the link-based based similarity defined in Eqn. (7) only associates a data point x_i to the cluster of which x_i has the maximum membership degree. If two data points happen to have the maximum degrees in the same cluster, then their similarity values assigned by $LS_{\widetilde{m}}$ is deemed to be the smaller degree value of the two, 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 smaller value of their respective maximum degrees times the weight of the link between those two base-clusters where x_i and x_j have the maximum degree values.

Note that non-zero weighted links may exist not only between base-clusters within a single base-clustering member, e.g., $\exists \widetilde{L}(\widetilde{C}^m_k, \widetilde{C}^m_l) > 0$, but also between base-clusters cross base-clustering members, e.g., $\exists \widetilde{L}(\widetilde{C}^m_k, \widetilde{C}^n_l) > 0, m \neq n$. As $LS_{\widetilde{m}}$ does not employ links cross base-clustering members, it can be computed efficiently in terms of both time and memory space required. However, in crisp clustering ensemble, links cross base-clustering members are employed to estimate the similarity within base-clustering members using means such as the connected-triple [22], thereby improving the quality of the final ensemble result. Inspired by this observation, and to test whether the cross links may indeed help refine $FLink(x_i, x_j)$ further while allowing for consistent comparison with link-based crisp clustering ensemble, the connected-triple is also applied to $\widetilde{L}$ in the present work as described below.

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. It assumes that if two nodes are both connected to a third node then it is indicative of similarity between those two nodes. The connected-triple is also applied to the weighted crisp cluster ensemble graph $< V, L >$ of Eqn. (5) to generate the similarity of nodes within clustering members [19]. Specifically, the weighted connected-triple deems the similarity of two base-clusters C_i and C_j as the sum of the minimum weights to every common neighbour of theirs:

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

where $n = \sum_{m=1}^{M} K_m$ denotes the total number of base-clusters of all base-clustering members. $w'(C_i, C_j)$ may also be normalised such that $n\_w'(C_i, C_j) = w'(C_i, C_j)/w'_{\max}$, 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 can be defined by

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

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): $CTS(x_i, x_j) = \sum_{m=1}^{M} S'_m(x_i, x_j)/M$.

In a similar way, the fuzzy version of CTS can be introduced, where $\widetilde{L}(\widetilde{C}_i, \widetilde{C}_j)$ is refined using the connected-triple to become $L'(\widetilde{C}_i, \widetilde{C}_j) = \sum_{t=1}^{n} \widetilde{L}(\widetilde{C}_i, \widetilde{C}_t) \wedge \widetilde{L}(\widetilde{C}_j, \widetilde{C}_t)$, and then 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 $\widetilde{C}^m$ can be modified to:

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

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 has the maximum membership. As before, if a draw incurs, one of those even clusters is randomly taken. The similarity of two data points in the overall fuzzy c-means clustering ensemble is computed by $FCTS(x_i, x_j) = \sum_{m=1}^{M} FS'_{\widetilde{m}}(x_i, x_j)/M$.

In spite of the CTS, other link-based methods such as the SimRank-based algorithm [19] can also be modified to support fuzzy c-means ensemble. However, the implementation of link-based similarity methods (including the CTS) similarly involve high computational complexity. This drawback is inherent to the algorithms, whose simplified variation may not be able to maintain the original performance [23]. Hence, the $FCTS$, which requires less computational time compared with the others, is developed in this work.

D. Link-based Fuzzy C-means Ensemble

The overall process of using the proposed matrices in building clustering ensembles is similar to that of the existing work that uses pairwise similarity matrices (e.g., [9]). To save space only the two main steps are outlined below:

1) Fuzzy c-means are used on the dataset X for M times to generate fuzzy base-clusters. The diversity of base-clustering members is ensured by a combination of re-sampling the original datasets, different numbers of learned clusters, and different initial centroids for fuzzy c-means. Note that in theory, many other methods used in crisp clustering ensemble can also be used in place of fuzzy c-means ensemble, though the current work only uses the latter for simplicity.
2) Any of the three proposed methods (FCO, $FLink$, $FCTS$) can be used to generate a pairwise similarity matrix of data points, exploiting the information embedded in base-clustering members. From this, a pairwise similarity based clustering algorithm, such as hierarchical clustering, can then be employed to generate the final partition of the dataset as the output of cluster ensemble.

IV. Experimentation and Evaluation

This section presents an experimental evaluation of the proposed work. It first outlines the set-up of the experiments carried out and then 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 to compare the performances of different methods.

A. Experimental Set-up

To evaluate the performance of proposed methods, they are experimentally tested over seven datasets obtained from UCI benchmark repository [20], where true labels of instances are known but are not explicitly used in the cluster ensemble learning process. The details of these datasets are summarised in Table III. The final results of the resulting cluster ensembles are evaluated in terms of accuracy as the group truth for each dataset is known.

TABLE III. Summary of 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

The fuzzy c-means clustering algorithm is used to generate the base clustering members. Thirty clustering-members are created ($M = 30$) and the cluster centroids are randomly initialised in each run. Two agglomerative clustering approaches (complete-linkage and average-linkage) [9] are selected to implement the consensus function. These consensus functions divide data points into clusters using the underlying similarity matrix $FLink$, FCO, $FCTS$, or CTS. For fair comparison, the number of final clusters on each dataset is set to that of its true classes and the decay factor (DC) of CTS is commonly set to 0.5 [19], and 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

1) Sensitivity of proposed methods: This is to check the robustness of the approach against the diversity of base-clustering members. To vary the base-clustering members, the maximum number of base-clusters $\max(K_m)$ in each test is set from 3 to 30 with an increment step of 3, and the number of base-clusters in each clustering-member K_m is randomly chosen from $[3, \max(K_m)]$. Figure 3 shows the change of accuracy with respect to the increase of diversity in base-clustering members where agglomerative clustering with average-linkage is used as the consensus function. Each point in Fig. 3 is an averaged value of 50 runs.

For five of the seven datasets, the accuracies of the three proposed methods ($FLink$, $FCTS$ and CTS) generally increase along with the increase of diversity. This indicates that the use of link-based pairwise similarity matrices in fuzzy c-means ensemble entails more differences in base-clustering members, which in turn allows the generation of better results. The outcome of using FCO seems to be more stable as compared with link-based methods. This indicates that 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 point has gained a certain membership to all the clusters. Thus, the base-clustering members which have a smaller number of clusters can retain as much information as the ones of a lager cluster number. However, the accuracy of FCO is not so high as that achievable by the link-based methods in general. This shows that although fuzzy c-means can help FCO to keep more information for building ensemble, the link-based refinements are helpful in generating more effective pairwise similarity matrices.

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

The results show that for both fixed and random K_m, the use of link-based pairwise similarity matrix $FLink$ leads to the best average accuracy over the seven datasets, in building fuzzy c-means ensembles. However, the performance of $FCTS$ is not significantly better than $FLink$ in general. This implies that the connected-tripe method does not necessarily further refine $FLink$ effectively. Note that both $FLink$ and $FCTS$ achieve a better accuracy than CTS on most of the datasets. Although the CTS which employs the 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 fuzzy links $\widetilde{L}$

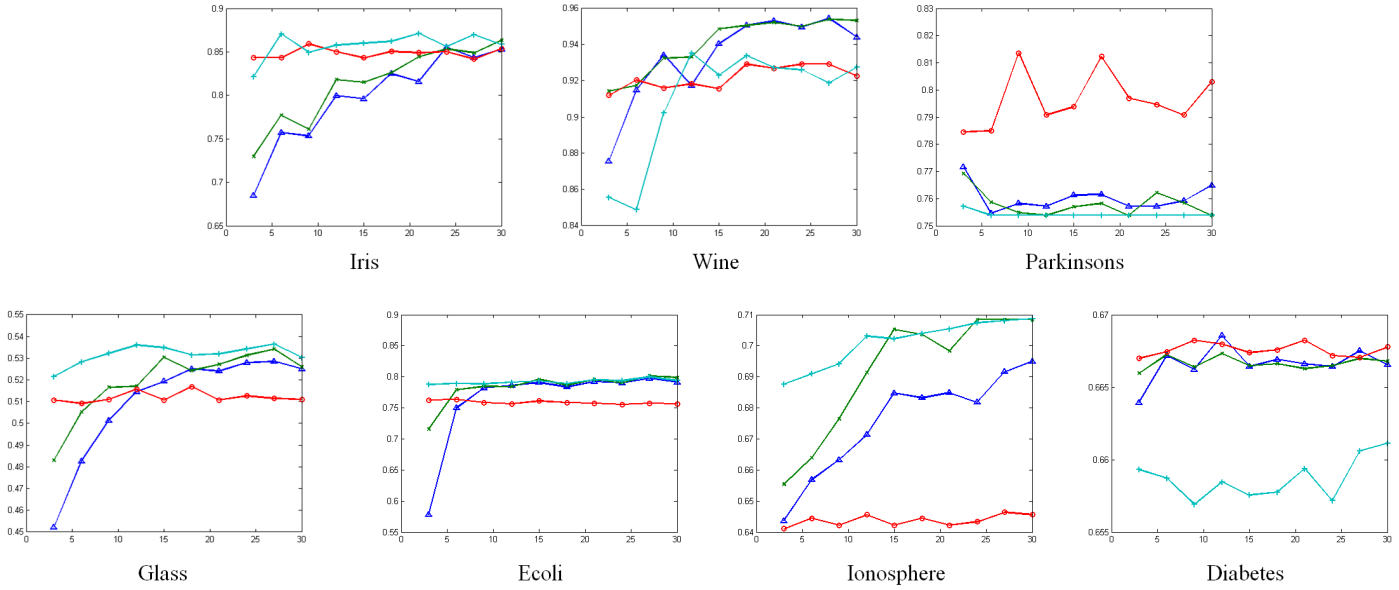

Fig. 3. Trend of Accuracy Change against Diversity (FCO: –o–, $FLink$: –x–, $FCTS$: –△–, CTS: –+–)

TABLE IV. 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 V. 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

in $FLink$ and $FCTS$. Particularly, $FLink$ can use the fuzzy links $\widetilde{L}(\widetilde{C}_k^m, \widetilde{C}_l^m)$ where $k, l = 1, \cdots, K_m$ directly without inferring them from $\widetilde{L}(\widetilde{C}_k^m, \widetilde{C}_l^n), m \neq n$, the time for running the connected-triple method (or the other similar refinement) is saved. In conclusion, $FLink$ entails higher accuracy but lower time-consumption than CTS.

V. CONCLUSION

This paper has presented the notion of co-association matrix and those of link-based pairwise similarity matrices for fuzzy c-means cluster ensemble. The proposed matrices take the advantage of fuzzy c-means in that each data point can have a membership to all clusters. A set of fuzzy links between base-clusters is defined and a fuzzy graph is employed to generate the link-based similarity matrices. Experimental results on seven UCI datasets indicate that the proposed approach generally outperforms the conventional CTS. Furthermore, the link-based methods also help to build better pairwise similarity matrices as compared to the non-link based matrix FCO.

Whilst promising, the present work 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 using different consensus functions. It is also interesting to examine whether any methods based on fuzzy graph theory rather than the connected-triple may be more suitable and efficient in dealing with the proposed 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] 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.

[3] 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.

[4] 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.

[5] R. Diao, F. Chao, T. Peng, N. Snooke, and Q. Shen, "Feature selection inspired classifier ensemble reduction," *Cybernetics, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.

[6] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," in *Proceedings of the Alan Turing centenary conference*, 2012.

[7] 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.

[8] 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.

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

[10] R. Diao and Q. Shen, "Fuzzy-rough classifier ensemble selection," in *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1516–1522.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

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

[19] 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.

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

[21] P. Su, C. Shang, and Q. Shen, "Link-based approach for bibliometric journal ranking," *Soft Computing*, vol. 17, no. 12, pp. 2399–2410, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00500-013-1052-4

[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.

[23] N. Iam-On and S. Garrett, "Linkclue: A matlab package for link-based cluster ensembles," *Journal of Statistical Software*, vol. 36, no. 9, pp. 1–36, 2010.