Neurocomputing 148 (2015) 487–497

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

An information theoretic approach to hierarchical clustering combination

Elaheh Rashedi [a], Abdolreza Mirzaei [b,*], Mohammad Rahmati [c]

[a] Computer Science Department, Collage of Engineering, Wayne State University, Detroit, MI, USA
[b] Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan 84156-83111, Iran
[c] Image Processing and Pattern Recognition Laboratory, Computer Engineering and Information Technology Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Hafez Ave., 424, Iran

ARTICLE INFO

Article history:
Received 14 July 2013
Received in revised form
7 April 2014
Accepted 5 July 2014
Communicated by: Deng Cai
Available online 14 July 2014

Keywords:
Clustering combination
Dendrogram descriptor
Divergence measure
Hierarchical clustering

ABSTRACT

In Hierarchical clustering, a set of patterns are partitioned into a sequence of groups represented as a dendrogram. The dendrogram is a tree representation where each node is associated with merging of two (or more) partitions and hence each partition is nested into the next partition. Hierarchical representation has properties that are useful for visualization and interpretation of clustering results. On one hand, different hierarchical clustering algorithms usually produce different dendrograms. On the other hand, clustering combination methods have received considerable interest in recent years and they yield superior results for clustering problems.

This paper proposes a framework for combining various hierarchical clustering results which preserves the structural contents of input hierarchies. In this method, first a description matrix is created for each hierarchy, and then the description matrices of the input hierarchies are aggregated to form a consensus matrix from which the final hierarchy is derived. In this framework, we use two new measures for aggregating the description matrices, namely Rényi and Jensen–Shannon Divergences. The experimental and comparative analysis of our proposed framework shows the effectiveness of these two aggregators in hierarchical clustering combination.

1. Introduction

Clustering is a process of forming groups (clusters) of similar patterns from a given set of inputs. A clustering algorithm seeks to group patterns such that patterns belonging to the same cluster are “similar” to each other, while patterns from different clusters are “dissimilar”. Clustering is used extensively as a fundamental data analysis tool in different fields such as data mining, image processing, machine learning, and bioinformatics [1,2].

There have been many approaches presented for data clustering [3,4]. The two main groups of clustering algorithms are hierarchical and nonhierarchical (partitional) clustering algorithms. In partitional algorithms, the number of clusters, k, is usually assumed to be known in advance. The inputs to partitional algorithms are the data, a distance metric, and the number of clusters, k. The output of each algorithm is a model of the data from which the memberships of patterns to different clusters can be derived.

The popular k-means algorithm belongs to the family of partitional clustering algorithms [4]. This method starts with a random set of centroids and assigns each pattern to its closest centroid. Then, repeatedly, for each group, based on its members, a new central point (new centroid) is calculated and pattern assignments to their closest centroids are changed, if necessary. The algorithm finishes when no pattern reassignments are needed or when certain amount of time elapses.

Hierarchical clustering algorithms work by merging the nearest clusters in the bottom-up fashion (agglomerative clustering) or splitting clusters into separate clusters in the top-down fashion (divisive clustering) [5]. In the Agglomerative Hierarchical Clustering (*AHC*) algorithms, each individual pattern is first assigned to a cluster containing only that pattern, then two clusters that are closest to each other are merged into a new group and this process continues until we reach to a cluster which contains all of the patterns. In the case of top-down, first, a cluster containing all patterns is created, and then this cluster is divided into two other clusters with respect to the amount of separation between patterns. This process is continued until the final clusters contain only one pattern. The relationship between the input patterns and the output of a hierarchical clustering algorithm, which is a hierarchy of

* Corresponding author. Tel.: +98 311 391 2450; fax: +98 311 391 2451.
E-mail addresses: fp0634@wayne.edu (E. Rashedi), mirzaei@cc.iut.ac.ir (A. Mirzaei), Rahmati@aut.ac.ir (M. Rahmati).

http://dx.doi.org/10.1016/j.neucom.2014.07.014
0925-2312/

clusters, is well represented by a tree which is known as dendrogram. Dendrograms offer better view of data distribution in different abstraction levels. This property makes the hierarchical clustering algorithms an ideal choice for data exploration and visualization. Furthermore, in some applications the number of clusters is not known in advance, and the dendrogram could provide a visualization method for user to decide on the number of clusters.

It can be shown that the performance of data clustering is improved by combining the results of several clustering algorithms [6–22]. These approaches are called ensemble methods. Ensemble methods contain two steps in general. In the first step, multiple clustering results are created, which is called the ensemble. And in the second step, the results from multiple clustering techniques are combined, using a consensus function which is called an aggregator [23], to create a single and integrated model for input dataset.

Many clustering combination techniques are introduced to create ensembles, with a comprehensive body of works on partitional clusterings [3,4,22] and a few are introduced on hierarchical clusterings. These methods are discussed in below. A categorization of different clustering combination methods, according to the consensus function which is used, is presented as the following.

1.1. Partitional clustering combination approaches

Many consensus functions are introduced for partitional clustering ensembles which used a variety of mathematical tools [22]. Some are introduced as follows: information theory, fuzzy clustering, genetic algorithms, relabeling and voting, co-association matrix, graph and hypergraph, Mirkin distance, finite mixture models, locally adaptive clustering algorithm, kernel and non-negative matrix factorization [22].

Most clustering combination methods are based on partitional base clusterers, i.e. the input of all combinational clustering algorithms is nonhierarchical and if one is interested in combining a set of hierarchical clusterings using aforementioned methods, the results of clustering techniques should be converted to non-hierarchical. In this conversion, only one level of each hierarchy, which itself is a clustering of data, is preserved and the structural contents of the hierarchy will be lost. Following this conversion, any combinational method such as stacked clustering [24] may be used to produce a consensus hierarchy. In this method, only the information of one level of primary hierarchical clusterings is used while useful information that exists in other levels of hierarchical clustering that may be used to improve the quality of combination methods are ignored. The objective of this paper is to propose a new framework for hierarchical clustering combination which preserves the structural contents of input hierarchies.

1.2. Hierarchical clustering combination approaches (HCC)

Some consensus functions are introduced for hierarchical clustering ensembles. Among them, an mean aggregator based method [25], a fuzzy similarity relation based method [26,27] and a boosting based method [28] can be found. These methods are described more in Section 3.

In the area of supervised learning, there exists a similar problem as hierarchical clustering combination, which is combining Decision Trees (DT) of classifiers. Different methods for combination of decision trees have been proposed. The boosting of DT [29] or random forests [30] are two examples to be mentioned. In these methods, the outputs of base trees are combined to produce the output for each new instance and, therefore, they do not create a DT from the ensemble. The only exceptions, to the best of our knowledge, are the works of [31–33] in which they use Fourier analysis to aggregate the trees in an ensemble to construct a single informative DT [31]. Nevertheless, all of aforementioned DT combination methods use pattern labels and therefore could not be used in an unsupervised scenario.

In this paper, we propose an *HCC* method in which the hierarchical clustering results are combined into a one representative consensus clustering. In this method, two new aggregators are use for combining the description matrices, namely Rényi and Jensen–Shannon divergences. The rest of this paper is organized as follows. In Section 2 the Hierarchical Clustering Combination problem, *HCC*, is introduced in its general framework. This framework includes the two main steps of ensemble methods, i.e. creation and the combination task. First, the hierarchies are created using clustering methods, and then, the hierarchy resulted from each clusterer is converted to a description matrix. The description matrices are used as middle structures. In the end, the combination task is performed on description matrices. In Section 3, different description matrices of a given dendrogram are introduced and following that the method used to recover the final hierarchy from the consensus matrix is described. Section 4 presents the theory and a description of the consensus method which is proposed for aggregating different descriptor matrices. In Section 5, the experimental set ups are declared via Section 5.1–5.3. Section 6 discusses the experimental results and Section 6 compares the performance of the developed techniques to the state of the art. Finally, a summarization of the main conclusions of this work is given in Section 8. Derivations of the formulas used in aggregation step are presented in Appendices A and B.

2. Hierarchical clustering combination

The problem of hierarchical clustering combination may be stated as follows:

2.1. Given a set of dendrograms, find a new dendrogram which Is a proper representative of the whole dendrograms set

In order to propose an algorithm for hierarchical clustering combination, as requires in the above statement, the term “proper representative” must be defined clearly. In this paper, a proper representative of a set of dendrograms is a dendrogram which is as close as possible to all dendrograms of the set. The distances between a dendrogram and a set of dendrograms are measured by middle structures which are called dendrogram descriptors. In other words, the proximity of two dendrograms is defined as the proximity of their descriptors. Fig. 1 illustrates the general framework proposed for Hierarchical Clustering Combination (*HCC*). Different *HCC* methods, under certain conditions, could be represented as a special case of this framework. Such a framework allows us to compare and highlight important common features among different combination methods and to draw new insights, thereby providing a basis for constructing new methods.

In this algorithm, the structural content of kth dendrogram is denoted by $H^{(k)}$ which is extracted and represented as a description matrix $T^{(k)}$. This step is performed by applying a predefined function f on the dendrogram. Then the description matrices $T^{(k)}$, $1 \le k \le L$ are aggregated into a final description matrix T which we call, hereafter, *consensus matrix*. Following this step, the final hierarchy is derived from T.

In the following sections, we present answers to two major questions regarding this algorithm, which are:

- Which description matrix is used to represent the hierarchical structure of a dendrogram?
- How to aggregate the description matrices of input dendrograms?
- How the final hierarchy is derived from the consensus matrix?

The General *HCC* Algorithm

Input: Dataset $Z = \{z_1, z_2, ..., z_N\}$
Ensemble size L

Variables:
Descriptor-type
Aggregator
Recovery-method

Start:
1- Generate L hierarchical clusterings of Z, $H^{(k)}$, $1 \le k \le L$, represented as dendrograms
2- Create a description matrix $T^{(k)}$ for each hierarchy $H^{(k)}$, according to the value of *Descriptor-type* variable.
$T^{(k)} = f(H^{(k)}) \quad 1 \le k \le L$
3- Aggregate the description matrices to a final matrix T, using the *Aggregator*
$T = \text{Aggregate}(T^{(1)}, T^{(2)}, ..., T^{(L)})$
4- Create the final dendrogram from the T, using the *Recovery-method*
$H = f^{-1}(T)$

Fig. 1. The general framework of Hierarchical Clustering Combination (HCC) algorithm.

The three variables, *Descriptor-type*, *Aggregator* and *Recovery-method* are described in Section 3 and 4.

3. Dendrogram descriptor matrix and hierarchy recovering method

A dendrogram associated to a hierarchical clustering of N input patterns can be represented by a description matrix of size $N \times N$. A description matrix expresses the relative position of a given pair of terminal vertices (i.e. patterns pair) in a dendrogram. If $T^{(k)} = \{t_{ij}^{(k)}\}$ is the description matrix of kth hierarchy, i.e. $H^{(k)}$, then the $t_{ij}^{(k)}$ represents how much the two patterns i and j within the kth dendrogram are different. Various descriptors are presented to reveal different structural aspects of a dendrogram. Some of these descriptors are Partition Membership Divergence (*PMD*), Cluster Membership Divergence (*CMD*), Sub-tree Membership Divergence (*SMD*), Cophenetic Difference (*CD*), Maximum Number of Edge Distance (*MNED*) [34].

All descriptors introduced above are ultrametric. A description matrix T is ultrametric if following inequality is satisfied [38]

$$t_{ij} \le \max(t_{il}, t_{lj}) \tag{1}$$

for all triplets of points i,j, and l. Any dendrogram is characterized by a unique ultrametric matrix and can be recovered from its corresponding ultrametric matrix through hierarchical clustering. In other words, hierarchical clustering and ultrametric description of a dendrogram are the inverse of each others. Although the description matrices of base clusterers are ultrametric, their aggregation which makes the consensus matrix is not necessarily ultrametric. However, if T is used as input to any hierarchical clustering algorithm that operates directly on a dissimilarity matrix, the Hcan be recovered from it which is considered as the resultant clustering. It should be noted that, in Fig. 1, applying the same f function of step 2 to H, will yield T', which is equal to T or has small deviation from it.

Now, the previous *HCC* methods can be declared clearly. In the combination method presented in [25], the descriptor matrices are aggregated using *mean* matrix operation. The aggregated matrix might not necessarily have an associated dendrogram. Therefore, a dendrogram recovery phase is then applied to derive the final dendrogram from the aggregated dissimilarity matrix. The quality of the consensus clustering proved to be better than any single clustering. Similar to this method, Another combination technique, *BobHic*, is presented in [28]. In this method, the descriptor matrices of the ensemble are combined using minimum matrix operation.

Another combination method based on fuzzy similarity relations, *MATCH*, is proposed in [26]. In this method, descriptors are aggregated into a transitive consensus matrix from which the final dendrogram can be directly formed. Unlike two previous methods, this technique does not need the additional recovery phase.

A hierarchical ensemble clustering technique, *HEC*, is also introduced in [27], which combines hierarchical clusterings into a one consensus clustering. *HEC* and *MATCH* are dual algorithms; the former generates minimum transitive dissimilarity matrix closure, and the latter generates maximum transitive similarity matrix closure [28].

4. Descriptor matrix aggregation

In clustering algorithms, it is generally desired that each pattern to be grouped into the same cluster as its nearest neighbor. It is noted that, row (column) i of a dendrogram description matrix shows the differences between pattern $i$and all other patterns. Given the pattern z_i, if we randomly select pattern z_j, then the probability that z_j not to be the nearest neighbor of z_i is obtained by

$$p_{ij} = \frac{t_{ij}}{\sum_{j=1}^{N} t_{ij}}, \tag{2}$$

where N is the number of patterns in the dendrogram and t_{ij} is the (i,j)th entry in the description matrix. It can easily be verified that $p_{ij}, 1 \le j \le N$ is a discrete probability distribution function i.e. $\sum_{j=1}^{N} p_{ij} = 1$ and $0 \le p_{ij} \le 1$. Hence, when the values of all elements at each row (column) are normalized, then this row (column) could be used to predict the nearest neighbors of this row's corresponding pattern. Each normalized row (column) of the description matrix is characterized to be a discrete probability distribution function (*PDF*) of a random variable.

The justification to consider each normalized row as a discrete probability distribution function of a random variable is as follows. Each description matrix is an ultrametric pair-wise distance matrix whose elements show the distance between different instances. The elements of row i of a description matrix show the distances of instance i from other instances in the corresponding dendrogram. If each column is assumed as a class label, the probability that instance i

do not belong to class j (column j) can be approximated by its distance to class j compared to its distance to other classes, which is equivalent to normalizing each row of description matrix to create a *PDF*.

In order to fulfill the objective of our stated problem, in the *HCC* algorithm, the description matrix of final hierarchy must have the minimum possible distance from the description matrices of the input dendrograms. Because each row (column) of a normalized description matrix is a *PDF*, an information theoretic measure can be used to assess the proximity of two normalized dendrogram descriptors. In regard to our objective and considering this aspect, the problem of aggregating a set of descriptors could be stated as follows. The desired final description matrix is the one in which each of its rows is a *PDF* whose distance from the corresponding *PDFs* of the input description matrices is minimized. The distance between two probability distribution functions may be computed by many divergence measures which have been proposed in the literature [35,36]. These measures represent the amount of information needed to transform one distribution to another one [35].

Let $p_{i.}^{k}$ and $p_{i.}^{*}$ be the ith row of the kth normalized description matrix and the final description matrix (consensus matrix), respectively. Let a specific divergence measure to be denoted by D. Also, recall that the number of clusterers in the combination is denoted by L. The average distance between $p_{i.}^{*}$ and the corresponding rows of different clusterers description matrices, $p_{i.}^{k}$ $\quad 1 \le k \le L$, is

$$D(i) = \frac{1}{L}\sum_{k=1}^{L} D(p_{i.}^{k}, p_{i.}^{*}). \tag{3}$$

The ith row of the final description matrix, denoted by $p_{i.}^{*}$, is the one that minimizes $D(i)$. In the following sections, we propose two new methods for estimating the final description matrices which are based on the Rényi [35] and Jensen–Shannon divergences [36].

It should be mentioned that the Euclidean distance was probably the most commonly used distance measure in aggregation process of the previous proposed methods. However, the fact that the distances in each dimension are squared before summation, place a great emphasis on those features for which the dissimilarity is large [37]. This makes Euclidean distance unsuitable for many applications. The above mentioned distance measures which will be introduced in the following sections are generalized distances with adjustable parameters. Consequently, this generalized framework leads to a class of new operators (adjustable by a parameter) none of them tried before.

4.1. Rényi divergence

The Rényi entropy, which is a generalization of the Shannon entropy, is a function for assessing the uncertainty or randomness of a random variable. The Rényi entropy of degree α is defined as [35]

$$H_{R\alpha}(P) = \frac{1}{1-\alpha}\log\left(\sum_{j=1}^{n} p_j^{\alpha}\right), \tag{4}$$

where P is a discrete probability distribution function of a random variable X with $p_j = \Pr\{X = x_j\}$ $\quad$ for $\quad j = 1, \ldots, n$.

The Rényi divergence of degree α of a distribution $Q(x)$ from another distribution $P(x)$ is defined as:

$$RD_{\alpha}(P, Q) = \frac{1}{\alpha-1}\log\left(\sum_{j=1}^{n}\frac{p_j^{\alpha}}{q_j^{\alpha-1}}\right) = \frac{1}{\alpha-1}\log\sum_{j=1}^{n} p_j^{\alpha} q_j^{1-\alpha}. \tag{5}$$

Substituting the Rényi divergence into (4), the $D(i)$ becomes,

$$D(i) = \frac{1}{L}\sum_{k=1}^{L} RD_{\alpha}(p_{i.}^{*}, p_{i.}^{k}). \tag{6}$$

Since the $p_{i.}^{*}$ is assumed to be a *PDF*, then it is required that:

$$\sum_{j=1}^{N} p_{ij}^{*} = 1, \quad p_{ij}^{*} \ge 0. \tag{7}$$

The Lagrange multiplier is used to introduce the constraint of (8) in the minimization of the divergence:

$$D'(i) = D(i) + \lambda\left(1 - \sum_{j=1}^{N} p_{ij}^{*}\right). \tag{8}$$

In appendix (A) the optimal p_{im}^{*} is derived by minimizing the Eq. (9), which is given by

$$p_{im}^{*} = \frac{1}{r}\left(\sum_{k=1}^{L}(p_{im}^{k})^{1-\alpha}\right)^{1/1-\alpha}, \tag{9}$$

where r is a proper constant which makes the approximations sum to 1. If we denote the term $1-\alpha$ by β the following general form is derived,

$$p_{im}^{*} = \frac{1}{r}\left(\sum_{k=1}^{L}(p_{im}^{k})^{\beta}\right)^{\frac{1}{\beta}}. \tag{10}$$

For some special values of β, Eq. (10) leads to simple aggregators such as: Min, Max, Product, etc. Some special cases of Eq. (10) are shown in Table 1.

4.2. Jensen–Shannon divergence

Jensen–Shannon divergence (JSD) is another measure of distance between two (or more) probability distributions which is introduced by Lin [36]. This measure is defined as

$$JSD(P, Q) = H_s(\overline{PQ}) - \left(\frac{H_s(P) + H_s(Q)}{2}\right), \tag{11}$$

where P, and Q are the probability distributions whose distance is to be evaluated. The parameter $\overline{PQ}$ denotes their average, and H_s is the Shannon entropy.

The Shannon entropy $H_s(P)$ measures the information content of $P(x)$ and is given by [38]

$$H_s(P) = -\sum_{j=1}^{n} p_j \log p_j. \tag{12}$$

Substituting Jensen–Shannon divergence, Eq. (12), into Eq. (4) the $D(i)$ becomes

$$\begin{aligned} D(i) &= \frac{1}{L}\sum_{k=1}^{L} JSD(p_{i.}^{*}, p_{i.}^{k}) \\ &= \frac{1}{L}\sum_{k=1}^{L}\left(H_s\left(\frac{p_{i.}^{*} + p_{i.}^{k}}{2}\right) - \frac{H_s(p_{i.}^{*}) + H_s(p_{i.}^{k})}{2}\right) \\ &= \frac{1}{L}\sum_{k=1}^{L}\left(-\sum_{j=1}^{N}\frac{p_{ij}^{*} + p_{ij}^{k}}{2}\log\frac{p_{ij}^{*} + p_{ij}^{k}}{2} + \frac{1}{2}\sum_{j=1}^{N} p_{ij}^{*}\log p_{ij}^{*} \right. \\ &\quad \left. + \frac{1}{2}\sum_{j=1}^{N} p_{ij}^{k}\log p_{ij}^{k}\right). \end{aligned} \tag{13}$$

Table 1
Different methods for combining the description matrices based on the Rényi divergence.

β	p_{im}^{*}	Name
$\beta \to -\infty$	$\min_k p_{im}^{k}$	Minimum
$\beta = -1$	$\frac{1}{r}\left(\sum_{k=1}^{L}\frac{1}{p_{im}^{k}}\right)^{-1}$	Harmonic mean
$\beta = 0$	$\frac{1}{r}\prod_{k=1}^{L} p_{im}^{k}$	Geometric mean (Product)
$\beta = 1$	$\frac{1}{r}\sum_{k=1}^{L}(p_{im}^{k})$	Arithmetic mean
$\beta = 2$	$\frac{1}{r}\sqrt{\sum_{k=1}^{L}(p_{im}^{k})^2}$	Euclidean length
$\beta \to +\infty$	$\max_k p_{im}^{k}$	Maximum

Minimizing Eq. (13), p^*_{im} is equal to the eigenvalue of the following matrix which satisfies $0 \le p^*_{im} \le 1$.

$$A = \begin{bmatrix} -a_L & -a_{L-1} & \cdots & -a_2 & -a_1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \tag{14}$$

where parameters of the entries are given as:

$$a_1 = \frac{1}{(1-2^L)} \frac{1}{L!} \sum_{k_1=1}^{L} \sum_{k_2=1}$$

$$k_1 \ne k_2 \ne \cdots \ne k_L{}^L \cdots \sum_{k_L=1}^{L} (p_{im}^{k_1} p_{im}^{k_2} \cdots p_{im}^{k_L}), : a_{L-1} = \frac{1}{(1-2^L)} \frac{1}{2!} \sum_{k_1=1}^{L} \sum_{k_2=1}$$

$$k_1 \ne k_2{}^L (p_{im}^{k_1} p_{im}^{k_2}), a_L = \frac{1}{(1-2^L)} \sum_{k_1=1}^{L} (p_{im}^{k_1}). \tag{15}$$

The complete derivation of these equations is given in appendix *B*.

5. Experimental setup

Several experiments have been conducted to evaluate our proposed framework. The plans made for performing the experiments are declared in this section. In Section 5.1, The evaluation method used for measuring the clustering quality is presented and the datasets used in the experiment are introduced. The behavior analysis of the proposed aggregator of our *HCC* technique is clarified in Section 5.2. And finally in Section 5.3, it is explained that how the set of experiments are performed. The experimental results are then described in Section 6.

5.1. Evaluation method and datasets

In order to experimentally verify the efficiency of our proposed methods, we need to incorporate a quality evaluation criterion. Different indices have been proposed to assess the validity of a cluster hierarchy generated from the proximity matrix of input data [5]. The aim of these indices is to verify whether the clustering structure produced by a clustering algorithm fits the data, using only information inherent in the data. The best known index is cophenetic correlation coefficient (*CPCC*) which compares the distance information in proximity matrix of input data and the distance information resulted from the cluster hierarchy. If W is the cophenetic matrix of the cluster hierarchy and Y is proximity matrix of input data, the cophenetic correlation between W and Y is defined as

$$CPCC = \frac{\sum_{i<j} (Y_{ij} - \overline{Y})(W_{ij} - \overline{W})}{\sqrt{\sum_{i<j} (Y_{ij} - \overline{Y})^2 \sum_{i<j} (W_{ij} - \overline{W})^2}} \tag{16}$$

Where Y_{ij} is the distance between objects i and j in Y, W_{ij} is the distance between objects i and j in W, and $\overline{Y}$ and $\overline{W}$ are the average of Y and W, respectively. The values of the *CPCC* are between -1 and 1. The closer the *CPCC* index to 1, the better the agreement between the cophenetic and the proximity matrix. The *CPCC* index may be used to compare the results for two different hierarchies of clusterings resulting from two different clustering algorithms.

Since the performance of a clustering combination method may be dependent on the inherent structure and the number of existing clusters of the input data, different datasets must be used to estimate the quality of the clustering (combination) method.

In this experiments 12 real datasets are used. The characteristics of these datasets are shown in Table 2 [39,40].

Table 2
The source and characteristics of datasets used in the experiments.

Number	Data names	#instances	#features
D1	Wine	178	13
D2	Ionosphere	351	34
D3	Wpbc	198	32
D4	Image_segmentation	210	19
D5	Liver_disorders	345	6
D6	Weaning	302	17
D7	Laryngeal1	213	16
D8	Laryngeal3	353	16
D9	Contractions	98	27
D10	Breast_cancer	263	9
D11	Flare_solar	144	9
D12	Titanic	24	3

The diversity between the outputs of base clusterings is an important issue for improving the quality of the combined clustering. If each clusterer derives an output similar to the output of the other clusterers, the combination of the outputs will not lead to a better result. In the following experiments, subsampling method is used to create the required diversity between the base clusterings of the ensemble. In other words, a random subsample (without replacement) of size $0.80N$ is used for creating each of the base clusterers, where N is the number of patterns in the original dataset.

It must be noted that since we use subsampling method to create a diverse ensemble of clusterings, not all patterns will be included in the datasets. Therefore, defining the difference between patterns which are missing in the clustering algorithm tree is impossible and the description matrices will not be completely filled. The missing values of the description matrix, which corresponds to the differences between patterns that are not included in the datasets and other patterns, are filled with 1. If subsampling is used with uniform probability, all patterns have the same chance to be excluded from a specific dendrogram. In the case that sufficiently large numbers of clusterers are combined, the effect of missing values is negligible. A more technical approach to handle unfilled entry is to fill the missing values with “don't cares”. In this situation, the aggregation operator skips don't care terms. But because it is not clear how JSD aggregator should work with don't care values, we skip this method and use the former simple method.

Because of existing randomness in the base clusterer creation, running one trial of combination method is not sufficient for our comparison. Thus 10 different ensembles were generated and the clustering quality results are averaged. In our set of experiments, each ensemble contains 10 Single Linkage (SL) hierarchical clusterers ($L = 10$).

Totally, 100 Single Linkage hierarchical clusterers (10 ensembles each with 10 clusterers) are created for each dataset. To get a better view about the quality of the base clusterers, Fig. 2 shows the *CPCC* histogram of base clusterers for different datasets. In these histograms, the *X*-axis represents the *CPCC* interval ($[-1,1]$) divided into a number of equally spaced sections and the *Y*-axis shows the number of the base clusterers which their *CPCC* values fall within each section. The poor performance of base clusterings of each dataset must not come as surprise (see Fig. 2), because they are built only on 80% of their training patterns. Next we will show that the performances of combined clusterings are superior to any of the clusterings performing alone.

5.2. Behavior analysis of Rényi aggregator in a HCC framework

In Section 4.1, we have shown that using different degrees of Rényi divergence leads to different aggregators. In order

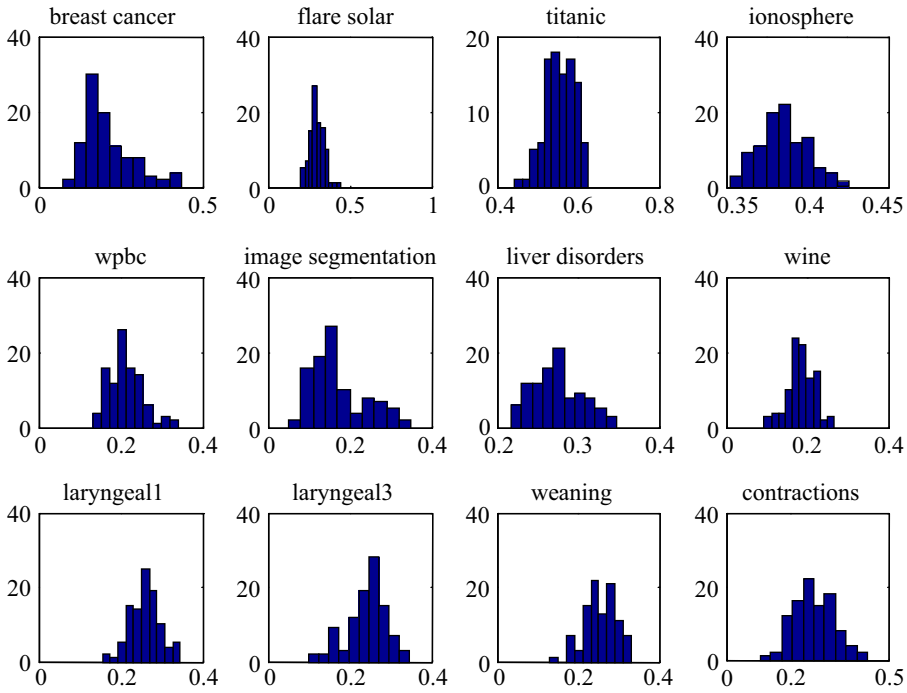

Fig. 2. Histogram of base clusterers *CPCC* for different dataset.

to illustrate the behavior of these aggregators, we have applied them to the Ionosphere dataset [39] and compared the clustering performance of resulted *HCC* methods with standard hierarchical clustering methods. The degrees of Rényi divergence that was used are $\beta \in \{-32, -24, -16, -8, -4, -2, 1, 0, 1, 2, 4, 8, 16, 24, 32\}$. Fig. 3 represents the *CPCC* values of these algorithms for different dendrogram descriptors. In Fig. 3 the final hierarchy is generated using the standard hierarchical clustering algorithms: Single Linkage [41], Complete Linkage [41], Average Linkage (unweighted pair-group) method using arithmetic average (UPGMA) [41], Weighted Linkage or weighted pair-group method using arithmetic average (WPGMA) [41], and Ward or minimum variance algorithm [41]. The straight lines in Fig. 3 show the performance of this algorithm. In this Figures, it is noticeable that in most cases, the *CPCC* of Rényi aggregator with negative values of β is superior in comparison with its positive values. It is also noted that in most cases the *HCC* methods with Rényi aggregators of degree $\beta < 0$, no matter which description matrix used, outperform the Single Linkage clustering algorithm.

5.3. Performing the set of experiments

Now, we perform a set of experiments. There are three variables in the general framework of *HCC* (see Fig. 1), which are *Descriptor-type*, *Aggregator* and the *Recovery-method*. For the first variable, *Descriptor-type*, the descriptors described earlier, i.e. *CD*, *PMD*, *CMD*, *SMD*, and *MNED*, are used in our experiments. The newly introduced *Aggregators* which used to combine the base clusterers description matrices are the Rényi aggregators of degree $\beta \in \{-32, -24, -16, -8, -4, -2, 1, 0, 1, 2, 4, 8, 16, 24, 32\}$ and Jensen-Shannon aggregator (JSD). Hence the final hierarchy is generated by applying a common hierarchical clustering algorithm to this dissimilarity matrix. Single Linkage, Complete Linkage, Average Linkage or UPGMA, Weighted Linkage or WPGMA, and Ward are the hierarchical clustering algorithms which are used as *Recovery-methods* in these set of experiments. There is another variable which is related to the missing values of the description matrix. Missing values can be filled with 0 or 1 (*Missing-value*). Accordingly, the *Descriptor-type* variable get 5 different values, the *Aggregator* variable gets 16 different values, the *Recovery-method* variable gets 5 different values, and the *filling-value* variable gets 2 different values. So, there are $5 \times 16 \times 5 \times 2 = 800$ experiments which are performed on each dataset. Putting all these together, there are $12 \times 800 = 9600$ experiments done on 12 different datasets. We aim to find the most suitable values for these variables. The experimental results and the optimized variables are determined in section *VI*.

6. Experimental results

Here, we perform a statistical analysis to find the most effective values among different values of the independent variables *Descriptor-type*, *Aggregator*, *Recovery-method*, and the *missing-value*, which leads to a higher dependant variable, *CPCC*. In this analysis, the variables are called **factors**, and the values are called the **levels** of the factors. The *CPCC* is also called the **response variable**.

First, we use the analysis of variance (ANOVA) to identify the significant factors and interaction effects. ANOVA is a powerful tool which anticipates the variations of a dependant response variable under the experimental conditions determined by independent factors.

After ANOVA has been performed, a statistical design of experiment (DOE) is used to find the optimal condition which determines the best setting level of each factor involved in the experiment. In our problem, the optimal setting of each factor is defined as the one which leads to the highest *CPCC*.

6.1. Analysis of variances

Here, we perform a full factorial ANOVA analysis model on all 9600 observations. ANOVA examines the variation of the dependant response variable, for the presence of valuable independent variables.

First, we should specify the independent categorization variables (factors) and the values they take (levels). The factors and

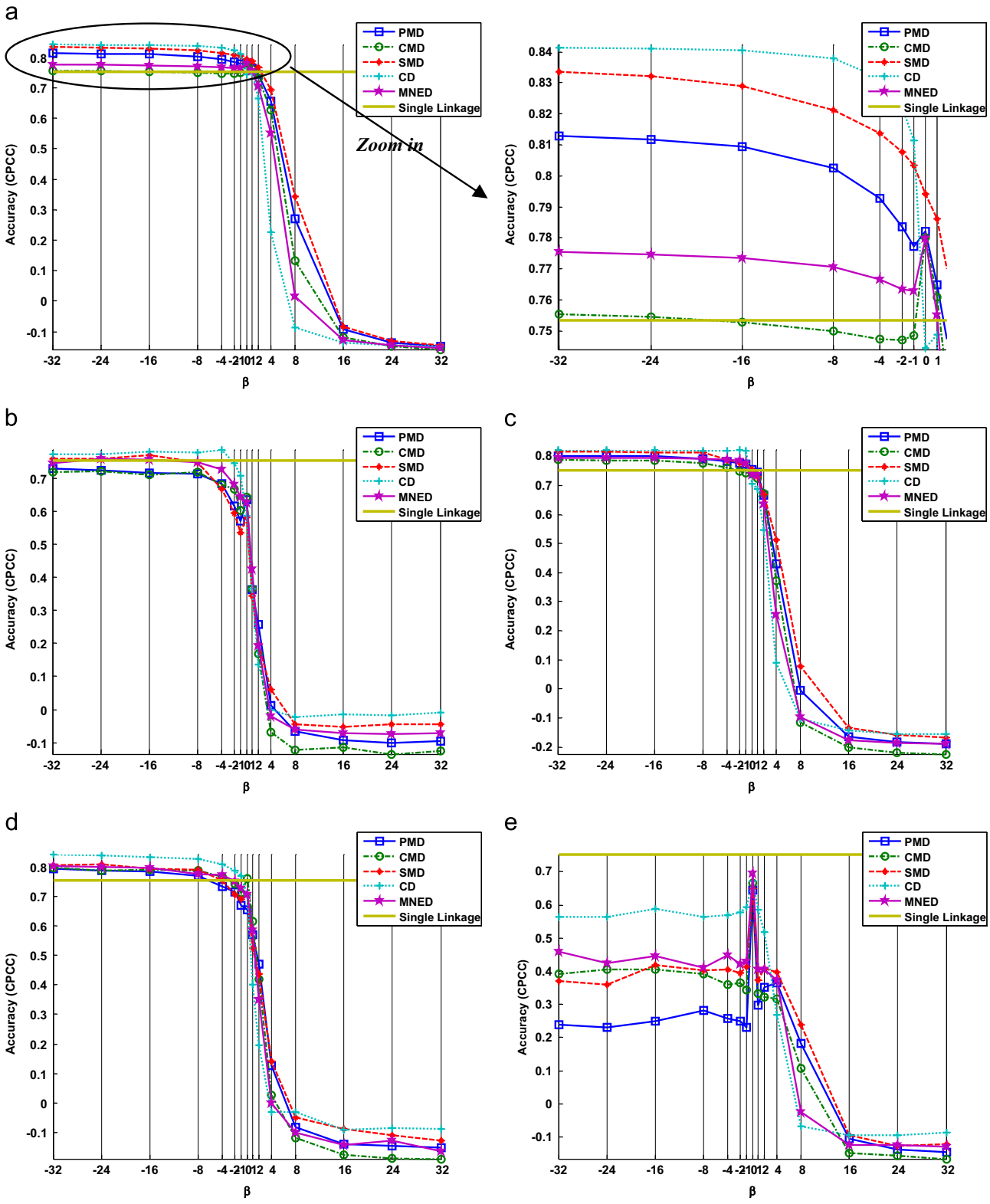

Fig. 3. The CPCC values of different Rényi aggregators applied to the Ionosphere dataset. In this Figure, the final hierarchy is recovered using standard hierarchical clustering algorithms: Single Linkage (a), Complete Linkage (b), UPGMA (c), WPGMA (d), and Ward (e). The performance of the standard Single Linkage clustering algorithm on the entire dataset in a none-ensemble based fashion is also shown in these plots with straight lines.

the factor levels are illustrated in Table 3. Then, we should specify the dependent response variable, which is the *CPCC* here.

Our ANOVA analysis contains a GLM (General Linear Model) procedure, an ANOVA model, and a Duncan post-hoc test. The result of the GLM performed on multiple factors is shown in Table 4. The result of ANOVA is shown in a standard Table 5, and the result of the Duncan test is shown in Table 6.

The GLM table (Table 4) includes six columns namely *Source*, *DF*, *Sum of Squares*, *Mean Square*, *f-value*, and *p*-value. The column *Source* stands for source of variation, *DF* column stands for the

Table 3
Factor level informations.

Factor	Level	Values
F1[a]	2	{0,1}
F2[b]	5	{Single, Complete, UPGMA, WPGMA, Ward}
F3[c]	5	{MNED, PMD, CMD, SMD, CD}
F4[d]	16	{β(−32), β(−24), β(−16), β(−8), β(−4), β(−2), β (−1), β(0), β(1), β(2), β(4), β(8), β(16), β(24), β(32), JSD}

[a] Fill the missing values.
[b] Recovery-method.
[c] Descriptor-type.
[d] Aggregator.

Table 4
The GLM Procedure.

Source	DF	Sum of squares	Mean square	*f*-value	*p*-value
Total	559	568.693	1.017	44.68	< **0.0001**
Error	9040	205.831	0.023		
Corrected total	9599	774.523			

Table 5
ANOVA performed on factors F1, F2, F3, F4.

Source	Degree of freedom	Type III SS	Mean square	*F*-value	*p*-value
Main effects					
F1	1	12.504	12.504	549.15	< **0.0001**
F2	4	8.675	2.169	95.26	< **0.0001**
F3	4	13.680	3.420	150.21	< **0.0001**
F4	15	423.601	28.240	1240.29	< **0.0001**
Two way interaction effects					
F1*F2	4	7.855	1.964	86.25	< **0.0001**
F1*F3	4	2.911	0.728	31.96	< **0.0001**
F1*F4	15	22.416	1.494	65.63	< **0.0001**
F2*F3	16	3.353	0.210	9.20	< **0.0001**
F2*F4	60	35.130	0.585	25.72	< **0.0001**
F3*F4	60	21.347	0.356	15.63	< **0.0001**
Three way interaction effects					
F1*F2*F3	16	1.573	0.098	4.32	< **0.0001**
F1*F2*F4	60	10.359	0.173	7.58	< **0.0001**
F1*F3*F4	60	2.565	0.043	1.88	< **0.0001**
F2*F3*F4	240	2.722	0.011	0.50	< **0.0001**

degrees of freedom, *f*-value column shows the division of mean square of the model by the mean square of the error, and *p*-value is the probability note which provide statistical significance [42]. The *p*-value which is resulted from the table is *p*-value < 0.0001; that, indicates the model is significant in explaining the variation of the *CPCC* at the significance level of 0.05 (α=0.05), that is, each *p*-value is "much" less than α=0.05. So, we say that the *p*-value in GLM table shows that the model is significant in explaining the variation of the *CPCC*.

According to calculated *p*-value shown in Table 5, it is observed that all parameters are significant in explaining the variation of the *CPCC*. It means that all factors are affected the response variable; and no factor can be omitted from the model. So, we continue the test with this model.

In the following, we perform Duncan's post-hoc test to obtain means of *CPCC* in multiple levels of *F1*, *F2*, *F3* and *F4*. Duncan is a type of multiple comparison procedure, which compares sets of means and determines the significant differences between means [43]. Multiple comparisons are used in the statistical analysis that includes a number of formal comparisons, with the attention focused on the strongest differences among all comparisons that are made. Table 6 shows that the strongest differences are caused by the main effects of F1={0}, F2={Single}, F3={*CD*}, F4={β(0)}.

Table 6
Duncan test for datasets D1 to D12.

Duncan grouping	Mean	**F1**
A	0.463	**0**
B	0.391	1
Duncan grouping	Mean	**F2**
A	0.469	**Single**
B	0.445	UPGMA
C	0.432	WPGMA
D	0.411	Complete
E	0.380	Ward
Duncan grouping	Mean	**F3**
A	0.491	**CD**
B	0.442	SMD
B	0.421	MNED
C	0.400	CMD
D	0.381	PMD
Duncan grouping	Mean	**F4**
A	0.646	β **(0)**
B	0.593	β (−32)
B	0.593	β (−16)
B	0.592	β (−24)
B	0.590	β (−8)
C B	0.582	β (−4)
C D	0.571	β (−2)
D	0.560	β (−1)
E	0.507	JSD
E	0.503	β (1)
F	0.439	β (2)
G	0.299	β (4)
H	0.148	β (8)
I	0.082	β (16)
J I	0.068	β (24)
J	0.064	β (32)

But the interaction effects are not considered here. So an optimal design is performed to find the values which interact together in order to put the strongest effect on *CPCC*.

6.2. Optimal design of experiment

The purpose of the optimal design of experiment in this study is to suggest a set of the factors' level which leads to the highest *CPCC*.

Performing the optimal DOE, the factor settings which yielded the highest *CPCC* in the experiment are conducted. According to the DOE test, the best factor settings are: F1={0}, F2={Single}, F3={*CD*}, F4={β(-32)}. It obvious that the results are intensively compatible with the previous Duncan test. The top ten optimal factors setting resulted from the test, are shown in Table 7.

Table 7
Top 10 optimal factor setting.

	F1	F2	F3	F4	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
1	**0**	**Single**	**CD**	**α(−32)**	0.735	0.850	0.838	0.841	0.827	0.868	0.923	0.707	0.942	0.909	0.765	0.622
2	0	Single	CD	α(−24)	0.735	0.851	0.838	0.841	0.827	0.867	0.922	0.7063	0.943	0.909	0.765	0.623
3	0	Single	CD	α(−16)	0.736	0.852	0.838	0.840	0.826	0.865	0.922	0.702	0.942	0.909	0.765	0.624
4	0	Single	CD	α(−8)	0.738	0.857	0.840	0.838	0.824	0.861	0.918	0.692	0.941	0.906	0.761	0.623
5	0	Single	CD	α(−4)	0.742	0.863	0.841	0.832	0.816	0.857	0.911	0.671	0.937	0.900	0.754	0.617
6	0	Single	CD	α(−2)	0.743	0.867	0.842	0.822	0.799	0.853	0.899	0.646	0.928	0.888	0.744	0.605
7	0	Single	SMD	α(−32)	0.735	0.813	0.885	0.834	0.705	0.690	0.735	0.666	0.759	0.690	0.732	0.645
8	0	Single	SMD	α(−24)	0.736	0.816	0.883	0.832	0.703	0.688	0.733	0.664	0.756	0.687	0.731	0.645
9	0	Single	SMD	α(−16)	0.735	0.818	0.879	0.829	0.699	0.683	0.729	0.658	0.751	0.684	0.729	0.645
10	0	Single	PMD	α(0)	0.695	0.796	0.780	0.782	0.7467	0.720	0.827	0.648	0.809	0.796	0.727	0.504

Table 8
The CPCC results of the proposed method in comparison with the *MATCH*.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	**Average**
Proposed *HCC*	**0.735**	**0.850**	**0.838**	0.841	0.827	**0.868**	0.923	0.707	**0.942**	**0.909**	0.765	0.622	0.820
MATCH	0.602	0.824	0.811	0.911	0.953	0.749	0.935	0.894	0.545	0.762	0.928	0.834	0.812
BobHic	0.6	0.83	0.805	0.91	0.905	0.736	0.895	0.911	0.796	0.762	0.856	0.8	0.817

7. Comparison of HCC approaches

Here, we compare the information-theory based aggregation approach with other *HCC* methods. We set the variables of proposed approach with the optimal factor settings F1={0}, F2={Single}, F3={CD}, F4={α(−32)} and compare the results. It should be mentioned that the combination method proposed in [25] is a *mean* aggregator based method. Therefore, it is a special case of our information-theory based aggregator with $\beta=1$ (see Table 1). So, the results are compared with *MATCH* and *BobHic*. The average *CPCC* obtained from the proposed method on 12 datasets is not significant from *BobHic* (significant level=0.05), and is higher than *MATCH* (see Table 8).

8. Conclusion

In this paper, we presented a framework for Hierarchical clustering combination problem. In this framework, each input dendrogram is represented by a dendrogram description matrix. Then the base clusterers descriptor matrices are combined by various aggregators where the aggregators are derived using information theoretic methods. This framework is a unified view of some of the existing algorithms, which leads to a class of new operators (adjustable by a parameter). We presented the mathematical derivation of the aggregation operators based on Rényi and Jensen-Shannon Divergences (RD, JSD) and used them to combine the description matrices. Following combination of description matrices into a consensus matrix, the final hierarchy is drawn by applying a hierarchical clustering algorithm to this matrix.

It can be noted that because the concept of combining dendrograms is new, it is difficult to give compelling evaluation of the proposed methods performance. However, because the output of each combination method is a hierarchy, we have used the cophenetic coefficient as an evaluation criterion. Experimental results also show that our proposed method, using *CD* descriptor matrices, Single linkage method, and Rényi aggregator with $\alpha=-32$ leads to a good degree of a consensus clustering quality. And the comparisons show that the average *CPCC* obtained from the proposed method on 12 datasets is not significant from *BobHic*, and is higher than *MATCH*.

Appendix

In the following two appendices we present how the combination of description matrices can be derived using Rényi Divergence and Jensen–Shannon divergence.

8.1. A. Rényi divergence

In order to minimize the function given in Eq. (8), we need to take its derivative with respect to p^*_{im} and set it to zero, that is $(\partial/\partial p^*_{im})D'(i)=0$. The derivative of the Eq. (8) with respect to the p^*_{im} is:

$$\frac{\partial}{\partial p^*_{im}}D'(i)=\frac{1}{L(\alpha-1)}\sum_{k=1}^{L}\frac{\partial}{\partial p^*_{im}}\left[\log\sum_{j=1}^{n}(p^*_{ij})^{\alpha}(p^k_{ij})^{1-\alpha}\right]-\lambda$$
$$=\frac{\alpha}{L(\alpha-1)\ln 2}\sum_{k=1}^{L}\frac{(p^*_{im})^{\alpha-1}(p^k_{im})^{1-\alpha}}{\sum_{j=1}^{n}(p^*_{ij})^{\alpha}(p^k_{ij})^{1-\alpha}}-\lambda. \quad (A.1)$$

If Eq. (A.1) is solved for p^*_{im}, the value of each p^*_{im} is dependent on the values of other $p^*_{im'}$, $m'\neq m$. This equation is simplified by assuming that $\sum_{j=1}^{n}(p^*_{ij})^{\alpha}(p^k_{ij})^{1-\alpha}=1$ for all $k=1,\ldots L$. In this case, the value of each p^*_{im} can be calculated separately from the other $p^*_{im'}$, $m'\neq m$. Thus, Eq. (A.1) becomes

$$\frac{\partial}{\partial p^*_{im}}D'(i)=\frac{\alpha(p^*_{im})^{\alpha-1}}{L(\alpha-1)\ln 2}\sum_{k=1}^{L}(p^k_{im})^{1-\alpha}-\lambda. \quad (A.2)$$

Solving for p^*_{im}, the following is obtained,

$$\frac{\partial}{\partial p^*_{im}}D'(i)=0\Rightarrow P^*_{im}=\left[\frac{\lambda L(\alpha-1)\ln 2}{\alpha\sum_{k=1}^{L}(p^k_{im})^{1-\alpha}}\right]^{\frac{1}{\alpha-1}}. \quad (A.3)$$

Substituting Eq. (A.3) into $\sum_{j=1}^{n}p^*_{ij}=1$, the λ may be determined from

$$\lambda^{1/\alpha-1}=\left(\frac{\alpha}{L(\alpha-1)\ln 2}\right)^{1/\alpha-1}\sum_{j=1}^{n}\left[\frac{1}{\sum_{k=1}^{L}(P^k_{ij})^{1-\alpha}}\right]^{1/1-\alpha}. \quad (A.4)$$

If $\lambda^{1/\alpha-1}$ is substituted into Eq. (A.3), the final formula for calculating p^*_{im} is

$$p^*_{im}=\frac{1}{\sum_{j=1}^{n}\left(\sum_{k=1}^{L}(p^k_{ij})^{1-\alpha}\right)^{\frac{1}{1-\alpha}}}\left(\sum_{k=1}^{L}(p^k_{im})^{1-\alpha}\right)^{1/1-\alpha}. \tag{A.5}$$

To simplify this, we denote the term $\sum_{j=1}^{n}\left(\sum_{k=1}^{L}(p^k_{ij})^{1-\alpha}\right)^{1/1-\alpha}$ by r, then Eq. (A.5) simplifies to

$$p^*_{im}=\frac{1}{r}\left(\sum_{k=1}^{L}(p^k_{im})^{1-\alpha}\right)^{1/1-\alpha}, \tag{A.6}$$

where r is a constant which causes that the values of each row of final description matrix sum to 1.

8.2. B. Jensen–Shannon divergence

To find consensus matrix based on Jensen–Shannon divergence, we require that $\frac{\partial}{\partial p^*_{im}}D(i)=0$, where $D(i)$ is defined according to Eq. (13). The derivative of the Eq. (13) with respect to the p^*_{im} is

$$\begin{aligned}\frac{\partial}{\partial p^*_{im}}D(i)&=\frac{1}{2L}\sum_{k=1}^{L}\left(\log p^*_{im}-\log\left(\frac{p^k_{im}+p^*_{im}}{2}\right)\right)\\&=\frac{1}{2\ln 2L}\sum_{k=1}^{L}(\ln p^*_{im}-\ln(p^k_{im}+p^*_{im}))+\frac{1}{2}.\end{aligned} \tag{B.1}$$

Using the $(\partial/\partial p^*_{im})D(i)=0$ gives

$$\frac{1}{\ln 2L}\sum_{k=1}^{L}(\ln p^*_{im}-\ln(p^k_{im}+p^*_{im}))+1=0 \tag{B.2}$$

or

$$\sum_{k=1}^{L}\left(\ln\left(1+\frac{p^k_{im}}{p^*_{im}}\right)\right)=\ln 2^L \tag{B.3}$$

It can be rewritten as

$$\prod_{k=1}^{L}(p^*_{im}+p^k_{im})=(p^*_{im})^L 2^L \tag{B.4}$$

Eq. (B.4) simplifies into the following polynomial form

$$\begin{aligned}&(1-2^L)(p^*_{im})^L+\left(\sum_{k_1=1}^{L}p^{k_1}_{im}\right)(p^*_{im})^{L-1}\\&+\frac{1}{2!}\left(\sum_{k_1=1}^{L}\sum_{\substack{k_2=1\\k_1\neq k_2}}^{L}p^{k_1}_{im}p^{k_2}_{im}\right)(p^*_{im})^{L-2}\\&+\frac{1}{3!}\sum_{k_1=1}^{L}\sum_{\substack{k_2=1\\k_1\neq k_2\neq k_3}}^{L}\sum_{k_3=1}^{L}(p^{k_1}_{im}p^{k_2}_{im}p^{k_3}_{im})(p^*_{im})^{L-3}\\&+\ldots+\frac{1}{L!}\sum_{k_1=1}^{L}\sum_{\substack{k_2=1\\k_1\neq k_2\neq\cdots\neq k_L}}^{L}\cdots\sum_{k_L=1}^{L}(p^{k_1}_{im}p^{k_2}_{im}\cdots p^{k_L}_{im})=0.\end{aligned} \tag{B.5}$$

where p^*_{im} is the polynomial root which satisfies $0\le p^*_{im}\le 1$. The roots of this polynomial can be obtained by finding the eigenvalues of a so called *companion matrix*. In other words, if we create the matrix A with the form of

$$A=\begin{bmatrix}-a_L & -a_{L-1} & \cdots & -a_2 & -a_1\\ 1 & 0 & \cdots & 0 & 0\\ 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & 0\end{bmatrix}, \tag{B.6}$$

where

$$\begin{aligned}a_1&=\frac{1}{(1-2^L)}\frac{1}{L!}\sum_{k_1=1}^{L}\sum_{\substack{k_2=1\\k_1\neq k_2\neq\cdots\neq k_L}}^{L}\cdots\sum_{k_L=1}^{L}(p^{k_1}_{im}p^{k_2}_{im}\cdots p^{k_L}_{im}),\\&\vdots\\a_{L-2}&=\frac{1}{(1-2^L)}\frac{1}{3!}\sum_{k_1=1}^{L}\sum_{\substack{k_2=1\\k_1\neq k_2\neq k_3}}^{L}\sum_{k_3=1}^{L}(p^{k_1}_{im}p^{k_2}_{im}p^{k_3}_{im}),\\a_{L-1}&=\frac{1}{(1-2^L)}\frac{1}{2!}\sum_{k_1=1}^{L}\sum_{\substack{k_2=1\\k_1\neq k_2}}^{L}(p^{k_1}_{im}p^{k_2}_{im}),\\a_L&=\frac{1}{(1-2^L)}\sum_{k_1=1}^{L}(p^{k_1}_{im}),\end{aligned} \tag{B.7}$$

then the eigenvalues of this matrix will be equal to the roots of polynomial (B.5). The eigenvalues of companion matrix can be found by an iterative QR algorithm [44].

References

[1] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood, Cliffs, N.J., 1988.
[2] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (1999) 264–323.
[3] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, John Wiley & Sons, USA, 2001.
[4] A.R. Web, Statistical Pattern Recognition, John Wiley & Sons, UK, 2002.
[5] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 2nd ed., Elsevier Academic Press, USA, 2003.
[6] A. Weingessel, E. Dimitriadou, K. Hornik, An Ensemble Method for Clustering, in: Presented at the DSC Working Papers, 2003.
[7] Z.H. Zhou, W. Tang, Clusterer ensemble, Knowledge-Based Systems 19 (2006) 77–83.
[8] A. Topchy, B. Minaei-Bidgoli, A.K. Jain, W.F. Punch, Adaptive Clustering Ensembles, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004, pp. 272–275.
[9] A. Fred, A.K. Jain, Combining multiple clusterings using evidence accumulation, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 835–850.
[10] H.G. Ayad, M.S. Kamel, On voting-based consensus of cluster ensembles, Pattern Recognit. 43 (2010) 1943–1953.
[11] A. Fred, Finding consistent clusters in data partitions, Multiple Classifier Syst. (2001) 309–318.
[12] J. Chang, D.M. Blei, Mixtures of Clusterings by Boosting, in: Learning Workshop, Hilton Clearwater, 2009.
[13] B.V. Dasarathy, A special issue on applications of ensemble methods, Inf. Fusion 9 (2008) 1.
[14] D. Frossyniotis, A. Likas, A. Stafylopatis, A clustering method based on boosting, Pattern Recognit. Lett. 25 (2004) 641–654.
[15] D. Frossyniotis, M. Pertselakis, A. Stafylopatis, A multi-clustering fusion algorithm, Methods Appl. Artif. Intell. 2308 (2002) 225–236.
[16] R. Ghaemi, M.N. Sulaiman, H. Ibrahim, N. Mustapha, A survey: clustering ensembles techniques, World Academy Sci., Eng. Technol. 38 (2009).
[17] A. Gionis, H. Mannila, P. Tsaparas, Clustering aggregation, ACM Trans. Knowl. Discov. Data (TKDD) 1 (2007) 1–4.
[18] Y. Hong, S. Kwong, H. Wang, Q.S. Ren, Resampling-based selective clustering ensembles, Pattern Recognit. Lett. 30 (2009) 298–305.
[19] R. Maclin, D. Opitz, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. 11 (2011) 169–198.
[20] B. Minaei-Bidgoli, A. Topchy, W.F. Punch, A comparison of resampling methods for clustering ensembles, in: Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, 2004, pp. 939–945.
[21] B. Minaei-bidgoli, E. Topchy, W.F. Punch, Ensembles of partitions via data resampling, in: Proceedings of the International Conference on Information Technology, Washington, DC, USA 2004, pp. 188–199.
[22] S. Vega-pons, J. Ruiz-shulcloper, A survey of clustering ensemble algorithms, Int. J. Pattern Recognit. Artif. Intell. 25 (2011) 333–372.
[23] M. Grabisch, J.L. Marichal, R. Mesiar, E. Pap, Aggregation Functions (Encyclopedia of Mathematics and its Applications), Cambridge University Press, UK, 2009.
[24] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, John Wiley & Sons, Canada, 2004.
[25] A. Mirzaei, M. Rahmati, M. Ahmadi, A new method for hierarchical clustering combination, Intell. Data Anal. 12 (2008) 549–571.
[26] A. Mirzaei, M. Rahmati, A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations, IEEE Trans. Fuzzy Syst. 18 (2010) 27–39.
[27] L. Zheng, T. Li, C. Ding, Hierarchical ensemble clustering, in: Proceedings of the 10th International Conference on Data Mining (ICDM), 2010, pp. 1199–1204.

[28] E. Rashedi, A. Mirzaei, A hierarchical clusterer ensemble method based on boosting theory, Knowledge-Based Syst. 45 (2013) 83–93.
[29] H. Drucker, C. Cortes, Boosting decision trees 8 (1996) 479–485Adv. Neural Inf. Syst. 8 (1996) 479–485.
[30] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32.
[31] B.H. Park, Constructing simpler decision trees from ensemble models using Fourier analysis, in: Proceedings of the 7th WorkShop Research Issues in Data Mining and Knowledge Discovery, 2002, pp. 18–23.
[32] H. Kargupta, B.H. Park, A Fourier spectrum-based approach to represent decision trees for mining data streams in mobile environments, IEEE Trans. Knowl. Data Eng. 16 (2002) 216–229.
[33] H. Kargupta, B.H. Park, H. Dutta, Orthogonal decision trees, IEEE Trans. Knowl. Data Eng. 18 (2006) 1028–1042.
[34] J. Podani, simulation of random dendrograms and comparison tests: some comments, J. Classif. 17 (2000) 123–142.
[35] A. Rényi, On measures of information and entropy, in: Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, 1960, pp. 547–561.
[36] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans Inf. Theory 37 (1991) 145–151.
[37] E. Gose, R. Johnsonbausg, S. Jost, Pattern Recognition and Image Analysis, Prentice Hall, USA, 1996.
[38] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423.
[39] C. Blake, C.J. Merz, UCI Repository of machine learning databases, Department of Information and Computer Science, vol. 55, University of California, Irvine, CA, 1998.
[40] L. Kuncheva, Real Medical Data Sets, Technical Report, School of Informatics: University of Wales, Bangor, UK, 2005.
[41] E. Rashedi, A. Mirzaei, A novel multi-clustering method for hierarchical clusterings based on boosting, in: Proceedings of the 9th Iranian Conference on Electrical Engineering (ICEE), 2011, pp. 1–4.
[42] D.A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press, USA, 2005.
[43] D.B. Duncan, Multiple range and multiple f tests, Biometrics 11 (1995) 1–42.
[44] R.J. Schilling, S.L. Harris, Applied Numerical Methods for Engineers Using MATLAB and C, Brooks/Cole Publishing Company, 2000.

Elaheh Rashedi received the Bachelor degree from University of Tehran, Tehran, Iran, in 2008 and the Master degree in electrical and computer engineering from the Isfahan University of Technology, Isfahan, Iran, in 2011. She is currently a Ph.D. student in Computer Science department at Wayne State University. Her current research interests include pattern recognition, multiple classifier systems, data fusion and learning methods.

Abdolreza Mirzaei was born in Isfahan, Iran. He received the B.S. (first-class honors) degree in computer engineering from Isfahan University, in 2001, the M.Sc. degree in artificial intelligence from Iran University of Science and Technology, Tehran, Iran, in 2003, and the Ph.D. degree in artificial intelligence from Amirkabir University of Technology, Tehran, in 2009, respectively. He is currently with the Department of Electrical and Computer Engineering, Isfahan University of Technology. His research interests include statistical and structural classification methods, digital image processing, computer vision, multiple classifier systems, and learning methods.

Mohammad Rahmati received his Ph.D. degree in electrical and computer engineering from the University of Kentucky, Lexington, KY, in 1993. He is currently an Associate Professor with the Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran. His research interests include pattern recognition, image and video processing, computer vision, and data mining. Dr. Rahmati is a member of the IEEE Signal Processing Society.