A Hierarchical Clusterer Ensemble Method Based on Boosting Theory

Elaheh Rashedi\textsuperscript{1}, Abdolreza Mirzaei\textsuperscript{2}

\textsuperscript{1} ECE Department, Isfahan University of Technology, Isfahan, Iran, elahehrashedi@gmail.com
\textsuperscript{2} ECE Department, Isfahan University of Technology, Isfahan, Iran, mirzaei@cc.iut.ac.ir

Abstract

Bagging and boosting are two successful well-known methods for developing classifier ensembles. It is recognized that the clusterer ensemble methods which utilize the boosting concept, can create clusterings with quality and robustness improvement. In this paper, we introduce a new boosting based hierarchical clusterer ensemble method called \textit{Bob-Hic}. This method is utilized to create a consensus hierarchical clustering (h-clustering) on a dataset which is helpful to improve the clustering accuracy. \textit{Bob-Hic} includes several boosting iterations, and in each iteration, first a weighted random sampling is performed on the original dataset and then an individual h-clustering is created on the selected samples. At the end of the iterations, the individual clusterings are combined to a final consensus h-clustering. The middle structures used in the combination are distance descriptor matrices which correspond to individual h-clustering results. This final integration is done through an information theoretic approach. Experiments on both synthetic and real popular datasets confirm that the proposed method improves the results of simple clustering algorithms. This method provides better consensus clustering quality in comparison to other available ensemble techniques.

Keywords: hierarchical clustering; multi-clusterer; ensemble; boosting.

1. Introduction

The idea of ensemble learning is to combine multiple learners’ predictions. In the area of supervised (classification) and unsupervised (clustering) learning algorithms, ensembles often lead to better results in comparison to single solutions. Classifier ensemble methods combine classifiers to achieve a classification solution of a higher predictive accuracy [1]. Similarly, clusterer ensemble methods improve quality of clusterings by aggregating clusterers. The most recent general ensemble methods that can be used to reduce errors, in both classification and clustering cases, are bagging and boosting algorithms [2-4].

The general idea of bagging, which is shortened form of the words “bootstrap aggregating”, is to generate an ensemble of learners built on bootstrap replicates of the training set and also to combine learners outputs [5]. Very similar to bagging, boosting is defined as a general method of machine learning which converts a weak learning algorithm into one with higher accuracy [5]. Also, boosting is one of the most powerful methods for creating classifier ensembles. There are also some bagging and boosting based multi clustering algorithms introduced on “partitional” clusterings [6-7]. Accordingly, there is a
potential of “hierarchical” clustering (h-clustering) quality improvement through using multi hierarchical clustering methods.

In this paper a boosting based hierarchical clusterer ensemble technique, Bob-Hic, is proposed, which is a new method for the framework previously presented by the same authors in [8]. The new method provides solutions of better quality compared to other h-clustering fusion methods.

This paper is organized as follows: The related works done on classification and partitional clusterings are demonstrated in section 2, followed by introduction to h-clustering combination approaches. In section 3, a new boosting based multi hierarchical clustering algorithm is presented. The experimental results of the proposed method on real datasets and synthetic datasets are discussed in sections 4 and 5. The comparative results between the proposed method and both of single and ensemble h-clustering methods are also illustrated in these sections. Finally, our concluding remarks are presented in Section 6.

2. Related works

There are four approaches to build and combine classifiers: i) methods which use different data subsets, ii) methods which use different feature subsets, iii) methods which apply different basic classifiers and, iv) methods which utilize different combiners to aggregate ensembles [5]. The quality of a classifier ensemble depends on both how the classifier ensembles are generated and how they are combined.

Numerous methods have been introduced to generate and integrate classifier ensembles. An extensive body of work already done on bagging and boosting, and their advantages over other classifier ensemble methods are confirmed [3-4, 9-12]. Various hierarchical classification ensemble methods are also proposed combining information of several labeled trees into one single representative tree [13-14]. Also, these techniques are widely used in phylogenetics [15-16]. A brief explanation of these methods is given in section 2.1.

In the area of clustering methods, various techniques are proposed to create clusterer ensembles [17-21], and some of them are based on bagging and boosting [3, 22-23]. These methods usually use partitional (non-hierarchical) clustering algorithms to create basic clusterings [7]. A brief explanation of these methods is presented in section 2.2.

The last point to be addressed is h-clusterer ensemble methods that can be used to combine a set of h-clusterings. We add a review of them in section 2.3.

2.1. Classifier ensemble methods based on bagging and boosting

Using the best of both classifier ensemble generation method and ensemble combination method leads a more accurate classification decision. However, this approach increases the complexity. The most popular ensemble methods utilized for reducing classification errors are bagging and boosting [5, 9, 24]. Bagging is a multiple learner combination method which always uses resampling. First, bootstrap replications of data points are created where the
sampling is based on the random selection, in which the samples are selected uniformly. Then, individual copies of the weak learner algorithm are applied on samples, and finally the results are combined using majority voting. It can be shown that for unstable classifiers, bagging leads to performance improvement [5, 9, 25].

Very similar to bagging, boosting is an ensemble method used to improve the accuracy of any learning algorithm [26]. The main difference is the use of reweighting. In boosting, the weights of data points are updated based on their learning complexity while in bagging the weights remain unchanged. A widely used adaptive boosting algorithm called AdaBoost is used to reduce error and boost any weak learner’s performance [3, 11, 24, 26].

2.2. Partitional clusterer ensemble methods

There are many consensus functions introduced in partitional clusterer ensembles area, which use several mathematical and computational tools [7]. Some of the functions are as follows: relabeling and voting [27-31], fuzzy clustering [32], genetic algorithms [33], Co-association matrix [18], graph and hyper-graph (CSPA, HPGA, MCLA) [34], Mirkin distance [35], information theory [36], finite mixture models [19, 37], locally adaptive clustering algorithm [38], Kernel [7], and non-negative matrix factorization [39].

These techniques mostly improve the result of single clustering algorithms. The main target of these methods is to find consistent clusters through clustering combination. In the combination step the information from all individual partitions of the ensemble are integrated and possible errors of single clustering methods are rectified. So the final combined clustering represents a better solution.

Similar to classification problems, constructing clusterer ensembles using bagging and boosting have a potential of clustering accuracy improvement, and this method can also provide more robust results [10, 27].

A clusterer ensemble method based on boosting, boost-clustering, is introduced in [1]. In this ensemble method, a simple partitional clusterer, e.g. k-means, is utilized to create multiple clustering results and then the clustering results are combined through weighted voting. Also, there are some resampling based combination solutions in which a selection of the clustering results are combined [17, 22-23].

Methods which are introduced to combine h-clusterings are described in next subsection, 2.3 [40-44].

2.3. Hierarchical clusterer ensemble methods

In h-clustering methods, data points are organized into a nested sequence of groups instead of simple partitions. These techniques are grouped into two categories, agglomerative and divisive. Some popular agglomerative h-clustering methods are Centroid Linkage, Single Linkage, Average Linkage, Complete Linkage, Weighted Linkage, Median Linkage and Ward Linkage. These methods construct the hierarchy by recursively merging the two clusters with minimum defined distance values.
Hierarchies are illustrated using a tree diagram called dendrogram listing all data points and indicating at which level of similarity any two clusters are joined. Accordingly, an h-clusterer ensemble method is defined as a technique combining the basic dendrograms. The dendrograms can be represented in the form of similarity or dissimilarity matrices where these matrices are called descriptors [44]. Some descriptors are Cophenetic Difference (CD), Partition Membership Divergence (PMD), Cluster Membership Divergence (CMD), Sub-tree Membership Divergence (SMD) and Path Difference (PD) [45]. The descriptor used in our experiments is CD. In CD descriptors, the distance between two data points is defined as the lowest level of the hierarchy where these pairs are joined together. According to this new definition, h-clusterer combination method is defined as a technique combining descriptors corresponding to basic dendrograms.

An h-clusterer combination method, HCC, is introduced in [44]. In HCC, different dissimilarity matrices are combined into one aggregated matrix by using matrix summation operator. The aggregated matrix might not necessarily have an associated dendrogram. Therefore, a dendrogram recovery phase is then applied to derive the final dendrogram from aggregated dissimilarity matrix. The recovery process uses the maximum transitive similarity property to derive the final dendrogram. The recovery phase has great effect on the quality of the final dendrogram [41, 43].

Another combination method based on fuzzy similarity relations, MATCH, is proposed in [41, 43]. In this method, descriptors are aggregated into one transitive consensus matrix from which the final dendrogram can be directly formed with no need to the additional recovery phase. By skipping over the recovery phase, the chaining effects are excluded [41, 43]. Some experiments show that MATCH is a better solution in comparison to HCC in terms of quality [44]. This quality is achieved in price of consuming more time. It should be mentioned that the time complexity of MATCH is $O(Ln^4)$, where $L$ is the ensemble size and $n$ is the number of data points.

Hierarchical ensemble clustering technique, HEC, is introduced in [46] combining partitional and hierarchical clusterings into one consensus h-clustering. HEC and MATCH are dual algorithms; the former generates minimum transitive dissimilarity matrix closure, and the latter generates maximum transitive similarity matrix closure.

An information-theory-based combination method is presented in [40]. In this method, each row of the basic descriptor matrix is supposed to be a probability distribution function, PDF. Accordingly, each row in the consensus matrix is calculated as a PDF with the most similarity to all PDFs associated to the same row in basic descriptors. The Rényi divergence is utilized as a measure to calculate the PDF values of the consensus matrix. The Rényi divergence is a family member of functions which quantify the diversity of a system and can be used to calculate the dissimilarity of two PDFs.

Let $P^*$ to be the consensus matrix, which stands for the nearest PDF to all basic PDF matrices $P^i$. Based on Rényi of order $\alpha$, $P^*$ is calculated as:
\[ p_{m,n}^* = \frac{1}{r} \left( \sum_{i=1}^{L} (p_{m,n}^i)^{1-\alpha} \right)^{\frac{1}{1-\alpha}} \]  

(1)

in which, \(L\) is the number of basic clusterings in the ensemble and \(p_{m,n}^i\) is the probability value placed in the \(m^{th}\) row and \(n^{th}\) column of the \(i^{th}\) basic PDF matrix. Similarly, \(p_{m,n}^*\) is the probability value placed in \(m^{th}\) row and \(n^{th}\) column of the consensus PDF matrix. Finally, \(r\) is the normalization constant. Setting \(\alpha\) to a fixed value, the consensus PDF matrix is calculated and is recovered to generate the final dendrogram [40]. HCC can be viewed as a special case of the information theory based combination method in which the Rényi divergence with \(\alpha = 0\) is used.

In this paper, the information theory based combination method is used to aggregate multiple descriptors. The proposed method is compared to MATCH.

3. The Boosting Based Hierarchical Clusterer Ensemble Method: Bob-Hic

In this section we propose a new clusterer ensemble method for h-clusterings. This is a general method in which any h-clustering algorithm can be used in ensemble generation and any hierarchical combination method can be used in individual clusterings aggregation. The algorithm is given in section 3.1, and time complexity analysis is discussed in section 3.2.

3.1. Bob-Hic algorithm

The steps of this algorithm are as follows: a) at the first iteration of boosting, a new sample set is provided by random sampling (without replacement) from the original dataset, b) the h-clustering algorithm is applied on the selected samples to create the first hierarchical clustering of data points, c) each data point’s weight is updated based on the efficacy of the previous hierarchical clusterings built on it, and the next basic single clustering is generated according to the new weights and d) the final clustering solution is produced by combining all hierarchies in the ensemble.

Here, standard h-clustering methods like Centroid, Single, Average, Complete, Weighted, Median and Ward Linkages are used for creating basic clusterings. Then, the similarity descriptors of basic single clustering results are combined into one descriptor. The combination stage is done based on the Rényi divergence approach. The pseudo code of the boosted combination algorithm is illustrated in figure 1.

The main framework of the boosted h-clusterer ensemble method is based on the general boosting algorithm, arc-x4 [5]. The algorithm starts to generate a consensus h-clustering on a data set of \(N\) data points, \(D=(x_1,x_2, \ldots , x_N)\).
**Bob-Hic:** A data set of $N$ data points, $D=(x_1,x_2, \ldots, x_N)$ is given. The Output is a consensus h-clustering, $H^*$. 

1. **Initialization** 
   
   $i=1$
   
   Initialize $T$, $L$
   
   $w_n^i = \frac{1}{N} \quad 1 \leq n \leq N$
   
   Choose $H_{clustering}$.
   
   Choose $H_{combiner}$.

2. **Iteration**

   2.1. **Sampling**
   
   $D' = Sampler(D, w^i)$
   
   (2)

   2.2. **Hierarchical clustering**

   $H' = H_{clustering}(D')$

   (3)

   2.3. **Clustering aggregation**

   $H_{agg}^i = H_{combiner}(H', H_{agg}^{i-1})$

   (4)

   2.4. **Calculating the boosting values**

   $BV_n^i = H_{quality}(x_n, H_{agg}^i) \quad 1 \leq n \leq N$

   (5)

   2.5. **Updating the weights**

   $reverseBV_n^i = 1 - \frac{1}{i} \sum_{n=1}^{i} BV_n^{ii}$

   (6)

   $loss = \frac{1}{2} \sum_{n=1}^{N} (w_n^i \times reverseBV_n^i)$

   (7)

   $\beta = (1 - loss)/loss$

   (8)

   $w_n^{i+1} = \frac{w_n^n \beta^{reverseBV_n^i}}{Z^i} \quad 1 \leq n \leq N$

   (9)

2.6. **Repeating**

   If $i \leq T$, go to step 2.

3. **Obtaining final consensus h-clustering**

   $H^* = H_{agg}^T$

   (10)

**Figure 1:** Pseudo code of the boosted h-clustering combination algorithm

**Step 1, initializing the variables:** At the beginning of the algorithm, some initializations are performed as follows: The iteration number, $i$, is set to 1, and the maximum number of iterations, $T$, is set to a predefined fixed number. Also, the size of the ensemble is denoted by $L$. $L$ is equal to $T$, as in each iteration one individual clustering of the ensemble is created. Then, a uniform distribution is put on the data points by assigning $1/N$ as the initial weight of each data point. Thereafter, a basic h-clustering algorithm, $H_{clustering}$, is selected for generating the basic single h-clusterings. Finally, an h-clustering combination method, $H_{combiner}$, is chosen to aggregate single h-clusterings built in each iteration into a one representative h-clustering result.
Step 2, performing the iterations: This step is an iterative process which is supposed to be done for $T$ times.

Step 2.1, generating sample sets: In this step, the boosting stage is iteratively done as follows: a sample set of the $i$th iteration, $D^i$, is prepared using the weighted sampling approach, in which the samples are selected randomly according to the distribution of data points. This distribution is in relation to the weights assigned to each data point (see eq. (2)). The pseudo-code of the Sampler algorithm used in eq. (2) is illustrated in figure 2. The Sampler is based on the weighted roulette wheel selection algorithm [47].

**Sampler**: A data set of $N$ data points, $D=(x_1, x_2, \ldots, x_N)$, and their corresponding weights, $w=(w_1, w_2, \ldots, w_N)$ are given. The output is the sample set $D'=(x'_1, x'_2, \ldots, x'_K)$.

1. Create a local copy of $w$
   for $n=1$ to $N$ do
   $\text{copy}w_n = w_n$
   end
2. Sample $K$ point out of the data set $D$
   for $k=1$ to $K$ do
   $s_0 = 0$
   for $n=1$ to $N$ do
   $s_n = s_{n-1} + \text{copy}w_n$
   end
   $r = \text{random}(0, s_n)$
   $x'_k = x_i$ such that $s_{i-1} \leq r < s_i$ and $\text{copy}w_i \neq 0$ ($x_i$ is not selected before)
   $\text{copy}w_k = 0$
   end
3. return $D'=(x'_1, x'_2, \ldots, x'_K)$

Figure 2. Pseudo code of the Sampler algorithm

Step 2.2, creating hierarchical clustering: In this step, one basic h-clustering is generated by applying basic h-clustering algorithm to $D'$. The single clustering generated in the $i$th iteration is denoted as $H^i$ (eq. (3)).

Step 2.3, combining hierarchical clusterings: $H_{\text{agg}}^i$ is the aggregated clustering obtained from the $i$th iteration (eq. (4)). At the starting point, this value is equal to $H^i$, and in the next iterations, it is equal to the combination of $H_{\text{agg}}^{i-1}$ and $H^i$. Where $H_{\text{agg}}^{i-1}$ is the aggregated clustering obtained from the previous iterations, and $H^i$ is the newly generated clustering.

In this step, $H^i$ and $H_{\text{agg}}^{i-1}$ are aggregated by applying the h-clustering combination method, $H_{\text{combiner}}$, and $H_{\text{agg}}^i$ is created.
As discussed in section 2.3, combination methods often perform the aggregation process by using descriptors instead of direct dendrograms. Hence, eq. (4) is redefined as eq. (11), in which $f$ is a conversion function creating a similarity matrix $M$ from the hierarchy $H$, and $f^{-1}$ is a reverse function recovering the hierarchy from the similarity matrix.

$$H_{agg}^{i} = f^{-1}(M_{comb}(f(H_{agg}^{i-1}), f(H^{i})))$$  \hspace{1cm} (11)$$

In eq. (11), $M_{comb}$ is an element wise matrix combination function based on the Rényi divergence approach introduced in eq. (1) (with $1 - \alpha$ equals to 1). $H_{agg}^{i}$ is generated by applying eq. (11).

**Step 2.4, calculating boosting values:** In this step, the h-clustering quality is calculated in order to update each data points' weight. The most challenging problem in h-clusterer ensembles is finding a measurement concept which predicts clustering quality of each instance; because it’s hard to detect how well a data point is clustered in a hierarchy. Here, the quality of each data point $x_n$ is calculated by the boosting value, $BV_n \ [5]$, where $BV_n$ takes high values when the data point $x_n$ has been well-clustered in the aggregated hierarchy $H_{agg}^{i}$ and low values when the clustering quality of the data point $x_n$ in the hierarchy is unpleasant (eq. (5)).

In our framework, $BV$ is measured by a comparison between the hierarchical distances of the data points (i.e. descriptors) and Euclidean distances of the original data set. Accordingly, we formulate the $BV$ of the data point $x_n$ in the $i^{th}$ iteration, $BV_n^{i}$, as a modulus correlation coefficient between these two distance sets:

1. Hierarchical distances of the sample $x_n$ from other data points in $H_{agg}^{i}$, i.e. the $n^{th}$ row of the descriptor of $H_{agg}^{i}$ (denoted by $Hd_n^{i}$),

2. And the Euclidian distances of the sample $x_n$ from other data points in the original data set, i.e. the $n^{th}$ row of Euclidean distance matrix, (denoted by $Ed_n$).

So the $BV_n^{i}$ in eq. (5) will be redefined as eq. (12):

$$BV_n^{i} = \left| \frac{\sum_{c=1}^{N} (Hd_{n,c}^{i} - \frac{1}{N} \sum_{c=1}^{N} Hd_{n,c}^{i}) (Ed_{n,c} - \frac{1}{N} \sum_{c=1}^{N} Ed_{n,c})}{\sqrt{\sum_{c=1}^{N} (Hd_{n,c}^{i} - \frac{1}{N} \sum_{c=1}^{N} Hd_{n,c}^{i})^2} \sum_{c=1}^{N} (Ed_{n,c} - \frac{1}{N} \sum_{c=1}^{N} Ed_{n,c})^2} \right|  \hspace{1cm} (12)$$
The BV takes value from [0,1] interval. According to eq. (12), better quality samples get higher boosting value, which is caused by high correlation between Euclidean distance matrix and h-clustering descriptor.

Step 2.5, updating the weights: In eq. (12), we consider the aggregated clustering in order to calculate the clustering quality. This approach is used in arc-x4 boosting algorithm [5]. Following this approach, we update each sample’s weight, where these weights have to be changed so that the instances which are harder to cluster get higher probability of being selected in the next iterations. Similarly, the instances which are well clustered should get lower probability to be selected. In order to reach this aim, we propose reverseBV, as the reverse of BV, which takes high values when a sample \( x_n \) has not been well-clustered in previous iterations (eq. (6)) [5]. Based on reverseBV, the loss and pseudo loss is computed using eq. (7,8) [5], and then, the distribution of the samples’ weight, \( w^{i+1}_n \), is calculated using eq. (9). In eq. (9), \( Z^{i+1} \) is a normalization constant used to enforce that \( \sum_{n=1}^{N} w^{i+1}_n = 1 \). This formula assigns higher weights to badly clustered samples and consequently lower weights to well clustered ones [5].

Step 2.6, stopping criteria: The algorithm terminates if the iteration number, \( i \), equals to the maximum number of iterations, \( T \). The newly assigned weights, i.e. \( w^{i+1} \), are used in the sampling algorithm for the next iteration.

Step 3, creating the final clustering: The consensus clustering generated in the \( T^{th} \) iteration, \( H^{T}_{agg} \), is the final clustering which is claimed to be better in quality (eq. (10)).

3.2. Time complexity analysis

In Bob-Hic algorithm, three methods, the Hclusterer, the Hcombiner and the recovery ( \( f^{-1} \)), are iteratively done for \( T \) times. Suppose that the time-complexities of the mentioned methods are consecutively denoted by \( T_{cluster} \), \( T_{combine} \) and \( T_{recovery} \). So, the overall time complexity of the proposed algorithm will be \( O(T \times (T_{cluster} + T_{combine} + T_{recovery})) \), or \( O(L \times (T_{cluster} + T_{combine} + T_{recovery})) \).

The worst computational complexity of the Hclusterer method and also the recovery method used in this framework (i.e. Centroid, Single, Average, Complete, Weighted, Median or Ward linkage methods), is \( O(n^2 \log n) \) [48]. Also, the underused Hcombiner methods are three common matrix operations, minimum, average and maximum (further discussed in 4.1), which have the computational complexity of \( O(n^2) \). So, the overall complexity of Bob-Hic is \( O(L \times (n^2 + n^2 + n^2 \log n)) = O(L \times n^2 \log n) \).

In the next section, the details of the experimental set up and results are given.

4. Experimental set up and results
The Bob-Hic has been evaluated on various benchmark datasets given in table 1. Datasets are collected from popular real dataset repositories, University of California Irvine Repository of Machine Learning Datasets [49] and Real Medical Datasets [50]. The trial datasets contains different number of data points, from 85 to 768. The experimental methodology, the quality measurement method, statistical analysis and the experimental results are presented in the next subsections, 4.1 to 4.4.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#col</th>
<th>#instance</th>
<th>#feature</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast_cancer</td>
<td>1</td>
<td>263</td>
<td>9</td>
<td>[49]</td>
</tr>
<tr>
<td>contractions</td>
<td>2</td>
<td>98</td>
<td>27</td>
<td>[50]</td>
</tr>
<tr>
<td>diabetes</td>
<td>3</td>
<td>768</td>
<td>8</td>
<td>[49]</td>
</tr>
<tr>
<td>heart</td>
<td>4</td>
<td>270</td>
<td>13</td>
<td>[49]</td>
</tr>
<tr>
<td>laryngeal2</td>
<td>5</td>
<td>692</td>
<td>16</td>
<td>[50]</td>
</tr>
<tr>
<td>laryngeal3</td>
<td>6</td>
<td>353</td>
<td>16</td>
<td>[50]</td>
</tr>
<tr>
<td>liver_disorders</td>
<td>7</td>
<td>345</td>
<td>6</td>
<td>[49]</td>
</tr>
<tr>
<td>pima.indians-diabetes</td>
<td>8</td>
<td>768</td>
<td>8</td>
<td>[49]</td>
</tr>
<tr>
<td>respiratory</td>
<td>9</td>
<td>85</td>
<td>17</td>
<td>[50]</td>
</tr>
<tr>
<td>thyroid aeberhard</td>
<td>10</td>
<td>215</td>
<td>5</td>
<td>[49]</td>
</tr>
<tr>
<td>thyroid</td>
<td>11</td>
<td>215</td>
<td>5</td>
<td>[49]</td>
</tr>
<tr>
<td>weaning</td>
<td>12</td>
<td>302</td>
<td>17</td>
<td>[50]</td>
</tr>
<tr>
<td>wpbc</td>
<td>13</td>
<td>198</td>
<td>32</td>
<td>[49]</td>
</tr>
</tbody>
</table>

4.1. Experimental methodology

At the starting point of the experiments, a basic single h-clustering algorithm is utilized to create dendrogram ensembles. The proposed method is evaluated by seven well known agglomerative h-clustering methods. We define the variable $F1$ to show the clusterer types in 7 levels as below:

$$F1 = \{ \text{Centroid, Single, Average, Complete, Weighted, Median, Ward} \}$$

The clustering algorithms are applied on a sample of data. Here, the sample is set to be 40% of the original dataset. Selecting the subsamples, the next step is to apply basic clustering algorithm on the samples to create the dendrogram. After that, we represent the dendrograms in form of descriptors. In our experiments, the cophenetic difference ($CD$) matrices are used.

A combination method which is based on the Rényi divergence approach is then applied to the descriptors. Setting $1 - \alpha$ parameter equals to $(-\infty,1,\infty)$, the combination function is converted to common operators minimum, average and maximum. So we define the variable $F2$ to show the combination types in 3 levels as below:

$$F2 = \{ \text{Max, Min, Average} \}$$

These combination methods are applied on descriptors. The consensus matrix is then put into a dendrogram recovery function, $f^{-1}$, to retrieve the final h-clustering result. The tested recovery functions are the common agglomerative h-clustering methods namely Centroid, Single, Average, Complete, Weighted, Median and Ward. So we define the variable $F3$ to show the recovery methods in 7 levels as below:
According to the following parameters, clusterer types, combination types and recovery methods, the experiments are conducted with $7 \times 3 \times 7 = 147$ different parameter values to evaluate the accuracy of consensus hierarchical partition results. So, the ensemble method is applied 147 times on each dataset. The number of boosting iterations, $T$, in each application is set to 100.

After performing all of the experiments, we design a statistical analysis to find the most effective parameter values among 147 different parameter values.

After finding the most effective parameter values, the experiment is reduced to these values, which we name it as the preferred Bob-Hic with a single value for each of the parameters clusterer type, combination type and recovery method.

Finally, the quality of final consensus h-clustering generated by the preferred method is compared to the standard h-clustering methods and also the fusing method, MATCH.

The quality measurement method, the statistical analysis and comparative results are explained further in sections 4.2 to 4.4.

### 4.2. Quality measurement method

In order to evaluate the proposed method, a quality measurement is needed to verify that whether the consensus h-clustering structure fits the original data [6, 51]. The most familiar measurement is cophenetic correlation coefficient (CPCC). This coefficient compares the two matrices containing defined distances between pairs of data points, one of them corresponds to the original data and the other corresponds to the hierarchy [6, 46, 51-53]. The defined distances on original data can be Euclidian and the distances on hierarchy can be described by cophenetic difference (CD). The CPCC between two different distance matrices $X$ and $Y$ is defined as eq. (13), in which $X_{ij}$ and $Y_{ij}$ are distances between objects $i$ and $j$ in matrices $X$ and $Y$ respectively. Also $\bar{X}$ and $\bar{Y}$ are the average distances.

\[
CPCC = \frac{\sum_{i,j} (X_{ij} - \bar{X})(Y_{ij} - \bar{Y})}{\sqrt{\sum_{i,j} (X_{ij} - \bar{X})^2 \sum_{i,j} (Y_{ij} - \bar{Y})^2}}
\]  

The CPCC takes a value from the interval [0,1], where a higher value shows a better agreement between two tested matrices. In our experiments, the CPCC between the Euclidian matrix of the original data and the cophenetic matrix of consensus dendrogram is measured to show the quality of the results.

The statistical analysis and experimental results are described in the following subsections, 4.3 and 4.4.

### 4.3. Statistical analysis
In this subsection, we design a statistical analysis method to find the most effective parameter values among the 147 different parameter values. Here, we perform a factorial analysis of variance (ANOVA) [54]. Factorial ANOVA is a technique which is used to prospect the variation of a continuous dependent response variable under the experimental conditions identified by independent categorization variables. The variation in the response is supposed to be caused by effects of the categorization variables, and the variation caused by random error. Similarly, factorial ANOVA is used when experimental data have multiple categorization variables and a continuous response variable. So the factorial ANOVA is believed to be appropriate in our experiments, with three multiple categorization variables (F1, F2 and F3) and one response variable (CPCC).

Here, we design a factorial ANOVA model, M1, on CPCC values of 147 experiments on all datasets. We compare CPCC levels for each F1, F2 and F3. The defined factorial model also includes the effects of up to 3-way interactions, F1×F2, F2×F3, F1×F3 and F1×F2×F3. In this analysis, the variation of the CPCC is expected to be the sum of the variation caused by multiple way interactions effects of the categorization variables (i.e. F1, F2 and F3), and the variation caused by random error.

The result of the analysis includes the standard ANOVA table. Table 2 is the result of analyzing the defined model M1. The resulting table includes six columns namely Source, DF, Sum of Squares, Mean Square, F-Value, and Pr>F. The column Source stands for source of variation, DF column stands for the degrees of freedom, F-value column shows the division of mean square of the model by the mean square of the error, and Pr>F is the probability note providing statistical significance [54]. Here, we focus on the last column, Pr>F, which is also called p-value. The p-value resulted from our ANOVA analysis is <0.0001 (see table 2). This value indicates that the model M1 is significant in explaining the variation in the CPCC at the $\alpha = 0.05$ level (that is, each p-value is “much” less than $\alpha = 0.05$ level).

The effect of the categorization variables on M1 model is shown in table 3. Description of each column is mentioned in details in [54]. According to calculated p-value shown in this table, it is observed that the effects of F2, F3 and their 2-way interaction effect, i.e. F2×F3, are significant whose corresponding p-values are <0.0001, which is “much” less than $\alpha$ level. In contrast, the effects of F1 and its multiple way interaction effects, i.e. F1×F2, F1×F3 and F1×F2×F3, are not significant whose corresponding p-values are much more than $\alpha$ level.

According to these results, it is observed that the differences between F1 levels are not significant. So, the analysis is continued upon a new model, M2, with two multiple categorization variables F2 and F3 and the response variable CPCC.

Table 4 shows the result of analysis of defined model M2. A small p-value <0.0001 indicates that the overall model M2 is significant in explaining the variation in the CPCC. According to the new calculated p-values shown in table 5, it is observed that the effects of F2, F3 and F2×F3 are significant.
Table 2. standard ANOVA table as the result of the analysis of M1 model

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>146</td>
<td>49.125</td>
<td>0.336</td>
<td>14.09</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>1763</td>
<td>42.102</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1909</td>
<td>91.227</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Tests of effects of categorization variables F1, F2 and F3 of M1 model

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>6</td>
<td>0.388</td>
<td>0.065</td>
<td>2.71</td>
<td>0.0130</td>
</tr>
<tr>
<td>F2</td>
<td>2</td>
<td>0.952</td>
<td>0.476</td>
<td>19.93</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>F3</td>
<td>6</td>
<td>30.850</td>
<td>5.142</td>
<td>215.31</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>F1xF2</td>
<td>12</td>
<td>0.572</td>
<td>0.048</td>
<td>2.00</td>
<td>0.0218</td>
</tr>
<tr>
<td>F1xF3</td>
<td>36</td>
<td>0.377</td>
<td>0.010</td>
<td>0.44</td>
<td>0.9985</td>
</tr>
<tr>
<td>F2xF3</td>
<td>12</td>
<td>15.435</td>
<td>1.286</td>
<td>53.86</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>F1xF2xF3</td>
<td>72</td>
<td>0.843</td>
<td>0.012</td>
<td>0.49</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

Table 4. Standard ANOVA table as the result of the analysis of M2 model

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>20</td>
<td>46.922</td>
<td>2.346</td>
<td>101.68</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>1889</td>
<td>43.170</td>
<td>0.023</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1909</td>
<td>90.093</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Tests of Effects of categorization variables F2 and F3 of M2 model

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>2</td>
<td>0.949</td>
<td>0.475</td>
<td>20.57</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>F3</td>
<td>6</td>
<td>30.871</td>
<td>5.145</td>
<td>223.00</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>F2xF3</td>
<td>12</td>
<td>15.435</td>
<td>1.286</td>
<td>55.75</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

In the following, we perform Duncan's multiple range test (MRT) to obtain means of CPCC in multiple levels of F2 and F3. Duncan MRT is a type of multiple comparison procedure which compares sets of means and determines the significant differences between means [55]. 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.

In the beginning of the Duncan’s MRT, some sets are indicated in relation to each parameter. Each set contains the results of the applied method taking one parameter’s value unchanged while those remaining are varied in the relevant indicated range of values. Then, the means of sets are arranged into some groups. The grouping indication is in such a way that means within the same group are not significantly different and, those from different groups are significantly different at an assumed $\alpha = 0.05$ level. Table 6 and 7 shows the results of Duncan's multiple range tests for F2 and F3 of model M2.

The Duncan grouping column in table 6 and 7 shows the means which are significantly different. According to table 6, we can conclude that the mean CPCC for Min combination type is higher than the means for all other combination types and differences between other
means are not significant. Similarly, from table 7 we can conclude that the mean of the CPCC for Average and Centroid recovery methods is higher than the means of all other recovery methods.

<table>
<thead>
<tr>
<th>Duncan Grouping</th>
<th>Mean</th>
<th>F2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.62890</td>
<td>Min</td>
</tr>
<tr>
<td>B</td>
<td>0.58024</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>0.55863</td>
<td>Max</td>
</tr>
</tbody>
</table>

Table 6. Duncan’s multiple range test for F2 of M2 model

<table>
<thead>
<tr>
<th>Duncan Grouping</th>
<th>Mean</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.74023</td>
<td>Average</td>
</tr>
<tr>
<td>B</td>
<td>0.68833</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td>0.67200</td>
<td>Weighted</td>
</tr>
<tr>
<td>C</td>
<td>0.63398</td>
<td>Single</td>
</tr>
<tr>
<td>D</td>
<td>0.38398</td>
<td>Complete</td>
</tr>
<tr>
<td>E</td>
<td>0.25932</td>
<td>Ward</td>
</tr>
</tbody>
</table>

Table 7. Duncan’s multiple range test for F3 of M2 model

Accordingly, it can be said that the main effect of Min level of F2 and Average level of F3 on CPCC is higher than all other levels of F2 and F3. But, as the 2-way interaction effect of F2×F3 is significant, it cannot directly be concluded that the two variable levels Min from F2 and Average from F3 cause the strongest variation on CPCC. In order to consider the 2-way interaction effect the mean plot of CPCC versus F2 and F3 is used. In figure 3 the means plot are illustrated, where the seven curves display CPCC values versus combination types. Here, each curve corresponds to one of the recovery methods. The means plot in figure 3 indicates that the CPCC values are higher in the case that F2 is on the Min level and F3 is on the Average level.

Figure 3. Means plot of CPCC versus F2 and F3

According to the analysis of the model it can be concluded that in the aggregation of any of the clusterer types, Min combination type and Average recovery method commonly
generate results of better quality (i.e. higher CPCC value). Hence, we can choose any clusterer type (here the Centroid), Min combination type and Average recovery method as the preferred level.

4.4. Experimental results and comparison

In order to show the quality improvement of the preferred level of Bob-Hic method, (i.e. Centroid clusterer type, Min combination type and Average recovery method) over basic h-clustering methods and MATCH algorithm, a comparison is done and the results are shown in table 8. The maximum CPCC values obtained from proposed solution and basic methods are also compared.

It is indicated from the table 8 that the preferred Bob-Hic performs better than basic h-clustering methods in all cases and it performs as well as MATCH. Comparing the time complexity of the proposed method with MATCH, i.e. $O(Ln^2 \log n)$ versus $O(Ln^4)$, we can say that Bob-Hic performs better in reducing time complexity.

Table 8. Comparison of CPCC values between proposed solution and seven well known agglomerative h-clustering methods namely Centroid, Single, Average, Complete, Weighted, Median and Ward with no subsampling and also MATCH fusing algorithm.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Preferred</th>
<th>Max</th>
<th>Standard methods preferred</th>
<th>Max</th>
<th>Ward preferred</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob-Hic</td>
<td>0.762</td>
<td>0.995</td>
<td>0.792</td>
<td>0.771</td>
<td>0.736</td>
<td>0.805</td>
</tr>
<tr>
<td>Centroid</td>
<td>0.715</td>
<td>0.853</td>
<td>0.604</td>
<td>0.785</td>
<td>0.575</td>
<td>0.653</td>
</tr>
<tr>
<td>Single</td>
<td>0.716</td>
<td>0.799</td>
<td>0.727</td>
<td>0.617</td>
<td>0.683</td>
<td>0.747</td>
</tr>
<tr>
<td>Average</td>
<td>0.644</td>
<td>0.676</td>
<td>0.594</td>
<td>0.724</td>
<td>0.530</td>
<td>0.554</td>
</tr>
<tr>
<td>Complete</td>
<td>0.631</td>
<td>0.669</td>
<td>0.554</td>
<td>0.786</td>
<td>0.666</td>
<td>0.683</td>
</tr>
<tr>
<td>Weighted</td>
<td>0.610</td>
<td>0.666</td>
<td>0.435</td>
<td>0.470</td>
<td>0.459</td>
<td>0.426</td>
</tr>
<tr>
<td>Median</td>
<td>-</td>
<td>0.762</td>
<td>0.790</td>
<td>0.881</td>
<td>0.921</td>
<td>0.812</td>
</tr>
<tr>
<td>MATCH</td>
<td>-</td>
<td>0.762</td>
<td>0.624</td>
<td>0.790</td>
<td>0.881</td>
<td>0.921</td>
</tr>
</tbody>
</table>

5. Experimental results on 2-dimensional labeled datasets

In this subsection, some 2-dimensional labeled datasets are used to better demonstration of the Bob-Hic performance. Datasets are illustrated in figure 4. Datasets named Banana, Long1, Spiral, Circle, Smile and Triangle2 include elongated clusters. These datasets are hard to be clustered with compactness based clustering algorithms. Dataset named Square1 includes four clusters which are equal in size and spreading of data points. Square4 and Size5 are variations of Square1 dataset with different degree of overlapping. These datasets are hard to be clustered with connectedness based clustering algorithms. Finally, datasets named Longspiral, Longsquare and Spiralsquare include different types of clusters which are hard to be clustered with both mentioned clustering algorithms [56].

The result of applying Single Linkage clustering method on each dataset is illustrated in figure 5. It can be seen from figure 5 that the Single Linkage method acts well on datasets which contains elongated clusters, but creates a low quality clusterings on datasets which contains either overlapped clusters or different clusters types. The illustration of our
consensus clustering results and evaluation of the results are presented in the next subsections, 5.1 and 5.2.

Figure 4. Datasets used to demonstrate the quality of the proposed h-clusterer ensemble method, Bob-Hic.

Figure 5. The resulting partition of experimental datasets shown in fig. 3 by Single Linkage h-clustering method.
5.1. Illustrated clustering results of Bob-Hic

In order to evaluate the proposed method, the Bob-Hic clusterer ensemble algorithm is applied on the data sets shown in figure 4. The initialization step of the Bob-Hic algorithm is as follows: First, the number of iterations, $T$, is set to 100. In order to create basic single h-clusterings, one of the agglomerative h-clusterings is selected randomly, since the effect of different basic standard h-clusterings is statistically proved to be not significantly different. Then, the subsample is set to be 40% of the original dataset and the distances are computed using the Euclidean distances. Finally, the $Min$ aggregation function and the Average linkage method are chosen as the h-clustering combination method and the recovery method respectively. Initializing the algorithm, a consensus h-clustering is generated on each dataset. The results of the consensus clustering on each dataset are illustrated in figure 6.

According to figure 6, the Bob-Hic algorithm acts well on the most of datasets consisting of both elongated and overlapped clusters. And it can be seen that the proposed method creates good clusters on some datasets containing different cluster types.

![Figure 6. The resulting partition of experimental datasets shown in Fig. 3 by Bob-Hic clusterer ensemble algorithm](image)

5.2. Evaluation criterion of clustering quality

In order to evaluate the quality of the final resulting clusters, we use the accuracy measure, $Fscore$. This measurement evaluates the similarity of a clustering to ground truth
information of classes \[57\]. Let \(c\) to be the number of individual classes. Then, the total \(F\text{score}\) will be computed as the weighted sum of these classes’ \(F\text{score}\) according to their size. So, the \(F\text{score}\) can be calculated by eq. (14), in which \(C_r\) is a class with the size of \(n_r\) and \(F(C_r)\) is the \(F\text{score}\) of class \(C_r\).

\[
F(C_r) = \sum_{r=1}^{c} \frac{n_r}{N} F(C_r) \tag{14}
\]

Assume that \(S_i\) is the \(i\)th cluster; For each class \(C_r\), \(F(C_r)\) finds a corresponding cluster \(S_i\) in hierarchy \(H\), that agrees with \(C_r\) more better than the other clusters. \(F(C_r)\) is calculated by eq. (15), where \(P_{S_i}\) is the precision (the number of objects in the cluster \(S_i\) belonging to the class \(C_r\), divided by the number of objects in the cluster \(S_i\)) and \(R_{S_i}\) is the recall (the number of objects in the cluster \(S_i\) belonging to the class \(C_r\), divided by number of objects in the class \(C_r\)).

\[
F(C_r) = \max_{S_i \in H} \left\{ \frac{2P_{S_i} R_{S_i}}{P_{S_i} + R_{S_i}} \right\} \tag{15}
\]

In the first step, we extract a predefined number of clusters for each dataset from the corresponding dendrogram, which is used as a representative data of the h-clustering result. In the next step, each data point is labeled according to the clustering result. Finally, the \(F\text{score}\) measure evaluates the similarity between the class label of every point in the dataset and the label extracted from the ensemble result.

### 5.3. Comparison of the results on 2-dimensional labeled datasets

Table 9 demonstrates the \(F\text{score}\) results of applying standard clustering methods, MATCH and Bob-Hic on datasets shown in Figure 4. The computed \(F\text{score}\) values of each method on different datasets are averaged and shown in the last row of table 9. This table reveals that the average \(F\text{score}\) value of Bob-Hic is significantly better than both non-ensemble standard methods and the MATCH ensemble method.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Standard methods</th>
<th>MATCH preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centroid</td>
<td>Single</td>
</tr>
<tr>
<td>Banana</td>
<td>0.945</td>
<td>1.000</td>
</tr>
<tr>
<td>Long1</td>
<td>0.517</td>
<td>1.000</td>
</tr>
<tr>
<td>Spiral</td>
<td>0.680</td>
<td>1.000</td>
</tr>
<tr>
<td>Square1</td>
<td>0.984</td>
<td>0.203</td>
</tr>
<tr>
<td>Square4</td>
<td>0.907</td>
<td>0.204</td>
</tr>
<tr>
<td>Sizes5</td>
<td>0.989</td>
<td>0.761</td>
</tr>
<tr>
<td>Circle</td>
<td>0.637</td>
<td>1.000</td>
</tr>
<tr>
<td>Smile</td>
<td>0.527</td>
<td>1.000</td>
</tr>
<tr>
<td>Triangle2</td>
<td>0.946</td>
<td>0.995</td>
</tr>
<tr>
<td>Longsquare</td>
<td>0.745</td>
<td>0.338</td>
</tr>
<tr>
<td>Lonspiral</td>
<td>0.516</td>
<td>0.501</td>
</tr>
<tr>
<td>Spiralsquare</td>
<td>0.590</td>
<td>0.600</td>
</tr>
<tr>
<td>Average (F\text{score})</td>
<td>0.748</td>
<td>0.716</td>
</tr>
</tbody>
</table>
From the other point of view, it can be seen that Bob-Hic can get good clusters on the datasets which are hard to be clustered with connectedness based clustering algorithms (i.e. Square1, Square4 and Size5). Also, Bob-Hic gets good clusters on some datasets which are hard to be clustered with compactness based clustering algorithms (like Banana, Long1, Circle, Triangle2 and Longspiral). But, the results on some other datasets which include elongated clusters (like Smile, Spiral, Spiralsquare and Longsquare) are not so satisfying. This might be cause of the effect of the Average method which is used to recover the final dendrogram. Nevertheless, in these databases, the results are better than those algorithms which are based on compactness clustering, like the Average standard method.

The variation of the two quality measurements CPCC and Fscore versus iteration number is shown in figure 7. The result in this figure comes from applying the Bob-Hic algorithm on the Size5 dataset. This figure shows that the algorithm successfully improves both the Fscore measurement and CPCC.

![Figure 7. Variation of two quality measures CPCC and Fscore at different iterations of Bob-Hic algorithm on the Size5 dataset](image)

6. Conclusions

In this paper a novel hierarchical clusterer ensemble method is proposed based on the boosting theory. There are several motivations beyond this hierarchical ensemble method. To the best of our knowledge, numerous ensemble methods exists which construct a set of classifiers or partitional clusterers, while few methods are designed to handle the situation in which a hierarchy of clusters is needed. Hence, the proposed algorithms can acquire this need. Boosting is a familiar classifier ensemble method and it can be successful when be applied to the clusterer ensembles. Boosting is a multiple learner combination method which always uses resampling and reweighting. In order to reweight the samples, we introduce a validation procedure to evaluate how well an individual data point has been clustered in the hierarchy.

Several real datasets are used to show that boosting is a good method to build hierarchical clusterer ensembles. Based on the experimental results, we show that the quality of the final clustering derived from Bob-Hic is superior to any of the clusterings performing alone. And
although the proposed method, Bob-Hic, and the MATCH method, achieve the same clustering quality in many cases, we prove that the time complexity of Bob-Hic is reduced in comparison to MATCH.

References

[38] C. Domeniconi, M. Al-Razgan, Weighted cluster ensembles: methods and analysis, ACM Transactions on Knowledge Discovery from Data, 2 (2009) 1-40.
[52] V.P. Lessig, Comparing cluster analyses with cophenetic correlation, Journal of Marketing Research, 9 (1972) 82-84.