Ensembled Heuristic Iterative Expected Maximization with BrownBoost Data Clustering for Uncertain Data Mining

Muruganantham S1, Elango N M2
1 Assistant Professor, Department of Computer Technology and Information Technology, Kongu Arts and Science College (Autonomous), Erode-638107, Tamil Nadu, India
2 Associate Professor, School of Information Technology and Engineering, VIT University, Vellore-632014, Tamil Nadu, India.

Abstract
Clustering on uncertain data needs to be handled in order to generate significant knowledge patterns. Clustering is the data mining tasks to group similar information or data. Though the conventional algorithm groups similar data for mining the uncertain data, minimizing the error rate and clustering time and improving the accuracy turn out to be the major issues in clustering of uncertain data. In this paper, Ensembled Heuristic Iterative Expected Maximization with BrownBoost Data Clustering (EHIEM-BBDC) technique is proposed in which the number of base learners are constructed using Iterative Expected Maximization and the heuristic is applied to speed up the process of clustering. The BrownBoost technique is applied for improving the clustering accuracy by combining the base learners to form strong cluster. Experimental evaluation of proposed EHIEM-BBDC technique and existing methods are carried out with the El Nino dataset taken from the UCI machine learning repository. The results have shown that the proposed technique outperforms well to mine the uncertain data through the high clustering accuracy with minimum time as well as less false positive rate.

Keywords: Uncertain data mining, Heuristic Iterative Expected Maximization, likelihood estimation, BrownBoost Data Clustering.

1. INTRODUCTION
Data uncertainty causes much more complexity in real-world applications due to certain reasons such as inaccurate measurement, sampling errors and so on. Recently, several data mining technique has been introduced for handling uncertain data. Among them, clustering technique is applied for mining the uncertain data to attain accurate mining results.

In [1], UK means clustering mechanism was developed to group the uncertain data. The mechanism increases the effectiveness of clustering and minimizes the time but it failed to identify the similar cluster member. A prototype-based agglomerative hierarchical clustering (U-AHC) of uncertain objects was developed in [2]. The clustering technique groups similar data depends on the information-theoretic and expected-distance measures. But the clustering time was not minimized.

A modified Affinity Propagation clustering algorithm based on K-nearest neighbor intervals (KNNI) was developed in [3] for uncertain data mining. Though the algorithm reduces the error, the clustering time was not minimized. A Kullback-Leibler divergence was presented in [4] to calculate the similarity for clustering the uncertain data. The method used density-based clustering but the computation cost was not minimized.

Variance for cluster Mixture Model (MMvar) was introduced in [5] for grouping the uncertain data by reducing the variance. The MMvar achieving better accuracy but the error rate was not minimized. An efficient algorithm was designed in [6] for uncertain patterns mining based on data structures. The algorithm increased the accuracy with the minimum false positive rate. But it failed to provide the appropriate mining results.

In [7], an intuitionistic fuzzy possibilistic C means (IFPCM) algorithm was designed to group the data objects with high accuracy. Though the clustering algorithm minimizes the computational complexity, the false positive rate was not minimized. The Kullback-Leibler divergence was developed in [8] for grouping the uncertain data. The method minimizes the time complexity but the accurate clustering was not attained.

Authentication of uncertain data was performed in [9] using a k-means clustering technique. This iterative approach increases the clustering performance and minimizes the error. But it fails to extend the technique for tackling the correlation. A multi-core density-based uncertain data clustering algorithm was presented in [10]. The algorithm increased the performance of uncertain data clustering but the time complexity was not minimized.

The major issues are identified from the above said methods such as lack of providing the accurate clustering results, high clustering time, high false positive rate, and so on. In order to address such kind of issues in the existing clustering, an ensemble of Heuristic Iterative Expected Maximization with BrownBoost Data Clustering (EHIEM-BBDC) is introduced. The EHIEM-BBDC technique provides a novel contribution for mining the uncertain data using ensemble clustering technique. The EHIEM-BBDC technique constructs the number of base learners i.e. Heuristic Iterative Expected Maximization technique group the similar data based on
maximum likelihood probability estimation. The maximum likelihood probability is computed between the data and cluster center to group the similar data. The heuristic is used to increase the data clustering process. The brownboost technique is applied for accurate data clustering by combining the base learners into strong. After that, the weight is set to all base learners and computing the potential loss function. The boosting technique provides strong clustering results with minimum loss function. This process increases the clustering accuracy and minimizes the false positive rate.

The remaining of this paper is organized into five different sections. Section 2 provides related works in uncertain data clustering. Description of the EHIEM-BBDC technique is elaborated with the neat diagram in Section 3. Experimental evaluation of proposed EHIEM-BBDC technique and state-of-art methods are described in section 4. Section 5 provides the results and discussion of certain parameters with table and graphical representation. Finally, section 6 concludes the proposed work.

2. RELATED WORKS

A model-based clustering analysis was performed in [11] for increasing the clustering performance. The model-based approach minimizes the problem of overlapping clusters. However, the model failed to obtain accurate results in the clustering process. A self-adapted mixture distance measure was performed in [12] for grouping the uncertain data based on the geometric distance and the probability distribution distance. The model was complex to attain satisfactory clustering results.

A novel hybrid clustering algorithm was introduced in [13] depends on the similarity. The accurate clusters were acquired using this algorithm. Though the algorithm minimizes the false positive rate, the clustering time was not minimized. The K-means divide and conquer approach based on vector model was developed in [14] to group the big data. The approach increases the clustering accuracy but it failed to perform uncertain data clustering.

A hybrid version of the well-known K-means clustering algorithm with the genetic algorithm was designed in [15] for effectively solve the empty clustering issue. But the algorithm takes more time to cluster the data. Clustering of uncertain data streams was performed in [16] using a sliding-window model. The model does not perform the clustering on high-dimensional uncertain streams. A superseding nearest neighbor (SNN) search was developed in [17] for grouping the uncertain objects with less computation time. The SNN search failed to accurately perform the clustering of uncertain objects.

Clustering the large graphs using neighborhood information was presented in [18]. The algorithm failed to attain the high-performance cluster with minimum computational cost. A Fuzzy c-means clustering algorithm was designed in [19] for grouping the uncertain data using quadratic penalty-vector regularization. The algorithm was more effective for handling the more number of uncertain data.

A clustering algorithm for Uncertain Data based on approximate backbone was presented in [20] for achieving the better clustering results. Though the algorithm minimizes the clustering time, the error rate was not computed.

The issues from the above said reviews are overcome by introducing the novel technique called EHIEM-BBDC. The description of the ensemble clustering technique is presented in the next section.

3. ENSEMBLE OF HEURISTIC ITERATIVE EXPECTED MAXIMIZATION WITH BROWNBOOST DATA CLUSTERING TECHNIQUE FOR UNCERTAIN DATA MINING

An ensemble of Heuristic Iterative Expected Maximization with BrownBoost Data Clustering (EHIEM-BBDC) Technique is introduced in this section to improve the clustering accuracy and mining the uncertain data. The EHIEM-BBDC technique is a machine learning ensemble meta-algorithm converts the weak learners into a strong one. In EHIEM-BBDC technique, an Iterative Expected Maximization Cluster is a weak learner which does not attain the accurate clustering results. In contrast, a BrownBoost is a strong learner which provides accurate results and minimizes the expected sum of the error.

![Figure 1: Architecture diagram of the EHIEM-BBDC technique](attachment:diagram.png)
The accurate clustering involves grouping similar data into the clusters. The strong clustering results increase the clustering accuracy and minimize the time complexity. Figure 1 illustrates the architecture diagram of the EHIEM-BBDC technique to mine the uncertain data through the ensemble clustering technique. Let us consider the number of data $D_1, D_2, D_3, ..., D_n$ taken from the dataset. The collected data are taken as input to the number of base clusters. The heuristic iterative expectation maximization algorithm is applied in EHIEM-BBDC technique for constructing the base cluster to group the similar data. After that, the number of base learners is combined to make strong clustering results by minimizing the error. The clustering results are described in the following subsection.

### 3.1 Iterative Expected Maximization with BrownBoost Data Clustering

Heuristic iterative Expected Maximization is a statistical model to compute the Maximum Likelihood in the presence of uncertain data. The likelihood expresses how the one variable is more similar to other. In Iterative Expected Maximization, an iterative method is a mathematical procedure used to generate a sequence of improving solutions for a class of problems. A heuristic is used to speed up the process of attaining an efficient solution based on the Maximum Likelihood estimation. The heuristic iterative expected maximization clustering technique employs the termination criteria to group all the similar data into the clusters. The existing clustering technique failed to deliver an exact solution while considering the number of data. In contrast, the proposed clustering method uses the heuristic iterative procedure which presents a better solution while handling a large number of data.

Figure 2 illustrates a heuristic iterative expectation and maximization based clustering the data. The number of data $D_1, D_2, D_3, ..., D_n$ are taken as input for the clustering process. Initialize the ‘n’ number of clusters $c_1, c_2, c_3, ..., c_n$ and cluster centers. The clustering technique includes two phases such as expectation and maximization. In the expectation, the expected likelihood between the data and cluster center is computed to group the similar data. The maximum likelihood probability is achieved by maximizing the expected likelihood based on the Maximum Likelihood estimation.

From (1), $eL(D, c_j)$ represents the expected likelihood between the ‘D’ and cluster center $c_j$. $P$ denotes a probability used to discover the data which is more likely belonging to the cluster. The expected likelihood is maximized in the maximization phase. In this phase, the heuristic method is used to attain the satisfactory clustering results by maximizing the expected likelihood.

From (2), $\theta_M$ denotes maximum log likelihood estimators which maximize the log likelihood value determined from the expectation step. A $\arg \max$ denotes an argument of the maximum function. The maximum likelihood probability indicates the more similar data are grouped into the particular cluster. This process is iterated until all the data are grouped into any of the cluster. As a result, the clustering is performed but the error rate is not minimized.

In order to attain the accurate clustering results, brownboost algorithm transforms a weak learner into a strong learner by combining the weak classifier generated results of the training data. At every iteration of a boosting algorithm, the weak learner is trained on the weighted training data and achieves less error rate.

The BrownBoost is a boosting algorithm that robust to noisy datasets and minimizes a loss. To construct a strong cluster for further improving the accuracy, the output of each base learner is summed and it is expressed as follows:

$$ Y = \sum_{i=1}^{n} w_i(D) $$

From (3), $Y$ denotes an output of the strong cluster. $w_i(D)$ denotes an output of the base cluster. After that, the weight is set to all the base clusters based on the remaining time of the base learner after clustering and margin of the data.

![Figure 2: Process of heuristic iterative expectation and maximization cluster](image)
The margin of the data is positive or negative real value which defines whether the data is correctly clustered under the base cluster. A positive margin defines the data is being correctly clustered and the negative value signifies the data is being incorrectly grouped. Therefore, the magnitude of the margin value shows that how effectively the base learner grouped the data into the particular cluster. The BrownBoost algorithm states that each base cluster takes a different amount of time which is directly related to the weight given to the base cluster. The weight is expressed as follows,

$$\beta_i = \exp \left( \frac{-\left(m_i(D_i)+\epsilon_r\right)^2}{v} \right) \quad (4)$$

From (4), $\beta_i$ represents the weight that the boosting algorithm assigns to the weak learner at iteration ‘i’. $m_i$ denotes margin of the data $D_i$. $t$ denotes an amount of remaining time of the base cluster ($t = v$). After setting the weight to base cluster, the potential loss for each data with margin $m_i$ is computed as follows,

$$P_L = 1 - \varepsilon_r \sqrt{v} \quad (5)$$

From (5), $P_L$ denotes a potential loss of the function, $\varepsilon_r$ denotes an error functions, $v$ denotes a positive real valued parameter. The margin of the each base cluster is updated based on the loss value.

$$m_i(t + 1) = m_i(D_i) + \sum_{i=3}^{n} \beta_i w_i(D) y_i \quad (6)$$

From (6), $m_i(t + 1)$ represent the updated margins of the data sample. $\beta_i$ denotes a weight of the base learner to ensure the final strong clustering results, $w_i(D)$ denotes a base cluster result, $y_i$ represents the actual output of the base cluster. The updated value shows that the boosting technique effectively groups the similar data into the cluster and minimizes the incorrect data clustering and forms the final strong results as the linear combination of the base cluster (i.e. training error is small). Then the strong clustering results are expressed as follows,

$$Y = \text{sign} \left( \sum_{i=1}^{n} \beta_i w_i(D) \right) \quad (7)$$

From (7), $Y$ denotes a final strong clustering result, $\beta_i$ represents the weight, $w_i(D)$ denotes an output of the base cluster. ‘$\text{sign}$’ represents the positive and negative results of output i.e. the data is correctly grouped and incorrectly grouped. By this way, the proposed ensemble clustering techniques effectively group the similar data into the clusters for mining the uncertain data.

Algorithm 1 describes the clustering the similar data to mine the uncertain data with minimum time complexity. The numbers of data are collected from the dataset. Initialize the number of cluster and cluster centers. For each cluster and data, compute the expected probability that the data belongs to the cluster. For each cluster, the likelihood probability is maximized to group all the data into the clusters. This clustering process is repeated until the convergence is met. The base clustering results are combined to make strong clustering results. For each base cluster, the weight is assigned based on the margin and remaining time. After that, the potential loss is computed to make strong results. The margin value is updated based on loss function. The updated margin value shows that the number of similar data correctly grouped into the particular cluster. This process increases the clustering accuracy and minimizes the false positive rate.

Input: Dataset $D$, Number of data $D_1, D_2, D_3, \ldots D_n$.

Output: Improve clustering accuracy.

Begin
1. Initialize number of clusters $c_1, c_2, c_3, \ldots c_n$ and centers $c_j$
2. for each $D_i \in D$
3. for each $c_j$
4. compute expected probability $eL(D, c_j)$ that the $D_i$ belongs to the cluster $c$
5. for each cluster $c$
6. Maximize expected log likelihood probability $eL(D, c_j)$ maximum similar data points belong to the cluster
7. end for
8. Got to step 4 until convergence is met
9. end for
10. End for
11. Combine all base cluster results $Y = \sum_{i=1}^{n} w_i(D)$
12. Initialize the margin $m_i(D_i) = 0$
13. For each $w_i(D)$
14. Set weight $\beta_i$
15. Compute potential loss $P_L$
16. Update margin $m_i(t + 1)$ for each data
17. Attain strong clustering results $Y = \text{sign} \left( \sum_{i=1}^{n} \beta_i w_i(D) \right)$
18. End for
End

Algorithm 1 Ensemble of Heuristic Iterative Expected Maximization with BrownBoost Data Clustering

4. EXPERIMENTAL EVALUATION

An experimental evaluation of EHIEM-BBDC technique and UK means clustering mechanism [1] and a prototype-based agglomerative hierarchical clustering method (U-AHC) [2] are implemented using Java language. The El Nino dataset is taken from the UCI machine learning repository. The weather dataset comprises the spatial information taken from the series of buoys positioned over the equatorial Pacific region. The dataset includes 12 attributes and 178080 instances. The attribute characteristics are an integer, real and their characteristics are spatio-temporal. The experiments are carried out with various parameters such as clustering accuracy, false positive rate and clustering time. Totally ten runs are carried out for each parameter. For the experimental consideration, the numbers of weather data are taken from 1000 to 10000. The experimental results are discussed in the following section.

5. PERFORMANCE RESULTS AND DISCUSSION

The results attained from the experimental evaluation of three clustering methods namely, EHIEM-BBDC technique and UK means clustering mechanism [1], U-AHC [2] are described in this section with a table and graphical results.
The different metrics such as clustering accuracy, false positive rate and clustering time are considered for evaluating the performance results of three different clustering methods.

5.1 Impact of clustering accuracy

Clustering accuracy is computed as the ratio of a number of (i.e. no. of) weather data are correctly grouped into the cluster to the total number of weather data.

\[
CA = \frac{\text{No. of data correctly grouped}}{\text{Total no. of data}} \times 100
\]

(8)

In Equation (8), ‘CA’ denotes a clustering accuracy. The clustering accuracy is measured in percentage (%). The mathematical computation of the clustering accuracy is presented with the three clustering methods namely, EHIEM-BBDC technique and UKmeans clustering mechanism [1], U-AHC [2].

The clustering accuracy is mathematically calculated as follows,

\[
CA = \frac{\text{No. of data correctly grouped}}{\text{Total no. of data}} \times 100
\]

(8)

In Equation (8), ‘CA’ denotes a clustering accuracy. The clustering accuracy is measured in percentage (%). Table 1 describes the experimental results of clustering accuracy versus a number of weather data collected from the EL-Nino dataset with three different clustering methods EHIEM-BBDC technique and UK means clustering mechanism [1], U-AHC [2]. The number of weather data is taken for the experimental consideration varied from 1000 to 10000. The results clearly show that the proposed EHIEM-BBDC technique improves the clustering accuracy when compared to existing clustering techniques.

5.2 Impact of false positive rate

The false positive rate is measured as the no. of data is incorrectly grouped into the cluster to the total no. of data. The mathematical formula for the false positive rate is computed as

\[
\text{False positive rate} = \frac{\text{No. of data incorrectly grouped into cluster}}{\text{Total no. of data}} \times 100
\]

(9)

Using Equation (9), the false positive rate is measured in terms of percentage (%). The mathematical computation of the false positive rate is presented with the three clustering techniques namely, EHIEM-BBDC technique and UKmeans clustering mechanism [1], U-AHC [2].
Let us consider the 1000 data taken as input, the above mathematical calculation shows that the EHIEM-BBDC technique incorrectly groups the weather data is 110. Then the corresponding percentage of false positive rate is 11%. The UKmeans clustering mechanism [1] incorrectly grouped data are 166 data and the percentage value is 17%. In addition, the other clustering technique U-AHC [2] incorrectly groups the weather data is 194 and the percentage value is 19%. Similarly, the nine results are computed and the values are tabulated.

Table 2: Tabulation for false positive rate

<table>
<thead>
<tr>
<th>Number of data</th>
<th>False positive rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EHIEM-BBDC</td>
</tr>
<tr>
<td>1000</td>
<td>11</td>
</tr>
<tr>
<td>2000</td>
<td>12</td>
</tr>
<tr>
<td>3000</td>
<td>7</td>
</tr>
<tr>
<td>4000</td>
<td>14</td>
</tr>
<tr>
<td>5000</td>
<td>11</td>
</tr>
<tr>
<td>6000</td>
<td>12</td>
</tr>
<tr>
<td>7000</td>
<td>10</td>
</tr>
<tr>
<td>8000</td>
<td>8</td>
</tr>
<tr>
<td>9000</td>
<td>7</td>
</tr>
<tr>
<td>10000</td>
<td>8</td>
</tr>
</tbody>
</table>

As shown in table 2, experimental results of the false positive rate with the number of weather data. For a different number of weather data, there are ten various false positive rates are attained. Here three clustering techniques are used to make a comparison for showing the significant improvements of the proposed EHIEM-BBDC technique. The table value clearly shows that the EHIEM-BBDC technique effectively performs the accurate clustering and minimizes the false positive rate when compared to existing clustering methods.

Figure 4 illustrates the experimental results of false positive rate based on the number of weather data taken from the EL-Nino dataset. In the two-dimensional graphical results, the number of weather data is taken as input and the corresponding false positive results are attained as an output. The graphical results obviously illustrate that the false positive rate of proposed EHIEM-BBDC technique is considerably minimized than the existing methods. The ensemble technique combines the results of all weak learners to make strong clustering results. After combining, the initial weight is set to all the base learners. Followed by, the loss is computed for each base learner result. Based on the loss, the initial margin gets updated and finds the number of data correctly grouped and incorrectly grouped. The EHIEM-BBDC technique minimizes the loss resulting in reduces the false positive.

The average results of false positive rate using EHIEM-BBDC technique are 44% minimized when compared to UKmeans clustering mechanism. The false positive rate is also minimized by 52% using EHIEM-BBDC technique when compared to U-AHC [2].

5.3 Impact of clustering time

Clustering time is defined as the amount of time taken to group the data into the clusters. The clustering time is mathematically computed as follows,

\[ CT = No. \text{ of data} \times T (\text{grouping the data}) \]  

In Equation (9), \( CT \) denotes a clustering time and \( T \) denotes a time. The clustering time is measured in milliseconds (ms). Mathematical computations of clustering time with three different methods are presented with the number of data.

Let us consider the number of input data is 1000. The proposed EHIEM-BBDC technique takes 21ms for grouping the similar weather data. The clustering time of the other two techniques UKmeans clustering mechanism [1] and U-AHC [2] are 32ms and 40ms respectively. Likewise, the remaining nine values are computed with various input data.

Table 3: Tabulation for clustering time

<table>
<thead>
<tr>
<th>Number of data</th>
<th>Clustering time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EHIEM-BBDC</td>
</tr>
<tr>
<td>1000</td>
<td>21</td>
</tr>
<tr>
<td>2000</td>
<td>22</td>
</tr>
<tr>
<td>3000</td>
<td>24</td>
</tr>
<tr>
<td>4000</td>
<td>28</td>
</tr>
<tr>
<td>5000</td>
<td>32</td>
</tr>
<tr>
<td>6000</td>
<td>41</td>
</tr>
<tr>
<td>7000</td>
<td>43</td>
</tr>
<tr>
<td>8000</td>
<td>48</td>
</tr>
<tr>
<td>9000</td>
<td>47</td>
</tr>
<tr>
<td>10000</td>
<td>52</td>
</tr>
</tbody>
</table>
As shown in table 3, experimental results of clustering time are described with the number of weather data. For a different number of input data, the various performance results of clustering time are attained. The number of data taken for computing the clustering accuracy is varied from 1000 to 10000. The results report that the EHIEM-BBDC technique takes minimum time to group similar data into the cluster than the UKmeans clustering mechanism [1] and U-AHC [2]. The results are plotted in the following two-dimensional graph.

Figure 5 shows the performance results of clustering time based on the number of weather data and it illustrates that the performance results of clustering time with three different clustering techniques. In the two dimensional graphical results, the number of data is taken for computing the clustering time. In figure 5, the three different colors namely blue, red, green indicates the performance results of clustering time using three techniques EHIEM-BBDC, UKmeans clustering mechanism [1] and U-AHC [2] respectively.

The above results and discussions show that the proposed EHIEM-BBDC technique effectively increases the clustering accuracy and minimizes the false positive rate and clustering time compared to the existing clustering techniques.

6. CONCLUSION

An efficient ensemble technique called Ensemble of Heuristic Iterative Expected Maximization with BrownBoost Data Clustering (EIEM-BBDC) is developed for mining the uncertain data with high clustering accuracy and minimal time. The uncertain data mining is carried out through the ensemble clustering technique. Initially, the numbers of data are collected from the dataset. Then, these data are trained with the base learner and groups the similar data into various clusters through the maximum likelihood probability. After that, the base learners are combined into strong one to group similar data by updating the margin. Experimental evaluation of EIEM-BBDC technique and existing clustering techniques are carried out using EL-Niño dataset. The performance results confirm that the EIEM-BBDC technique increases the clustering accuracy and minimizes the clustering time and false positive rate when compared to existing UKmeans clustering mechanism [1] and U-AHC [2].

REFERENCES


