An integrated robust semi-supervised framework for improving cluster reliability using ensemble method for heterogeneous datasets

Smita Prava Mishra\textsuperscript{a,\ast}, Debahuti Mishra\textsuperscript{b}, Srikanta Patnaik\textsuperscript{b}

\textsuperscript{a} Dept. of Computer Sc. \& Information Technology, Institute of Technical Education and Research, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India
\textsuperscript{b} Dept. of Computer Sc. \& Engineering, Institute of Technical Education and Research, Siksha ‘O’ Anusandhan University, Bhubaneswar, Odisha, India

Received 28 September 2015; revised 9 November 2015; accepted 9 November 2015
Available online 18 December 2015

Abstract

Data mining literature offer some clustering techniques. But when we implement even an effective clustering technique, the results are found unreliable. The efficacy of the technique come under scrutiny. Here, the proposal is about an integrated framework, which ensures the reliability of the class labels assigned to a dataset whose class labels are unknown. The model uses PSO-$k$-means, $k$-medoids, $c$-means and Expectation Maximization for data clustering. This model integrates their results through majority voting cluster ensemble technique to enhance reliability. The reliable outcomes serve as the training set for the classification process through Bayesian classifier, Multi Layer Perceptron, Support Vector Machine and Decision tree. The predicted class labels by majority of classifiers through bagging classifier ensemble method are included with the training set and in combination, designated as the set with known class labels. Heterogeneous datasets with unknown class labels but known number of classes, after being treated through this model would be able to find the class labels for a significant portion of the data and may be accepted with reliability. The evaluation procedure has been performed by following the Dunn’s, Davies–Bouldin and Modified Goodman–Kruskal indexing techniques for internal validation and probabilistic measures such as Normalized Mutual Information, Normalized Variation of Information and Adjusted Random Index which are appropriate measures of goodness-of-fit and robustness of the final clusters. The predictive capacity of the model is also validated through probabilistic measures and external indexing techniques such as Purity Measure, Random Index and F-measure.

© 2015 The Authors. Production and hosting by Elsevier B.V. on behalf of University of Kerbala. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Data clustering; Cluster ensemble; Internal indexing; External indexing; Probabilistic measures

1. Introduction

Dealing with unclassified data is a real time challenge for data miners. After years of research on several clustering algorithms, researchers could not succeed in designing a standalone robust clustering algorithm for heterogeneous datasets which could
assign reliable class labels to unclassified data. With the aim to improve reliability and robustness of the clustering outcomes, this paper proposes a semi-supervised method for clustering data where the class labels of the data are unknown. Multiple clustering techniques [1–4] such as PSO-\(k\)-means [5,6], \(k\)-medoids [7], \(c\)-means [8] and Expectation Maximization (EM) [4] are applied on datasets of diverse domains. To improve the reliability of the clustering results, the majority voting, otherwise known as bagging [2,9,10] cluster ensemble technique is adopted. Through the voting method, each dataset is segregated into two partitions. One having pure majority upon the obtained clustering results and the other data partition, without pure majority. Subsequently, a learning environment is simulated with multiple classifiers such as Bayesian classifier [11–14], Multi Layer Perceptron (MLP) [13–15], Support Vector Machine (SVM) [13–15] and Decision tree [13–15] classifiers being individually trained with the data partition with pure majority where label obtained from agreed clustering techniques is treated as the class label of the training set. After training, the classifiers are tested with the remaining partition of the data without pure majority. The testing results of multiple classifiers are again ensembled through majority voting [2]. The remaining data without pure majority after ensemble is discarded and rest of the data is accepted with their class labels. The evaluation procedure is performed to verify reliability of the results through various validation methods such as internal indexing techniques, external indexing techniques and statistical methods like probabilistic measures for clustering results. Our experimentation includes: (a) internal indexing techniques [16–19] such as Dunn’s index [20], Davies–Bouldin index [21] and Modified Goodman–Kruskal (\(GK_{modified}\)) index [22] which does not take any reference of the known class labels and only consider tightness of the intra-cluster elements and separation among inter-cluster elements for measuring the quality of the clusters; and (b) the probabilistic measures [23–25] taken for clustering result validation are Normalized Mutual Information (NMI), Normalized Variation of Information (NVI) and Adjusted Random Index (ARI) which relies upon statistical methods for measuring overlapping of comparative classes. The above two validation strategies are appropriate measures of goodness-of-fit and robustness of the final clusters. The external indexing techniques [18,19] applied for validation with reference to the real class labels of the dataset are Purity Measure [26], Random Index [27] and F-measure [26]. They are supervised methods of result validation and exploit the known information about a dataset for comparison purpose. The predictive capacity of a model is validated through probabilistic measures and external indexing techniques.

After treatment through the model, the class labels obtained for a significant partitions of the datasets can be accepted as reliable with credible class labels. As the framework relies both on unsupervised as well as supervised methods, it may be designated as a semi-supervised method of data clustering.

The article is structured as follows: first, the schematic description section shows the layout of the proposed integrated semi-supervised framework for class label determination for heterogeneous datasets. Second, the method selection and parameter discussion is presented which also highlights the clustering, classification, internal and external indexing mechanisms along with probabilistic measures with reasoning and justifications. In the experimentation section, the description of datasets along with stepwise empirical evaluation is discussed. The result analysis section critically evaluates the significance of the findings described in experimentation section. Finally, the conclusions of this work are summarized and future directions are highlighted.

2. Schematic description

Fig. 1 describes the schematic representation of the environment simulated for the work. It also identifies the datamining and validation techniques used for the same. Datasets with removed class labels are taken at first for the purpose of clustering. Multiple clustering techniques are applied on those datasets individually. The clustering results are then integrated through cluster ensemble technique on per tuple basis. Then based on the results, each dataset is segregated into the training set with majority agreed consensus cluster determined and testing set whose class/cluster labels are not yet known. Then, each training set is used to train multiple classifiers. The testing sets are now given to the classifiers for identification of their class/cluster labels. Again the consensus is taken to obtain a single class label for each tuple. The training tuples and the tuples with consensus in classification techniques are designated as the final dataset with known class/cluster labels. The remaining tuples with still ambiguity in their class determination are discarded from the dataset.

The final set of tuples with their class labels are then treated for verification and acceptability of the results
through external indexing techniques, purity and probabilistic measures. A dataset with unknown class labels has to undergo a set of unsupervised clustering techniques and another set of supervised classification techniques for knowing the class labels of its tuples which is a clustering task. Hence, the overall process is designated as a semi-supervised method of data clustering.

3. Method selection and parameter discussion

Heterogeneous data clustering problem begins with the initial challenge of identifying number of cluster, $k$. The framework used in the work becomes functional, once the number of clusters in a dataset are known. A lot many literature [28-34] address the issue of finding optimal value of $k$, which is a priory assumption for this work.

Five of the datasets are collected from UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science [35]. The datasets are from different domain and vary greatly in their characteristics. The Iris dataset comprises of 150 tuples belonging to three classes with four features describing the sepal and petal lengths and widths. This is the most structured dataset consisting of

![Fig. 1. Schematic representation of the semi-supervised method of data clustering.](image-url)
real values. The Wine dataset is meant for chemical analysis of wine quality for three types of wines. It has 13 features which are the chemical constituents of all the three types of wine. The Diagnostic Wisconsin Breast Cancer (WDBC) dataset has two classes such as Malignant, Benign. The features are characteristics of the cell nuclei collected from a digitized image of a fine needle aspirate test of a breast mass. The dataset comprises of 32 features of 569 instances. The Parkinson's disease detection dataset is composed of a range of biomedical voice measurements from 31 people. Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals. The data discriminates healthy people from those with Parkinson's disease. The Connectionist Bench (Sonar, Mines vs. Rocks) dataset is meant for classification of sonar signals obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. The label associated with each record is a rock or a mine (metal cylinder). The dataset comprises of 60 features of 208 patterns.

The clustering techniques used for the work are PSO-k-means [5,6], k-medoids [7], c-means [8] and Expectation Maximization (EM) [4] used for data clustering. The reason necessitating as follows:

i. The normal k-means clustering is a partition based heuristic technique, which may give different results for each iteration and the cluster quality is very much dependent on the initial choice of seeds and at times may converge with local optimas. In PSO-k-means, the seeds have been chosen by applying Particle Swarm Optimization technique [36], so that the above mentioned drawbacks can be overcome [5,6]. The PSO based searching procedure is described in equation (1) and the position of the particles are updated as per equation (2).

\[ V_i^{(t+1)} = \omega \cdot V_i^{(t)} + c_1 \cdot rand_i(o) \left( Pbest_i^{(t)} - X_i^{(t)} \right) + c_2 \cdot rand_i(o) \left( Gbest^{(t)} - X_i^{(t)} \right) \]

\[ X_i^{(t+1)} = V_i^{(t)} + X_i^{(t)} \]  

where, \( V_i^{(t)} \) is \( i \)th particle velocity in iteration \( t \), \( X_i^{(t)} \) is \( i \)th particle position in iteration \( t \), \( c_1 \), \( c_2 \) are constant weight factors, \( Pbest_i \) is local best position achieved so far by particle \( i \), \( Gbest \) is global best position achieved so far by particle \( i \), \( rand_i(o) \), \( rand(o) \): are random factors within the interval 0 to 1 and \( \omega \) is Inertia Weight.

ii. The k-medoids clustering is also a partition based technique but is unbiased towards outliers and its methodology is independent of the initial choice of seeds.

iii. The c-means clustering inherently makes use of fuzzy partitioning concept and determines the likelihood of each instance belonging to a particular cluster. In this parer, we have used the c-means clustering for crisp partitioning by taking maximum likelihood of cluster to be the clustering criteria.

iv. The EM technique used here for clustering takes care of the missing values if any and uses probabilistic maximum likelihood method to be the clustering criteria.

The PSO-k-means algorithm used here chooses randomly the population of size and iterates until convergence. The local and global optima are fixed at 0.1 both. The k-medoids algorithm adopts a random initialization process and is iterated until convergence. The c-means clustering algorithm though is basically a fuzzy clustering algorithm, here it is used for crisp clustering. The weighing exponent is taken 2 and error threshold is set to 0.01. The data tuple is assigned to that cluster whose likelihood probability is maximum. In EM method used in this work for data clustering, the error threshold is set to 0.001. All these afore said algorithms converge within 100 iterations.

The internal indexing techniques of cluster validation are preferred for the cluster quality verification because they are unsupervised methods and they determine the cluster qualities based on the feature characteristics of the datasets. The internal indexing techniques used here are Dunn's index, Davies–Bouldin index and Modified Good-man–Kruskal (GKmodified) index.

i. Maximized inter-cluster distance and minimized intra-cluster distance maximizes the contribution to the Dunn's index value. The cluster that obtains maximized index value is taken as optimal. For each cluster partition, the Dunn index can be calculated using equation (3):

\[ D = \min_{1 \leq i \leq n} \left\{ \frac{d(i,j)}{\max_{1 \leq k \leq n} d(k)} \right\} \]

where, \( d(i,j) \) represents the distance between clusters \( i \) and \( j \), \( d(k) \) measures the intra-cluster distance of cluster \( k \) and \( n \) represents the number of clusters.
ii. Davies–Bouldin index is determined by the ratio of the sum of the within-cluster scatter to between-cluster separation. Smaller values of the calculated index indicate good clustering. It can be calculated using equation (4):

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{i \neq j} \left\{ \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right\}$$  

(4)

where, $n$ is the number of clusters, $c_x$ is the centroid of cluster $x$, $\sigma_x$ is the average distance of all objects in cluster $x$ to centroid $c_x$, and $d(c_i, c_j)$ is the distance between centroids $c_i$ and $c_j$.

iii. The original Goodman–Kruskal indexing technique (Mishra et al., 2015) considers all possible quadruples for a given dataset. The quadruples may be either concordant or disconcordant, however the quadruples are disjoint in nature. A good cluster is one with many concordant and few disconcordant quadruples. In contrast Modified Goodman–Kruskal ($GK_{\text{modified}}$) index takes triplets instead of quadruples to validate the clusters by avoiding disjointedness. A good cluster is one with many concordant and few disconcordant quadruples. Let $N_c$ and $N_d$ denote the number of concordant and disconcordant triplets, respectively. The cluster with larger value of $GK_{\text{modified}}$ indicates a good clustering. Then the modified GK index, $GK_{\text{modified}}$, is defined as equation (5):

$$GK_{\text{modified}} = \frac{N_c - N_d}{N_c + N_d}$$  

(5)

A set of four prominent and widely used classification techniques such as Bayesian classifier, MLP with backpropagation used for learning, SVM and Decision tree classifiers are chosen for classification purposes.

i. Bayesian classifier is a probabilistic inference model which classifies the new incoming tuples based on the prior probabilities of training data assigned to individual classes. It is a mathematically verifiable model for classification.

ii. MLP is a neural network model which after being trained through backpropagation method, turns into a robust classification model resembling biological neurons. It also without any prior assumption can classify data with missing values and can deal uncertainties.

iii. SVM is a mathematical model used here as a multiclass classifier simulated in the form of an extension of binary classifier achieved through one-against-all method designed with a Radial Basis Function kernel. Though its training process is slow, still it is preferred as a good classification model because of its good generalization ability and as it performs better when the number of training samples are small.

iv. Decision Tree is a non-parametric, predictive model for classification of data which very much resembles to human inference method. It describes the training data characteristics and maps them to respective class labels with finite number of steps. Its classification performance is good even when the data contains missing values.

The neural network used here as MLP classifier with backpropagation method used for training the network, the network is dynamically constructed for each dataset. The number of input neurons are taken same as that of the number of attributes available in the dataset. The number of hidden layers is one with same number of neurons as that of input neurons. The output layer has only one neuron where class label is determined by the range of equi-spaced values separated within a range of $0 - 1$. The learning rate and acceleration coefficients are set at 0.02 and 0.01 respectively. The attribute selection method used in this work for Decision Tree classifier is Information gain.

For establishing the acceptability of the dataset with known class labels, the external indexing techniques, purity and probabilistic measures of cluster validation are used.

i. External indexing techniques used in this work are Random index, F-measure and Purity. The external validity methods evaluate the clustering based on the class information known earlier. Both Random index and F-measure compute the quality on per class/cluster basis.

a. The Random indexing technique computes the similarity of a clustering result to that of the known classification results. It may hence be regarded as a measure of correct decisions taken by the method under study. Rand index [27] measures the number of pair wise agreements between a clustering $K$ and a set of class labels $C$, normalized so that the
value lies between 0 and 1. It can be calculated using equation (6).

\[ R(C,K) = \frac{a}{a+b+c+d} \]  
(6)

where \( a \) denotes the number of pairs of points with the same label in \( C \) and assigned to the same cluster in \( K \), \( b \) denotes the number of pairs with the same label, but in different clusters, \( c \) denotes the number of pairs in the same cluster, but with different class labels and \( d \) denotes the number of pairs of different labels in \( C \) that were assigned to a different cluster in \( K \).

b F-measure is the harmonic mean of Precision and Recall which is suitable for clustering quality for imbalance datasets. The recall and precision of the cluster for each given class can be calculated from equations (7) and (8) respectively.

\[ \text{Recall}(i,j) = \frac{k_{ij}}{k_i} \]  
(7)

\[ \text{Precision}(i,j) = \frac{k_{ij}}{k_j} \]  
(8)

where \( k_{ij} \) is the number of objects of class \( i \) that are in cluster \( j \), \( k_j \) is the number of objects in cluster \( j \), and \( k_i \) is the number of objects in class \( i \).

ii Purity is a simple and straightforward method of cluster quality evaluation which computes the ratio of correctly assigned tuples to that of the total number of tuples for any clustering technique with respect to the known class labels. It assigns a single numeric value to measure all the clusters together, unlike that of Random index and F-measure which are computed for each class. For a cluster \( i \), the purity of the clusters may be computed as per equation (9).

\[ P_i = \frac{1}{k_i} \max_j (k_i^j) \]  
(9)

where, \( k_i^j \) is the number of objects in cluster \( i \) with class label \( j \). Purity of the entire clustering solution can be obtained from equation (10).

\[ \text{Purity} = \frac{1}{n} \sum_{i=1}^{k} \frac{n_i}{n} P_i \]  
(10)

where \( n_i \) is the size of cluster \( i \), \( k \) is the number of clusters, and \( n \) is the total number of data points. The F-measure of cluster \( j \) and class \( i \) is then calculated by equation (11).

\[ F(i,j) = \frac{2 \text{Recall}(i,j) \text{Precision}(i,j)}{\text{Precision}(i,j) + \text{Recall}(i,j)} \]  
(11)

iii Probabilistic measures used here to validate the clustering ability of the overall process and measure the cluster quality are NMI, NVI and ARI.

a NMI is a measure of the similarity between two candidate clustering outcomes. In this case, the result of the final clusters obtained are compared to the original class to which the instance belongs. When clusters are identical, the value approaches 1. It is computed on the basis of entropy of individual clustering technique. Let the information contained, which is also called entropy of cluster \( C_o \) be \( E(Co) = -\sum \frac{n_i}{n} \log \left( \frac{n_i}{n} \right) \) and \( C_c \) be \( E(Cc) = -\sum \frac{n_c}{n_c} \log \left( \frac{n_c}{n_c} \right) \). Similarly, the joint entropy of \( C_o \) and \( C_c \) be \( E(Co,Cc) = -\sum \frac{n_{oc}}{n} \log \left( \frac{n_{oc}}{n} \right) \). The NMI value may be computed as per the formula given in equation (12).

\[ \text{NMI}(Co, Cc) = \frac{E(Co) + E(Cc) - E(Co,Cc)}{\sqrt{E(Co) + E(Cc)}} \]  
(12)

b NVI is calculated on the basis of variation of information. It is an information theoretic approach for evaluation of comparable clustering techniques. NVI values decrease as class and cluster become more similar and that would be 0 when they are identical. NVI value may be computed as per the formula given in equation (13).

\[ \text{NVI}(Co, Cc) = 1 - \frac{2 \times [E(Co) + E(Cc) - E(Co,Cc)]}{E(Co) + E(Cc)} \]  
(13)

c ARI is a measure of the similarity between two data clusters. It has a value between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same. ARI value may be computed as per the formula given in equation (14). Let, \( C_o \) = original set of clusters, \( C_c \) = set of clusters obtained by applying clustering, \( n_o \) = number of objects in \( C_o \), \( n_c \) = number of objects in \( C_c \), and \( n_{oc} \) = the number of objects in \( |C_o \cap C_c| \).
No too much of parameter tuning is done for this work, as the framework is established for a generalization procedure. Domain specific parameter tuning could be done for achieving the acceptable threshold values of various validity methods.

4. Experimentation

The experimental setup simulated for empirically establishing the reliability of the work is based on the assumption that the number of classes for each dataset is known. Section 3 refers to several methods of finding the number of classes of a dataset whose number of classes are unknown. The datasets taken up for this work are all classified datasets with known class labels. To keep the experimental results verifiable, we have removed the class labels since beginning of the work and only use them for verification purpose. The flow structure of the entire experimentation is stated as in Fig. 2.

Also, the experimentation process is explained in stepwise process as follows:

**Step 1:** Five different datasets [35] from different domains as specified in Table 1 are taken so that the methods and findings can be generalized and would remain independent of application domains. The datasets with their number of instances, number of attributes and number of classes are described in Table 1. The class label of each dataset was removed, so that clustering can be applied on them.

**Step 2:** Four different clustering techniques such as PSO-\(k\)-means, \(k\)-medoids, \(c\)-means and Expectation Maximization are chosen for clustering the data. The diverged clustering techniques are expected to minimize data set specific clustering ability independent of the overall methodology. The clustering results on a per cluster basis instance assignment are given in Table 2.

**Step 3:** After applying the clustering techniques, the numeric random class labels assigned to each instance

\[
ARI(Co,Cc) = \frac{1}{2} \left[ \sum_{\alpha} \binom{\alpha C}{2} \right] - \frac{\left[ \sum_{\alpha} \binom{\alpha n}{2} \sum_{c} \binom{c n}{2} \right]}{\binom{n}{2}} \frac{\left[ \sum_{\alpha} \binom{\alpha c}{2} \right]}{n}
\]

Also, the experimentation process is explained in stepwise process as follows:

**Step 1:** Five different datasets [35] from different domains as specified in Table 1 are taken so that the methods and findings can be generalized and would remain independent of application domains. The datasets with their number of instances, number of attributes and number of classes are described in Table 1. The class label of each dataset was removed, so that clustering can be applied on them.

**Step 2:** Four different clustering techniques such as PSO-\(k\)-means, \(k\)-medoids, \(c\)-means and Expectation Maximization are chosen for clustering the data. The diverged clustering techniques are expected to minimize data set specific clustering ability independent of the overall methodology. The clustering results on a per cluster basis instance assignment are given in Table 2.

**Step 3:** After applying the clustering techniques, the numeric random class labels assigned to each instance

\[
ARI(Co,Cc) = \frac{1}{2} \left[ \sum_{\alpha} \binom{\alpha C}{2} \right] - \frac{\left[ \sum_{\alpha} \binom{\alpha n}{2} \sum_{c} \binom{c n}{2} \right]}{\binom{n}{2}} \frac{\left[ \sum_{\alpha} \binom{\alpha c}{2} \right]}{n}
\]

Also, the experimentation process is explained in stepwise process as follows:

**Step 1:** Five different datasets [35] from different domains as specified in Table 1 are taken so that the methods and findings can be generalized and would remain independent of application domains. The datasets with their number of instances, number of attributes and number of classes are described in Table 1. The class label of each dataset was removed, so that clustering can be applied on them.

**Step 2:** Four different clustering techniques such as PSO-\(k\)-means, \(k\)-medoids, \(c\)-means and Expectation Maximization are chosen for clustering the data. The diverged clustering techniques are expected to minimize data set specific clustering ability independent of the overall methodology. The clustering results on a per cluster basis instance assignment are given in Table 2.

**Step 3:** After applying the clustering techniques, the numeric random class labels assigned to each instance

\[
ARI(Co,Cc) = \frac{1}{2} \left[ \sum_{\alpha} \binom{\alpha C}{2} \right] - \frac{\left[ \sum_{\alpha} \binom{\alpha n}{2} \sum_{c} \binom{c n}{2} \right]}{\binom{n}{2}} \frac{\left[ \sum_{\alpha} \binom{\alpha c}{2} \right]}{n}
\]

Also, the experimentation process is explained in stepwise process as follows:

**Step 1:** Five different datasets [35] from different domains as specified in Table 1 are taken so that the methods and findings can be generalized and would remain independent of application domains. The datasets with their number of instances, number of attributes and number of classes are described in Table 1. The class label of each dataset was removed, so that clustering can be applied on them.

**Step 2:** Four different clustering techniques such as PSO-\(k\)-means, \(k\)-medoids, \(c\)-means and Expectation Maximization are chosen for clustering the data. The diverged clustering techniques are expected to minimize data set specific clustering ability independent of the overall methodology. The clustering results on a per cluster basis instance assignment are given in Table 2.

**Step 3:** After applying the clustering techniques, the numeric random class labels assigned to each instance

\[
ARI(Co,Cc) = \frac{1}{2} \left[ \sum_{\alpha} \binom{\alpha C}{2} \right] - \frac{\left[ \sum_{\alpha} \binom{\alpha n}{2} \sum_{c} \binom{c n}{2} \right]}{\binom{n}{2}} \frac{\left[ \sum_{\alpha} \binom{\alpha c}{2} \right]}{n}
\]

Also, the experimentation process is explained in stepwise process as follows:

**Step 1:** Five different datasets [35] from different domains as specified in Table 1 are taken so that the methods and findings can be generalized and would remain independent of application domains. The datasets with their number of instances, number of attributes and number of classes are described in Table 1. The class label of each dataset was removed, so that clustering can be applied on them.

**Step 2:** Four different clustering techniques such as PSO-\(k\)-means, \(k\)-medoids, \(c\)-means and Expectation Maximization are chosen for clustering the data. The diverged clustering techniques are expected to minimize data set specific clustering ability independent of the overall methodology. The clustering results on a per cluster basis instance assignment are given in Table 2.

**Step 3:** After applying the clustering techniques, the numeric random class labels assigned to each instance

\[
ARI(Co,Cc) = \frac{1}{2} \left[ \sum_{\alpha} \binom{\alpha C}{2} \right] - \frac{\left[ \sum_{\alpha} \binom{\alpha n}{2} \sum_{c} \binom{c n}{2} \right]}{\binom{n}{2}} \frac{\left[ \sum_{\alpha} \binom{\alpha c}{2} \right]}{n}
\]
of each dataset are relabeled such that a consensus may be taken from the result. The necessary relabeling of assigned cluster labels is done by finding closest distance from the cluster mean values obtained after applying PSO-k-means clustering technique which is taken as the base value. As per closeness from the base value, the relabeling of each class is done.

**Step 4:** No single clustering technique may be claimed to perform better as there may be biasing towards application specific datasets or particular clustering techniques may cluster specific class of data better. For generalized adoption of the methodology, an ensemble technique can be relied upon. In the present scenario, majority voting is adopted as the ensemble method because of its simple, straightforward approach and unbiased preference to each clustering technique. The results after applying majority voting on per class basis are shown in Table 3. It also contains the number of instances which could not obtain pure majority and hence are treated as testing tuples for the classification step.

**Step 5:** To verify the validity of the individual clustering results as well as the result of the ensemble clustering along with a comparison on clustering abilities, internal clustering techniques such as Dunn's, Davies Bouldin and Modified Goodman–Kruskal (GKmodified) indexing techniques have been applied. The results of Dunn's, Davies Bouldin and Modified Goodman–Kruskal (GKmodified) indexing techniques with respect to individual clustering technique along with the consensus clustering results are as presented in Tables 4–6 respectively.

### Table 2

<table>
<thead>
<tr>
<th>Dataset \ Clustering techniques</th>
<th>(k)-medoids</th>
<th>(PSO-k)-means</th>
<th>(c)-means</th>
<th>Expectation maximization</th>
</tr>
</thead>
<tbody>
<tr>
<td>(C_1)</td>
<td>(C_2)</td>
<td>(C_3)</td>
<td>(C_1)</td>
<td>(C_2)</td>
</tr>
<tr>
<td>Iris</td>
<td>50</td>
<td>31</td>
<td>69</td>
<td>50</td>
</tr>
<tr>
<td>Wine</td>
<td>71</td>
<td>53</td>
<td>54</td>
<td>66</td>
</tr>
<tr>
<td>WDBC</td>
<td>192</td>
<td>377</td>
<td>NA</td>
<td>177</td>
</tr>
<tr>
<td>Parkinson Disease</td>
<td>104</td>
<td>91</td>
<td>NA</td>
<td>98</td>
</tr>
<tr>
<td>Connectionist Bench (Sonar, Mines vs. Rocks)</td>
<td>130</td>
<td>78</td>
<td>NA</td>
<td>175</td>
</tr>
</tbody>
</table>

**Step 6:** The dataset after being treated with majority voting ensemble technique gets segregated to two partial datasets. One with pure majority, where pure majority means more than half (i.e. three votes). Four different classification techniques such as Bayesian classifier, MLP with Backpropagation used for learning, SVM and Decision tree classifiers are chosen for classifying the data. The diverged classification techniques are expected to neutralize data specific classification results. All the classifiers are simulated separately. The data tuples with class labels known from pure majority of ensemble method from each dataset are used as training tuples for each classifier. The remaining data tuples of each dataset are used for testing purpose. After the testing is done with each classifier, for each data tuple four predicted class labels are obtained. Again, a consensus is taken by majority voting otherwise known as bagging method. If a tuple is able to bag a pure majority consensus vote from the classifiers then the agreed upon class label is accepted as its final class. Out of the testing tuples, the classifiers as contributed to the ensemble method of majority voting are presented in Table 7.

**Step 7:** After this process is completed, still some data tuples are left out for some datasets, whose class labels can not be predicted with certainty. Those data tuples are discarded for acceptance as well as further analysis. The final set of tuples for which the class labels could be obtained and those a few for which the class label could not be obtained as per the semi-supervised method of data clustering are as shown in Table 8.

**Step 9:** Now each dataset has more than 99% or all of its tuples with known class labels obtained after treatment of the dataset with unsupervised followed by supervised method for finding the class labels which is evident from the results of Table 8.

**Step 10:** The validity of the results are established through external indexing techniques, purity and probabilistic measures. The external indexing techniques used are Random index, F-measure which give...
a supervised acceptance reason to the identified class labels. Purity gives a guided cluster quality determination ability to the method. The probabilistic measures such as NMI, NVI and ARI analyze the results from a information theoretic view point. The results of external indexing techniques and purity are as shown in Table 9.

The results of probabilistic measures are as shown in Table 10.

5. Result analysis

The intermediate and final results obtained from the experimental setup are presented in Tables 2–10. The facts and figures shown in the tables may be interpreted as follows.

- The clustering results shown in Table 2 state that no single clustering techniques is capable to
segregate the data as per its natural classes as specified in Table 1. For a particular class of a specific dataset one clustering technique may perform better. But the end result is data specific. To avoid the effect of random class label assignment, relabeling of the class labels is done before representing the results.

- To avoid data specific clustering ability, the ensemble technique i.e, majority voting is applied. For a particular tuple, if three of the clustering techniques agree in the class label assignment then a consensus is obtained. The result after consensus clustering on a per class basis is presented in Table 3. However, the tuples which could not obtain a majority might not be acceptable with a known class label.

- As the clusters are not compact enough, neither they are well separated. The Dunn's index values are too small as indicated in Table 4. It is also an indicator of the fact that it is difficult to separate such data through clustering techniques. The Davies Bouldin index values in Table 5 also indicates the same fact as a smaller value of index indicates better clustering. The G_k modified index considers point to point distance to determine cluster membership. The values in Table 6 state that the ensemble clustering is able to achieve satisfactory results.

- The partial data from each dataset whose class/cluster labels are known and subsequently used as training tuples for each classifier. The data tuples with unknown classes are treated as testing tuples.

Table 7 cites the number of tuples under each classification technique that could contribute to majority voting ensemble method. The impure tuples though classified by individual classifier but could not obtain agreed consensus results by other classifiers. The final outcome is represented in Table 8. Excepting one or two instances from all the datasets the class labels of almost all the tuples are known. For a dataset with no class labels known, such small amount of data may be discarded.

- Satisfiability about the acceptability of the results could be controlled through setting threshold values for validity indicators specified in Tables 9 and 10. Further enhancement and parameter tuning may be done to obtain satisfactory domain specific threshold levels for individual datasets under study.

- The generalized framework emphasizes on integration of unsupervised as well as supervised techniques for data clustering and hence exploiting their capabilities to segregate and determine the class labels of a dataset whose class labels are unknown. The overall process could be called a semi-supervised method of data clustering. The different validity measures may be treated as technical indicators of acceptability of the results.

6. Conclusion and future directions

The issue addressed in the current work is the usability of the clustering techniques on real world data.

---

**Table 8**
Classification results of the applied classification techniques.

<table>
<thead>
<tr>
<th>Dataset \ Clustering techniques</th>
<th>Classified</th>
<th>Unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>148</td>
<td>2</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>0</td>
</tr>
<tr>
<td>WDBC</td>
<td>569</td>
<td>0</td>
</tr>
<tr>
<td>Parkinson disease</td>
<td>194</td>
<td>1</td>
</tr>
<tr>
<td>Connectionist Bench</td>
<td>208</td>
<td>0</td>
</tr>
</tbody>
</table>

(Sonar, Mines vs. Rocks)

**Table 9**
Outcomes of applied external indexing techniques on the classified data.

<table>
<thead>
<tr>
<th>Dataset \ Clustering techniques</th>
<th>Random Index</th>
<th>F-measure</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C_1</td>
<td>C_2</td>
<td>C_3</td>
</tr>
<tr>
<td>Iris</td>
<td>1</td>
<td>0.8090</td>
<td>0.8411</td>
</tr>
<tr>
<td>Wine</td>
<td>0.9593</td>
<td>0.9323</td>
<td>0.9600</td>
</tr>
<tr>
<td>WDBC</td>
<td>0.9169</td>
<td>0.9555</td>
<td>NA</td>
</tr>
<tr>
<td>Parkinson Disease</td>
<td>0.6013</td>
<td>0.7404</td>
<td>NA</td>
</tr>
<tr>
<td>Connectionist Bench (Sonar, Mines vs. Rocks)</td>
<td>0.6290</td>
<td>0.4524</td>
<td>NA</td>
</tr>
</tbody>
</table>

**Table 10**
Outcomes of applied probabilistic measures on the classified data.

<table>
<thead>
<tr>
<th>Dataset \ Clustering techniques</th>
<th>NMI</th>
<th>NVI</th>
<th>ARI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>0.7289</td>
<td>0.4269</td>
<td>0.7154</td>
</tr>
<tr>
<td>Wine</td>
<td>0.8330</td>
<td>0.2862</td>
<td>0.8489</td>
</tr>
<tr>
<td>WDBC</td>
<td>0.6947</td>
<td>0.4680</td>
<td>0.7794</td>
</tr>
<tr>
<td>Parkinson Disease</td>
<td>0.2240</td>
<td>0.8746</td>
<td>0.1329</td>
</tr>
<tr>
<td>Connectionist Bench (Sonar, Mines vs. Rocks)</td>
<td>0.0214</td>
<td>0.9892</td>
<td>0.0086</td>
</tr>
</tbody>
</table>

(Sonar, Mines vs. Rocks)
Unless we know the class labels of the data tuples, classification cannot be done. However, when we do clustering of the data, the result remains unacceptable. The major challenges of clustering techniques are their clustering ability towards methodological bias, data and application domain specific bias and their varying function criteria. Data mining literature gives ample evidence where data segregation is done for diverse applications. But, when it comes to a generalized framework for class label assignment to data tuples, the research works are d than cited. The present work builds an integrated framework for assigning class labels to data tuples whose class labels are unknown and even they do not have any prior evidence of data tuples with class labels. The framework exploits the separability characteristics of multiple clustering techniques and enhances the reliability of the result through cluster ensemble method, which are unsupervised learning techniques. The obtained partial dataset with acceptable result is treated further for classifier training, which is a supervised learning techniques. Hence, the overall methodology may be designated as a semi-supervised framework for improving cluster reliability using ensemble methods for datasets whose class labels are unknown. Further refinement is done in increasing the tuples with known class labels by treating them as testing tuples. The acceptability is further enhanced by ensembling multiple classification techniques. Verifiability is established through external indexing techniques and probabilistic measures. However, the proposed framework comes with a constraint that it works well when the number of classes in the dataset is low, whereas for higher number of classes it may or may not hold good. This is because one class may not be identified at all by the clustering techniques through pure majority voting (i.e. possibility of class imbalance in the dataset). In case, the number of classes increase the classifier may not be able to deal with the class imbalance problem leading to small amount of training samples and large amount of testing samples. Further for unknown datasets internal indexing techniques can be used with user specific threshold values for acceptability of the results. Depending on the data characteristics, parameter tuning can also be done to achieve optimum threshold limits for index values.

References


[28] G.W. Milligan, M.C. Cooper, An Examination of Procedures for Determining the Number of Clusters in a Dataset, The Ohio State University, 1985, pp. 159–179, 50(2).


[33] M. Yan, Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion (Ph.D. thesis), Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2005.

