International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume6(2014) pp.494–504

Development of An External Cluster Validity Index using Probabilistic Approach and Min-max Distance

Abhay Kumar Alok[1], **Sriparna Saha**[2] , **Asif Ekbal**[3]

[1]Indian Institute of Technology Patna, Computer Science Engineering,
Software Technological Park of India , Patna 800013, India
abhayalok@iitp.ac.in

[2]Indian Institute of Technology Patna, Computer Science Engineering,
Software Technological Park of India , Patna 800013, India
sriparna@iitp.ac.in

[3]Indian Institute of Technology Patna, Computer Science Engineering,
Software Technological Park of India, Patna 800013, India
asif@iitp.ac.in

Abstract: **Validating a given clustering result is a very challenging task in real world. So for this purpose, several cluster validity indices have been developed in the literature. Cluster validity indices are divided into two main categories: external and internal. External cluster validity indices rely on some supervised information available and internal validity indices utilize the intrinsic structure of the data. In this paper a new external cluster validity index, MMI and its normalized version NMMI have been implemented based on Max-Min distance along data points and prior information using structure of data. A new probabilistic approach has been implemented to find the correct correspondence between the true and obtained clustering. Different possibilities for probabilistic approaches have been considered and tried to rectify their problems. Genetic K-means clustering algorithm (GAK-means) and single linkage clustering technique have been used as the underlying clustering techniques. Results of proposed index for classifying the true partitioning results have been shown for six artificial and two real-life data sets. GAK-means and single linkage clustering techniques are used as the underlying partitioning techniques with the number of clusters varied in a range. The MMI and NMMI index are then used to determine the appropriate number of clusters. Performance of MMI along with its two versions MMI_old and MMI_new along with its normalized version NMMI are compared with the existing external cluster validity indices, F-measure, purity, normalized mutual information (NMI), rand index (RI), adjusted rand index (ARI). Proposed MMI index works well for two class and multi class data sets.**

Keywords: Cluster validity, External cluster validity index, Genetic K-means clustering algorithm, Single linkage clustering.

I. Introduction

Clustering is used to partition the unlabelled data into different groups such that data objects within the same group are similar to each other according to some criteria of similarity and dissimilar to each other according to the same criteria [24]. Clustering is also known as unsupervised learning and mainly used in the fields of bioinformatics, web data analysis, text mining, and scientific data exploration. So in recent years many clustering algorithms have been developed [25, 26]. But the main challenging research question is which one of the clustering algorithms is best suitable for identifying the true partitioning of a given data set.

So cluster validity indices have been developed to validate the partitioning obtained by the clustering algorithms. In the literature there are mainly two different types of cluster validity indices available[11]: external and internal. External validity indices use the supervised information. These indices mainly quantify how good is the obtained partitioning with respect to prior class labelled information available. Adjusted rand index, Rand index, F-measure, Purity, NMI are some common examples of external validity indices. Internal validity indices are based on intrinsic information of the data. Most of the internal validity indices quantify how good a particular partitioning is in terms of the compactness and separation between clusters:

- Compactness: The proximity among the cluster elements can be measured using this. Variance is a commonly used measure of compactness.
- Separability: In order to differentiate between two clusters, this measure is used. One commonly used measure of separability is the distance between two cluster centers. This measure is easy to compute and can detect hyperspherical-shaped clusters well.

Among the internal cluster validity indices some well-known cluster validity indices are the BIC-index [13], CH-index [14], Silhouette-index [17], DB-index [15], Dunn-index [16], XB-index [19], PS-index [20], and *I*-index [18]. In this paper we have developed an external cluster validity index.

Here two new external validity indices, namely MMI[1] and NMMI, have been explored based on Max-Min distance between obtained clustered data and true prior available class labeled data. These validity indices depend on main three important factors given below:-

1. Relative ratio of misclassified data items.
2. True classified data items belonging to a particular class and cluster.
3. Distances between data items belonging to a particular class and cluster according to class and cluster contingency table.

Basically it gives a mapped value which quantifies how good is the obtained clustering result with respect to the true solution. Here in MMI [1] some changes have been employed to evaluate truly classified data items. Different probabilistic approaches are applied for evaluation of misclassified and truly classified data items. Results obtained through MMI[1] have been divided into two parts like MMI_old and MMI_new. These indices, MMI and NMMI, ensure the identification of the appropriate partitionings from data sets, i.e., its obtained maximum value for MMI-version and lowest value for NMMI ensure that cluster solution is very similar to true partitioning. For evaluation of the indices, genetic K-means (GAK-means)[2] and single linkage [6] clustering algorithms are used as the underlying clustering techniques. Results have been shown for six artificial and two real-life data sets. Comparisons have been done with five existing and well known external cluster validity indices, purity[9], F-measure[22], NMI[23], adjusted rand index(ARI)[3] and rand index[12].

This paper is organized as follows: Section II describes in detail the background and related works; Section III tells about the newly proposed external cluster validity index MMI; Section IV describes the newly proposed method of determining the correspondence between the true partitioning and the obtained partitioning; Section V discusses about the essential parameters related to MMI index and finally defines the index; Section VI describes in detail essential parameters related to normalized MMI index: NMMI; Section VII discusses the data sets used for experimental results; Section VII explores the experimental results; and finally Section IX concludes the paper.

II. Background and Related Work

In most of the papers, external cluster validity indices try to compare the true supervised information with the obtained clustering result[11]. But it has been seen that most of the cluster validity indices fail if there are some overlaps among original clusters in some data sets. For evaluation of the obtained cluster solution with respect to the true supervised information, different external cluster validity indices have been developed.

Rand Index (RI) proposed by Rand [12][24], is a popular cluster validation index given as:

$$RI = \frac{a+d}{a+b+c+d} \tag{1}$$

or

$$RI = \frac{TP+TN}{TP+FP+FN+TN} \tag{2}$$

Where TP shows true positive, TN is of true negative, FP is of false positive and FN is of false negative. These all are similar to a, b, c and d. RI value lies between [0,1]. When two solutions perfectly match, value of Rand index is 1. The problem associated with Rand index is that it does not show constant value for random partitions[4] . So Hubert and Arabic[3] overcome the deficiency of Rand index and assume randomness for partitions. So modified Rand index(RI) is defined as follows:

$$Adjusted\_Index = \frac{Index - ExpectedIndex}{Max\_Index - Expected\_Index}, \tag{3}$$

ARI

$$= \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}, \tag{4}$$

or

$$= \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}, \tag{5}$$

Here U denotes the set of all true classes and V denotes the set of all clusters; where a:- number of data items belonging to same class of U and same cluster of V, b:-number of data items that are placed in same class of U but different clusters of V, c:-number of data items that are placed in same cluster of V but different classes in U, d:-number of data items that are placed in different classes of U and in different clusters of V, and n_{ij} represents the total number of common data items which are in same class U_i and same cluster V_j. In contingency Table I, data item pairs have been shown corresponding to their true solution and obtained cluster solution. ARI[3] gives value between [0,1], 1 for best partitioning result and 0 for worst partition, -1 value shows random partitioning result.

F-measure: inherits the concepts of precision and recall from information retrieval [22]. For cluster V_j and class U_i

$$Recall(U_i, V_j) = \frac{n_{ij}}{n_i} \tag{6}$$

$$Precision(U_i, V_j) = \frac{n_{ij}}{n_j} \tag{7}$$

where n_{ij} equals number of data items belonging to class U_i and cluster V_j, n_j is the number of data items belonging to cluster V_j and n_i denotes number of data items belonging to class U_i.

$$F(U_i, V_j) = \frac{2 \times Recall(U_i, V_j) * Precision(U_i, V_j)}{Precision(U_i, V_j) + Recall(U_i, V_j)} \tag{8}$$

The F-measure value lies within [0,1] and high value indicates the higher clustering quality.

Purity[9]:- This external cluster validity index is similar to entropy. It counts the fraction of common data items of cluster V_j with respect to particular class U_i. Evaluation of

purity for obtained partitioning result is done on the weighted average of all individual cluster purities.

$$\textit{Purity} = \frac{1}{N}\sum_{j}\max_{i} \mid U_i \cap V_j \mid \tag{9}$$

NMI[23]:- Normalized Mutual Information for evaluation of true partitioning results can be formulated as below-

$$\textit{NMI}(U_i, V_j) = \frac{I(U_i, V_j)}{[H(V_j) + H(U_i)]/2} \tag{10}$$

where,

$$I(U_i, V_j) = \sum_{i}\sum_{j} P(U_i \cap V_j) log \frac{P(U_i \cap V_j)}{P(U_i)P(V_j)} \tag{11}$$

$$H(V_j) = -\sum_{j} P(V_j) log P(V_j) \tag{12}$$

Here I(U_i, V_j) denotes the mutual information between true assigned class and obtained cluster label, and H(V_j) is the entropy of cluster V_j while information about U_i classes are available .

In this paper a new approach to develop an external cluster validity index has been proposed. The newly developed cluster validity index MMI[1] and its normalized version NMMI are based on the unsupervised and supervised learning methods. Unsupervised learning process is based on the computation of Euclidean distance. Distances between every pair of clusters in U and V are computed. The maximum distance across columns and minimum distance across rows of distance matrix are calculated. Thereafter true classified data points n_{ij} and misclassified data points are determined based on different types of probabilistic approach. Essential parameters are calculated based on the above mentioned approaches to develop some new external cluster validity indices: MMI and NMMI.

III. Proposed External Cluster Validity Index: MMI

Let X be the set of N data items: X= { x_1, x_2,.....,x_N }. For evaluation of obtained partitioning result with appropriate number of clusters, K, agreement between cluster solution against external information is needed. Suppose here U = { u_1,u_2,.....,u_C } with C clusters, and V = { v_1,v_2,.....,v_K } with K clusters show two partitionings of X. Partitions should be like this $\bigcup_{i=1}^{C} u_i = \bigcup_{j=1}^{K} v_j$ = X, and $u_i \bigcap u'_i = \emptyset = v_j \bigcap v'_j$. The inequality constraints are $1 \leq i \neq i' \leq C$ and $1 \leq j \neq j' \leq K$. The corresponding contigency table which shows the matching between different clusters of U and V is provided in Table 1. This is a matrix of size $C \times K$. Here n_{ij} indicates number of pairs of data points which are in same ith cluster of U and same jth cluster of V.

Here first Euclidean distance between each pair of data items in cluster u_i of U and cluster v_j of V is computed. Let $x_i \in u_i$ and $y_j \in v_j$. Then Euclidean distance between(x_i,y_j) is computed as:

$$dist\{x_i, y_j\} = \left[\sum_{p=1}^{d} ||{x_i}^p - {y_j}^p||^2\right]^{\frac{1}{2}}, \tag{13}$$

Table 1: Contingency table for two partitions U and V

U/V	V_1	V_2	...	V_K
U_1	n_{11}	n_{12}	...	n_{1K}
U_2	n_{21}	n_{22}	...	n_{2K}
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
U_C	n_{C1}	n_{C2}	...	n_{CK}

Table 2: Distance Matrix between U and V partitions

U/V	V_1	V_2	...	V_K
U_1	[d_1,d_2,..dn_{11}]	[d_1,d_2,...dn_{12}]	...	[d_1,d_2,...dn_{1K}]
U_2	[d_1,d_2,..dn_{21}]	[d_1,d_2,..dn_{22}]	...	[d_1,d_2,..dn_{2K}]
.	.	.	...	.
.	.	.	...	.
.	.	.	...	.
U_C	[d_1,d_2,..dn_{C1}]	[d_1,d_2,..dn_{C2}]	...	[d_1,d_2,..dn_{CK}]

Here d is the dimension of the data items. Now dist(u_i,v_j)=[d_1, d_2,.....,dn_{ij}], where d_1= dist(x_1, y_1) and $x_1 \in u_i$, $y_1 \in v_j$. Now another distance matrix $S_{C\times K}$ is calculated as follows S_{ij}= max[d_1, d_2,.....dn_{ij}], where $1 \leq i \leq C, 1 \leq j \leq K$ and n_{ij} equals to total number of pairs of data points belonging to the ith partitioning of U and jth partitioning of V. This distance matrix is shown in Table 2.

Now based on these distances the compactness and separability of these partitions are calculated. After evaluation of maximum distance measurement, selection of maximum distance from S_{ij} matrix across each row is calculated. Let P be the maximum distance vector obtained where p_i= max[S_{i1}, S_{i2},.....S_{iK}] for $1 \leq i \leq C$. Similarly another matrix Q is constructed as follows: q_j= max[S_{1j}, S_{2j},.....S_{Cj}] for $1 \leq i \leq K$. Now R^{min}, and C^{max} are determined as follows:

$$R^{min} = \min_{i=1}^{C} P_i, \tag{14}$$

$$C^{max} = \max_{i=1}^{K} Q_i, \tag{15}$$

IV. Evaluation of Degree of Membership Between Obtained Partitioning and True Clustering

Here GAK-means and single linkage clustering algorithms are executed on all data sets. Thereafter obtained partitioning results are compared with available true partitioning information as shown in Table 1. Let $U = \bigcup_{i=1}^{C} U_i$ and $V = \bigcup_{j=1}^{K} V_j$ denote, respectively, the available true partitioning and obtained clustering results. With the help of membership functions of GAK-means/single linkage clustering technique common data points are evaluated for every U_i and V_j as shown in Table 1. Four different types of approach have been tried for extracting actual true classified data items.

case1:- Relative ratio of obtained common data points for every pair of class and cluster is calculated with respect to actual number of data points in the true class. Let T_i= [$\frac{n_{ij}}{\|U_i\|}$]is the vector of respective relative ratios where $1 \leq i \leq C$, $1 \leq j \leq K$ and $\|U_i\|$ are data points of true cluster i. Common points with high relative ratios are selected for every

class and cluster data item pairs. So common points n_{ij} are selected according to max[T_i]. But problem associated with this approach is that for some data sets accompanied with high relative ratios, the approach provides less number of truly classified data items. So it can cause poor evaluation results for some data sets.

case2:- Relative ratio of obtained common data points for every pair of class and cluster is calculated with respect to actual number of data points in the obtained cluster. Let T_i= [$\frac{n_{ij}}{\|V_j\|}$]is the vector of respective relative ratios where $1 \leq i \leq C, 1 \leq j \leq K$ and $\|V_j\|$ are data points of true cluster i. Common points with high relative ratios are selected for every class and cluster data item pairs. So common points n_{ij} are selected according to max[T_i]. But this approach also behaves like case 1 for some data set consisting of some symmetrically spherical shaped clusters.

case3:- Now the concept of geometric mean has been applied to find the correspondence between obtained common data points for class and cluster. Fraction of obtained common data points with geometric mean of their corresponding class and cluster label is calculated. Let T_i= $\frac{n_{ij}}{\sqrt{\|U_i\|}\sqrt{\|V_j\|}}$ is the vector of relative ratios where $1 \leq i \leq C, 1 \leq j \leq K$, and $\|U_i\|$, $\|V_j\|$ are the respectively true class data and cluster partitioned data. Here max[T_i] ensures the actual truly classified data items but also gives less number of truly classified data items for some data set.

case4:- Here combination of the above mentioned approaches have been incorporated. Now relative ratio of obtained common data points for every pair of class and cluster is calculated with respect to the total number of points in the obtained cluster. Let T_i= [$\frac{n_{ij}}{\|V_j\|}$] is the vector of respective relative ratios where $1 \leq i \leq C$, $1 \leq j \leq K$ and $\|V_j\|$ are data points of obtained cluster j. Common points with high relative ratios are selected for every class and cluster data item pairs. So common points n_{ij} are selected according to max[T_i]. But due to overlapping these selected common data points are not truly classified data items. To overcome this severe problem, again relative ratio of obtained common points is evaluated with true labelled data items. Let L_i=[$\frac{n_{ij}}{U_i}$] shows relative ratio of obtained common data points with their truly partitioned class labelled data items U_i. Here ith true cluster was selected for jth obtained partitioning by the first method. In case two obtained partitions got their maximum relative scores for the same true cluster we can resolve the tie by this method. The true cluster corresponds to that obtained partition for which L_i is maximum. This helps us to identify the correct correspondence in case of overlapping clusters Now number of true classified data items can be extracted as follows:

n_{ij}=$argmax_{j=1}^{K}(\mathbf{case1}, \mathbf{case2}, \mathbf{case3}, \mathbf{case4})$.

These extracted numbers correspond to the actual assigned cluster label data items and are more efficient to calculate the validity index.

V. Essential Parameters for Evaluation of Newly Developed Index:-MMI

External cluster validity index depends on the ratio of expected index with minor deviation to expected index with major deviation. So for evaluating MMI cluster validity index following parameters are defined as follows:

$\mathbf{a_1}$= Minor deviation of truly classified data points influenced by ratio of misclassified data points with max-min distance value from obtained maximum distance vector.

$\mathbf{b_1}$= Deviation of truly classified data points associated with maximum distance C^{max} due to misclassified data points.

$\mathbf{c_1}$= Major deviation of truly classified data points influenced by ratio of misclassified data points with max-min distance value from obtained maximum distance vector.

$\mathbf{d_1}$= Average deviation of truly classified data points from misclassified data points.

$\mathbf{e_1}$= Deviation of truly classified data points associated with minimum distance R^{min} due to misclassified data points.

Expected index with minor deviation= $\mathbf{a_1}$ + $\mathbf{b_1}$,

Expected index with major deviation= $\mathbf{c_1}$ + $\mathbf{d_1}$ + $\mathbf{e_1}$.

Above defined parameters are calculated as follows:

$$a_1 = \prod_{i=1}^{C}\prod_{j=1}^{K} n_{ij} - \{\frac{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}}{N} \times C^{max} \times R^{min}\}^2 \quad (16)$$

$$b_1 = \{\{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}\} \times C^{max} \times \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij} \times \sum_{i=1}^{C} P_i\}^{\frac{1}{2}} \quad (17)$$

$$c_1 = \prod_{i=1}^{C}\prod_{j=1}^{K} n_{ij} - \{\frac{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}}{N} \times C^{max} \times R^{min}\}^3, \quad (18)$$

$$d_1 = C^{max} \times R^{min} \times \{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}\} + \frac{\sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij} \times \{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}\}}{2} \quad (19)$$

$$e_1 = \{\sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij} \times \{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}\} \times R^{min} \times \sum_{i=1}^{C} P_i\}^{\frac{1}{2}} \quad (20)$$

where $\prod_{i=1}^{C}\prod_{j=1}^{K} n_{ij}$ shows multiplication of truly classified data points which belong to same class and cluster, $\{\frac{N-\sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}}{N}\}$ is relative ratio of misclassified data points present in total N data points in a dataset. $\sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}$ equals to summation of data points belonging to same class-cluster pairs, $\sum_{i=1}^{C} P_i$ equals the summation of the maximum Euclidean distance in distance matrix.

Now MMI cluster validity index can be defined as follows:-

$$MMI = \frac{a_1 + b_1}{c_1 + d_1 + e_1} \quad (21)$$

MMI value lies between [0,1]; value of 1 indicates true partitioning results, value of 0 is its lower value of clustering result and if somehow -1 value is obtained than it is totally worst result.

VI. Normalized External Cluster Validity Index:-NMMI

For evolution of true clustering result essential factor is to calculate truly classified object pairs, misclassified data items, relative ratio of misclassified data items in contrast with their unsupervised property like max-min distance separation among class cluster object pairs. So for normalized version of MMI say NMMI, essential parameters are defined as follows:

x = filter of misclassified data items with respect to total sum of truly classified data items

y = summation of maximum separable distance across class-cluster contingency table associated with truly classified data items. Separable distance ensures to prevent overlap between clusters and leads to selection of true cluster result.

z = selection of misclassified data items with relative ratio of deviation while keeping their maximum and minimum separation distances while finding correspondence between class-cluster data pairs.

$$x = \sqrt{\sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}} + [N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}] \tag{22}$$

$$y = \sqrt{\prod_{i=1}^{C}\prod_{j=1}^{K} n_{ij}} + \sum_{i=1}^{C} P_i \tag{23}$$

$$z = \frac{C^{max} + R^{min} + [N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}]}{\frac{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}}{N}} \tag{24}$$

where $\{\frac{N - \sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}}{N}\}$ is the relative ratio of misclassified data items, $\sum_{i=1}^{C}\sum_{j=1}^{K} n_{ij}$ equals the summation of data points belonging to same class-cluster pairs, $\sum_{i=1}^{C} P_i$ equals the summation of the maximum Euclidean distance in distance matrix.

So NMMI can be defined as follows:-

$$NMMI = \frac{x}{y + z} \tag{25}$$

NMMI value lies between [0,1]. Value close to 0 gives the truly partitioning results, higher value does not give the good evaluation result.

VII. Data Set Used for Experiment

Here six artificial data sets and two real-life data sets are used for evaluation. AD_10_2, AD_5_2, AD_6_2, AD_4_3, $Mixed\_5\_2$, AD_9_2, are artificial data sets and Iris, Newthyroid belong to real-life data sets.

1. AD_10_2: This data set, used in [7], consists of 500 two-dimensional data points and is distributed over 10 different clusters. Each cluster consists of 50 data points and, clusters are highly overlapping in nature. This data set is shown in Figure 3(a).

2. AD_5_2: This data set, used in [8], consists of 250 two-dimensional data points and is distributed over 5 different spherically shaped clusters. These clusters are highly overlapping and, each consists of 50 data points. This data set is shown in Figure 3(b).

3. AD_6_2: This data set consists of 300 two-dimensional data points, and is distributed over 6 different hyper-spherical shaped clusters. Each cluster consists of 50 data points. This data set is shown in Figure 3(c).

4. AD_4_3: This data set has 400 three-dimensional data points, and is distributed over four clusters. Each cluster consists of 100 data points. This data set is shown in Figure 3(d).

5. AD_9_2: This data set consists of 900 two-dimensional data points and is distributed over nine clusters. This data set is shown in Figure 3(f)

6. $Mixed\_5\_2$: This data set, used in [21], consists of 850 two-dimensional data points and is distributed over five spherically shaped clusters. This data set is shown in Figure 3(e).

7. Iris: This data set, obtained from [10], consists of 150 four-dimensional data points, and is distributed over three different clusters. Each cluster consists of 50 data points. It shows four different categories of irises. Out of three classes, two of them, Vergincia and Versicolor, are highly overlapping with each other and last one Setosa is linearly separable from others.

8. Newthyroid: This data set, obtained from [10] consists of total 215 data items with five features. Three classes are classified as euthyro idism, hypothyroidism or hyperthyroidism under the five laboratory test results.

VIII. Experimental Result

Here Genetic K-means algorithm(GAK-means)[2] is used for partitioning the data sets. Number of clusters is varied from K=2 to K=12 for artificial data sets and K=2 to K=7 for real-life data sets. To determine the appropriate partitioning from these data sets, we have computed the values of MMI_old, MMI_new, NMMI, F-measure, Purity, NMI, ARI and RI for different number of clusters K. The partitioning for which a particular index obtains its maximum value is considered to be the optimum partitioning. The values of different indices MMI_old[1], MMI_new, NMMI, F-measure, Purity, NMI, ARI and RI for different values of number of clusters are shown in Tables 3, 4 and 5 respectively for artificial and real-life data sets. For data set AD_10_2, RI index attains its maximum value for $K = 10$ and $K = 11$; thus it identifies $K = 10$ and $K = 11$ as the optimal number of clusters. The corresponding value of

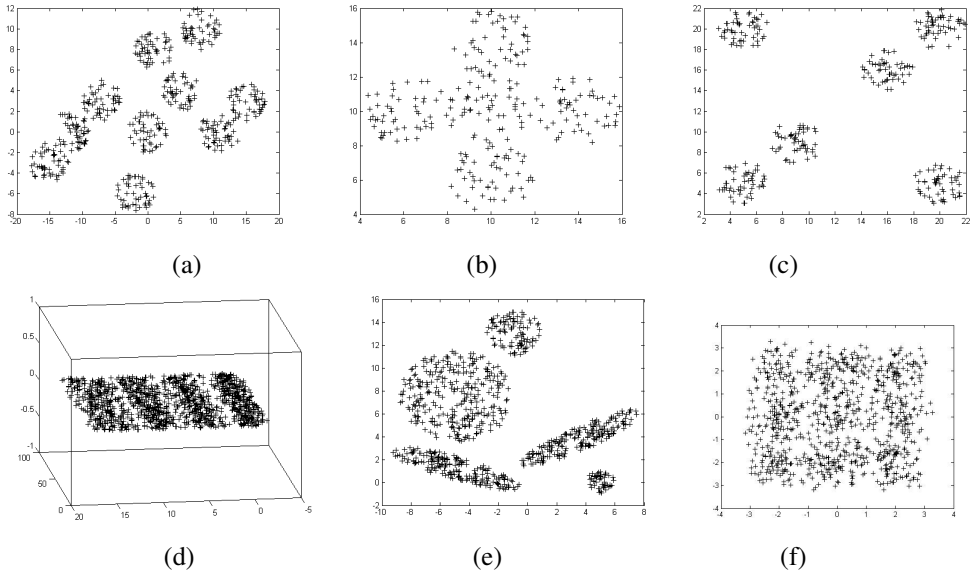

Figure. 1: Artificial Data Sets (a) AD_10_2 (b) AD_5_2 (c) AD_6_2, (d)AD_4_3 (e) $Mixed\_5\_2$, (f)AD_9_2

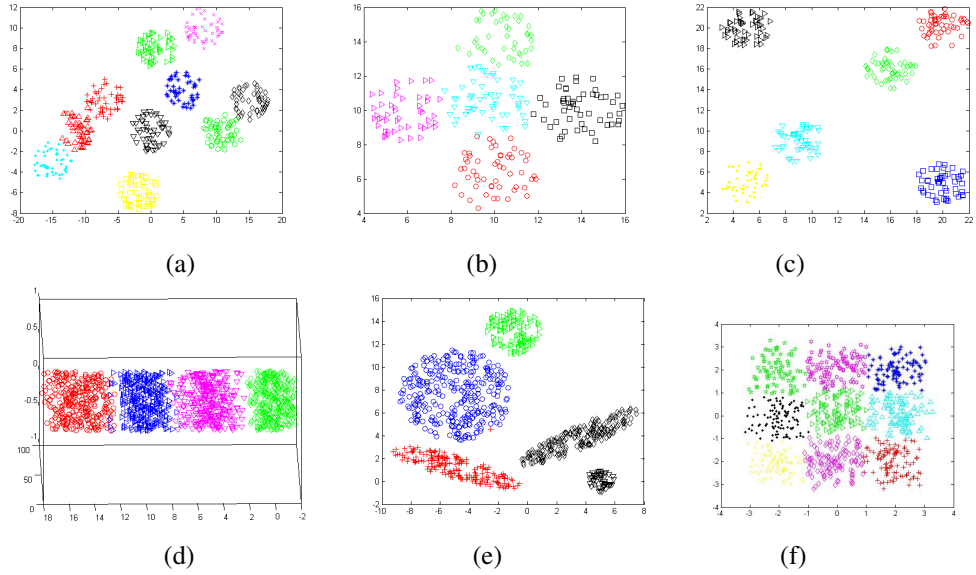

Figure. 2: Partitioning results obtained after application of GAK-means clustering technique on (a) AD_10_2 (b) AD_5_2 (c) AD_6_2, (d)AD_4_3, (f)AD_9_2, and Single linkage on (e)$Mixed\_5\_2$ corresponding to the highest values of MMI, ARI, RI, F-measure, Purity, NMI and lowest value of NMMI

RI index is 0.983038. For data set AD_4_3, F-measure attains its highest value 0.750000 at K=5, NMI also attains its highest value 0.943452 at $K = 5$. Thus according to these indices $K = 5$ is the optimal number of clusters. But actual partitioning result comprises of 4 different clusters. For data set AD_9_2, both RI and ARI indices attain their highest value say 1.000000 with cluster number 8 and 9, respectively. These have been shown in Table 3. But there are actually 9 different clusters in this data set. For Iris data set NMI index attains its highest value 0.667305 with $K = 4$. Thus optimum value of NMI index suggests that there are 4-different clusters but actual number of clusters in this data set is 3. Results have been shown in Table 4. For Newthyroid data set F-measure attains its highest value 0.201219 at $K = 5$ and NMI index obtains its highest value 0.390785 at $K = 4$. Thus according to F-measure Newthyroid has 5 different clusters and according to NMI it has 4 different clusters. But the true partitioning of this data set has 3 different clusters. Results have been shown in Table 4. Optimum partitionings identified by the MMI-index and NMMI index are shown in Figure 2 for all artificial data sets. Graph plots of values obtained by different indices with respect to number of clusters for different data sets are shown in Figure 3. Results show that MMI and NMMI indices are also better than ARI. It is evident from the results on AD_5_2. Figure 2(b) shows the partitioning obtained by GAK-means for K=5 on this data set. After visual inspection we can say that this partitioning is very near to the optimal one. But ARI obtains value of 0.853434, whereas two different versions of MMI provide values of 0.999877 and 0.931684, respectively and NMMI index attains value of 0.000109. This proves that NMMI and MMI-versions are more robust in detecting the

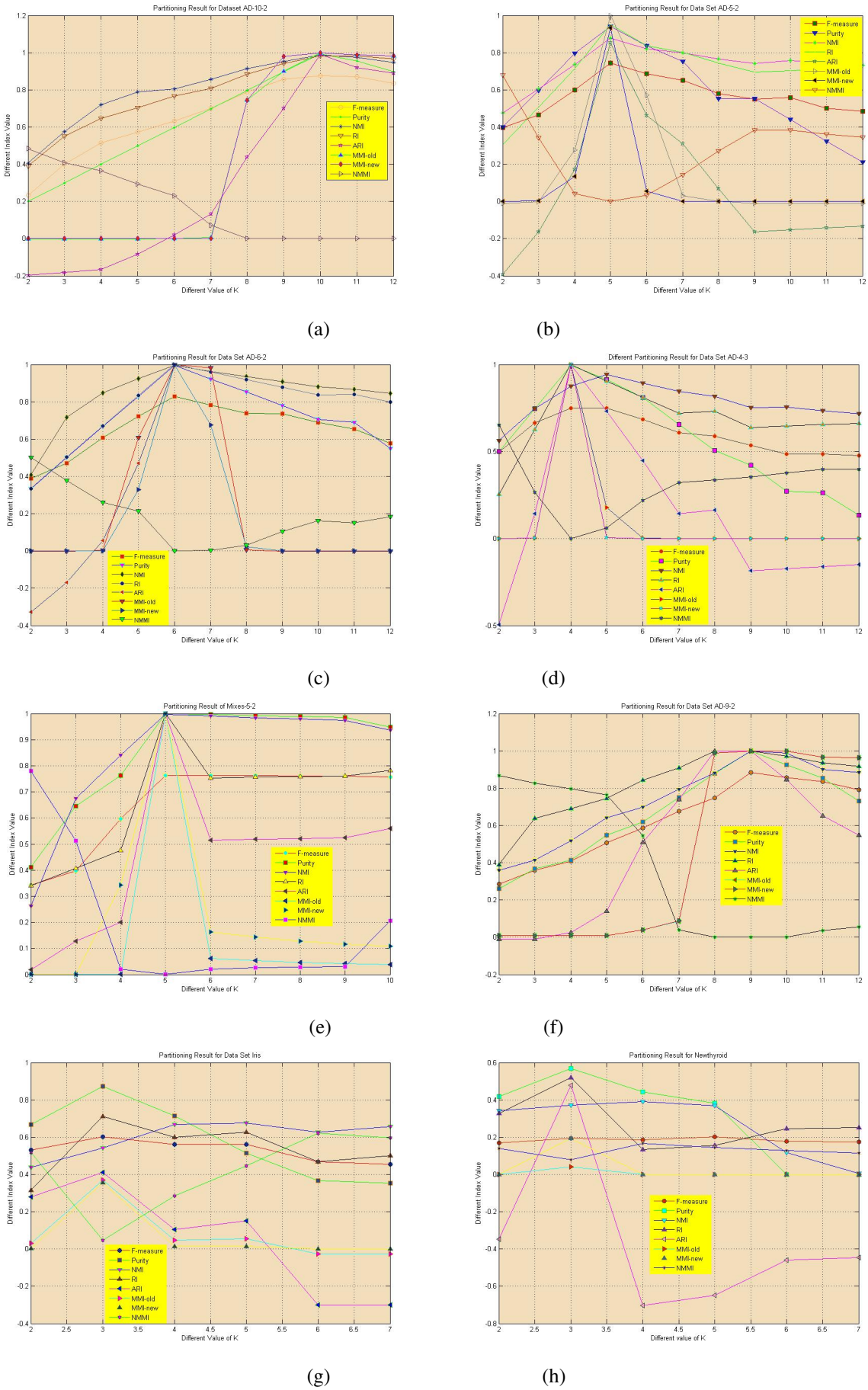

Figure. 3: Graphs showing the obtained values of cluster validity indices versus number of clusters for data sets (a) AD_10_2 (b) AD_5_2 (c) AD_6_2, (d)AD_4_3 (e) $Mixed\_5\_2$, (f)AD_9_2 ,(g)Iris and (h) Newthyroid

optimal partitioning. Results have been shown in Table 5.

Here again Single linkage algorithm[6] is used for partitioning the data set $Mixed\_5\_2$. Number of clusters is varied from K=2 to K=10. The values of different indices F-measure, Purity, NMI, MMI_old, MMI_new, ARI and RI for different values of number of clusters are shown in Table V, respectively for this artificial data set $Mixed\_5\_2$. Figure 2(f) shows the partitioning obtained by Single linkage for K=5 on this data set. Number of misclassified data points is one in this optimal partitioning result. MMI-versions provide values of 0.999786 and 0.999922, respectively and NMMI provides value of 0.000109 whereas ARI and RI both provide value of 1.00000. This shows that MMI-version and NMMI are more robust in detecting the optimal partitioning. As there is one misclassified data item, MMI versions are not getting values of 1. In this new probabilistic approach of finding correspondence between obtained partitioning and true partitioning, MMI_new dominates over MMI_old[1] with respect of determining of true partitioning from different data sets.

IX. Conclusion

In this paper a novel approach has been adopted to develop a new external cluster validity index, named MMI and its normalized version say NMMI to quantify the obtained partitionings after application of a given clustering technique with respect to the true partitioning. It works well for two class and multi-class data sets. We have used different probabilistic concepts to identify the correspondence between the obtained and true partitionings. MMI_new ensures that the MMI index value lies between [0,1] with respect to the truly classified data items in comparison with MMI_old. NMMI index always works well for true optimum solution. A genetic K-means clustering algorithm (GAK-means) and Single linkage clustering technique are used as the underlying partitioning techniques.

Results and comparison graph plots of different index values are shown for six artificial and two real-life data sets. NMMI and MMI-version index are always able to determine the optimum partitioning for these data sets. Results are compared with already existing external cluster validity indices F-measure, Purity, NMI, adjusted rand index (ARI), and rand index (RI).

It has been observed in some cases that values of F-measure, NMI, Purity, RI, ARI for optimum partitioning are lower or higher than corresponding MMI-version and NMMI values. Use of both true classification information of all data items and also some unsupervised information help MMI-version and NMMI to obtain the highest value for optimum partitioning. For worst partitionings MMI obtains lower values and NMMI obtains higher values close to 1.

References

[1] A. K. Alok, S. Saha and A. Ekbal, *A min-max distance based external cluster validity index: MMI*. In Hybrid Intelligent Systems (HIS), IEEE 12th International Conference, pp. 354-359, December 2012.

[2] K. Krishna, M. Murty, *Genetic K-means algorithm*, IEEE Trans. Syst., Man, Cybern. B, Cybern., 29(3), pp. 433 - 439, 1999.

[3] L. J. Hubert, P. Arabie, *Comparing partitions*, Journal of Classification, 2(1), pp. 193 -218, 1985.

[4] J.M. Santos, M. Embrechts, *On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification*,ICANN, Artificial Neural Networks, ICANN, pp. 175 - 184, 2009.

[5] K.Y. Yeung, W.L. Ruzzo, *Details of the adjusted rand index and clustering algorithms. supplement to the paper an experimental study on principal component analysis for clustering gene expression data*, Bioinformatics 9(17) , pp. 763 - 774, 2001.

[6] Jain, A. K., Murty, M.N. Flynn, P.J, *Data clustering: A review*, ACM Computer. Surveys 31 (3), pp. 264 - 323, 1999.

[7] S. Bandyopadhyay, S. K. Pal, *Classication and Learning Using Genetic Algorithm: Application in Bioinformatics and Web Intelligence*, Springer, Heidelberg, 2007.

[8] S. Bandyopadhyay, U. Maulik, *Genetic Clustering for automatic evolution of clusters and application to image classication*, Pattern Recognition, 2(2002), pp. 1197 - 1208, 2002.

[9] E.Rendon, I.Abundez, A.Arizmendi, E.M.Quiroz, *Internal versus external cluster validation indexes*, International Journal of Computers and Communications 5(1), pp. 27 - 34, 2011

[10] http://archive.ics.uci.edu/ml

[11] M. Halkidi, Y. Batistakis, M. Vazirgiannis, *On Clustering Validation Techniques*, Intelligent Information Systems, pp. 107 - 145, 2001.

[12] W. M. Rand, *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association , pp. 846 - 850, 1971.

[13] A.Raftery, *A note on Bayes factors for log-linear contingency table models with vague prior information*, Journal of the Royal Statistical Society. 48(2), pp. 249 - 250, 1986.

[14] R.B.Calinski, J.Harabasz, *A Dendrite Method for Cluster Analysis*, Communication in Statistics Simulation and Computation, Vol. 3(1), pp. 1 - 27, 1974.

[15] D. L. Davies, D. W. Bouldin, *A cluster separation measure*, IEEE Trans. Pattern Anal. Machine Intell., 1(4), pp. 224 227, 1979.

[16] J. C. Dunn, *A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters*, J. Cybern., 3(3), pp. 32 57, 1973.

Table 3: Computed values of F-measure, Purity, NMI, RI, ARI, MMI, NMMI indices with different values of number of clusters, K, for Artificial data sets when GAK-means is used as the underlying clustering technique

Data set	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
AD_10_2	K=2	0.232670	0.200000	0.405863	0.389980	0.197288	-0.001324	0.000061	0.484093
	K=3	0.399084	0.300000	0.576266	0.549764	-0.181418	-0.001343	0.000062	0.409725
	K=4	0.513832	0.400000	0.720318	0.646244	-0.165335	-0.00132	0.000062	0.364346
	K=5	0.572640	0.500000	0.789261	0.704810	-0.084125	-0.001380	0.000122	0.292242
	K=6	0.633669	0.598000	0.804696	0.767825	0.019063	0.001456	0.000241	0.230100
	K=7	0.700220	0.698000	0.857865	0.808569	0.131703	0.007062	0.000999	0.071607
	K=8	0.787203	0.798000	0.915233	0.883559	0.438978	0.744340	0.743367	0.001710
	K=9	0.857307	0.898000	0.953484	0.942485	0.700072	0.899798	0.980791	0.000197
	K=10	**0.875010**	**0.994000**	**0.988616**	**0.983038**	**0.991084**	**0.999999**	**0.999999**	**0.000001**
	K=11	0.871000	0.954000	0.974763	**0.983038**	0.919900	0.989444	0.989444	0.000178
	K=12	0.834959	0.898000	0.947892	0.966116	0.888888	0.983647	0.983647	0.002181
AD_5_2	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
	K=2	0.397714	0.400000	0.475800	0.305221	-0.392451	-0.010820	0.000549	0.677894
	K=3	0.465900	0.596000	0.605798	0.507277	-0.162124	-0.002890	0.002150	0.342402
	K=4	0.599254	0.796000	0.737453	0.713060	0.172911	0.276355	0.1348898	0.042124
	K=5	**0.744365**	**0.940000**	**0.878824**	**0.953510**	**0.853434**	**0.999877**	**0.931684**	**0.001900**
	K=6	0.687464	0.836000	0.821830	0.838490	0.462769	0.571607	0.054731	0.032528
	K=7	0.651268	0.752000	0.799924	0.800964	0.308924	0.031852	0.001286	0.142945
	K=8	0.581553	0.552000	0.766465	0.743743	0.068070	0.001428	0.000016	0.272488
	K=9	0.550421	0.552000	0.742524	0.694104	-0.163308	-0.010820	0.000548	0.383433
	K=10	0.557782	0.440000	0.759351	0.704225	-0.151700	-0.010820	0.000545	0.383439
	K=11	0.499668	0.324000	0.743406	0.712739	-0.141497	-0.010820	0.000521	0.362360
	K=12	0.485010	0.212000	0.734569	0.719550	-0.133028	-0.010811	0.000051	0.344573
AD_6_2	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
	K=2	0.389018	0.333333	0.408449	0.335095	-0.327979	-0.002107	0.001054	0.502297
	K=3	0.470556	0.500000	0.717614	0.503902	-0.170229	-0.002364	0.000833	0.379295
	K=4	0.608889	0.666677	0.847498	0.670011	0.054444	0.002633	0.000545	0.259691
	K=5	0.72222	0.830000	0.925549	0.835006	0.470353	0.608066	0.328981	0.033617
	K=6	**0.830000**	**0.996667**	**0.994825**	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**0.000000**
	K=7	0.783550	0.923333	0.963916	0.959220	0.845960	0.983158	0.676604	0.003489
	K=8	0.739246	0.853333	0.935249	0.918172	0.679871	0.006656	0.0219708	0.031316
	K=9	0.736398	0.780000	0.908142	0.877191	0.501660	-0.002414	0.000542	0.104917
	K=10	0.689462	0.706667	0.882385	0.836343	0.309973	-0.002414	0.000309	0.162423
	K=11	0.653115	0.690000	0.867828	0.841093	0.320855	-0.0023819	0.000287	0.151950
	K=12	0.578329	0.550000	0.844935	0.799688	0.110093	-0.002235	0.000469	0.184111
AD_4_3	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
	K=2	0.497776	0.500000	0.562345	0.253133	-0.495623	-0.001521	0.000761	0.654163
	K=3	0.663744	0.747500	0.745689	0.626566	0.143974	0.002305	0.002234	0.264910
	K=4	0.748744	**0.995000**	0.874543	**1.000000**	**1.000000**	**1.000000**	**1.000000**	**0.000000**
	K=5	**0.750000**	0.912500	**0.943452**	0.904060	0.732719	0.179385	0.004856	0.060903
	K=6	0.686244	0.810000	0.893421	0.810075	0.447567	0.002151	0.000357	0.219938
	K=7	0.608553	0.655000	0.845462	0.720489	0.143734	-0.001861	0.000519	0.320509
	K=8	0.589577	0.507500	0.817537	0.731905	0.163977	-0.001849	0.000476	0.336074
	K=9	0.536792	0.420000	0.752375	0.637268	-0.185953	-0.001521	0.000761	0.353894
	K=10	0.487112	0.272500	0.754152	0.647368	-0.172434	-0.001521	0.000076	0.377144
	K=11	0.487112	0.262500	0.735821	0.654411	-0.162734	-0.001521	0.000007	0.396685
	K=12	0.478384	0.135000	0.718333	0.662426	-0.151366	-0.001521	0.000057	0.396687
AD_9_2	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
	K=2	0.285111	0.26000	0.359702	0.388990	-0.010586	0.009320	0.009320	0.869118
	K=3	0.358264	0.367382	0.413303	0.638282	-0.009224	0.009321	0.009321	0.827821
	K=4	0.408621	0.413333	0.516467	0.689247	0.025272	0.009321	0.009337	0.796973
	K=5	0.506406	0.547778	0.641392	0.745611	0.139379	0.009321	0.009398	0.764963
	K=6	0.586715	0.617778	0.699262	0.843970	0.509873	0.039758	0.039758	0.544740
	K=7	0.676495	0.748889	0.793442	0.909195	0.738936	0.084977	0.089767	0.040404
	K=8	0.747031	0.877778	0.881010	**1.000000**	**1.000000**	0.988095	0.988095	0.001469
	K=9	**0.884444**	**0.998880**	**0.998452**	**1.000000**	**1.000000**	**1.00000**	**1.000000**	**0.000000**
	K=10	0.856419	0.924444	0.987624	0.970458	0.845317	0.988999	0.999876	0.001867
	K=11	0.835657	0.853333	0.899504	0.936258	0.652702	0.968755	0.966743	0.037120
	K=12	0.791489	0.730000	0.883586	0.918342	0.548303	0.953395	0.963395	0.055777

Table 4: Computed Values of F-measure, Purity, NMI, RI, ARI, MMI, NMMI indices with different number of clusters (K) for Real-life data sets when GAK-means is used as the underlying clustering technique

Dataset	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
Iris	K=2	0.530373	0.666667	0.437563	0.315441	0.279232	0.030780	0.002054	0.518562
	K=3	**0.602943**	**0.873333**	0.542764	**0.711136**	**0.411854**	**0.370015**	**0.356167**	**0.046082**
	K=4	0.561181	0.713333	**0.667305**	0.600296	0.104867	0.046616	0.014823	0.286323
	K=5	0.561181	0.513333	0.677302	0.627483	0.151115	0.054077	0.013020	0.446234
	K=6	0.467052	0.366667	0.627699	0.467672	-0.299095	-0.025795	0.001369	0.621545
	K=7	0.455766	0.353333	0.657293	0.499651	-0.299095	-0.025791	0.001360	0.595419
Newthyroid	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
	K=2	0.170603	0.418685	0.342427	0.327190	-0.347957	-0.000291	0.000148	0.139021
	K=3	0.195297	**0.567442**	0.370785	**0.519278**	**0.478610**	**0.040400**	**0.194261**	**0.079206**
	K=4	0.186356	0.441860	**0.390785**	0.132623	-0.704227	-0.00018	0.000093	0.166352
	K=5	**0.201219**	0.381395	0.369231	0.156401	-0.649533	-0.000170	0.000085	0.144232
	K=6	0.176483	0.000231	0.117932	0.245333	-0.460122	0.000170	0.000007	0.127143
	K=7	0.175890	0.000001	0.006392	0.252597	-0.444837	0.000170	0.000007	0.115446

Table 5: Computed values of F-measure, Purity, NMI, RI, ARI, MMI, NMMI indices with different values of number of clusters, K, for Artificial data set $Mixed\_5\_2$ when single linkage is used as the underlying clustering technique.

Dataset	K	F-measure	Purity	NMI	RI	ARI	MMI_old	MMI_new	NMMI
$Mixed\_5\_2$	K=2	0.342674	0.411765	0.263188	0.340687	0.018197	0.000256	0.000616	0.779549
	K=3	0.398207	0.645882	0.674339	0.406834	0.127287	0.000487	0.000141	0.513656
	K=4	0.596387	0.763529	0.840466	0.474818	0.200897	0.000578	0.341904	0.021391
	K=5	**0.762657**	**0.998824**	**0.997496**	**1.000000**	**1.000000**	**0.999786**	**0.999922**	**0.000109**
	K=6	0.762156	0.996471	0.992154	0.753058	0.513732	0.060946	0.162768	0.020932
	K=7	0.761759	0.991765	0.983824	0.75631	0.51862	0.053217	0.144129	0.025540
	K=8	0.761558	0.989412	0.979184	0.757936	0.521059	0.046657	0.127883	0.028770
	K=9	0.761254	0.985882	0.972778	0.760337	0.524698	0.042213	0.116677	0.029991
	K=10	0.757812	0.948235	0.935943	0.783126	0.560012	0.038687	0.107625	0.206314

[17] P.Rousseeuw, Silhouettes, *a graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics, 20(1), pp. 53 - 65, 1987.

[18] U. Maulik, S. Bandyopadhyay , *Performance evaluation of some clustering algorithms and validity indices*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), pp. 1650 - 1654, 2002.

[19] X. Xie, G. Beni, *A Validity Measure for Fuzzy Clustering*, IEEE Transactions, Pattern Analysis and Machine Intelligence (PAMI), 13(8), pp. 841 - 847, 1991.

[20] C. H. Chou, M. C. Su, E. Lai, *Symmetry as a new measure for cluster validity*, in: Second WSEAS International Conference on Scientic Computation and Soft Computing, pp. 209 - 213, 2002.

[21] S. Bandyopadhyay, S. Saha, *A point symmetry based clustering technique for automatic evolution of clusters*, IEEE Transaction on Knowledge and Data Engineering, 20(11), pp. 1 - 17, 2008.

[22] B. Larsen, C. Aone, *Fast and effective text mining using linear-time document clustering*, In proc. 5th ACM SIGKDD Int. conf. on Knowledge Discovery and Data Mining, pp. 16 - 22, 1999.

[23] A. Strehl, J. Ghosh, *Cluster Ensembles - a knowledge reuse framework for combining multiple partitions*, Journal on Machine Learning Research(JMLR) 3, pp. 379 - 423, 2002.

[24] D. Tsarev, M. Petrovskiy, I. Mashechkin, *Supervised and Unsupervised text classification via generic summarization*, International Journal of Computer Information Systems and Industrial Management Application(IJCISIM), 5, pp. 509 - 515, 2013.

[25] R. Sarmah, *Gene expression data clustering using a fuzzy link based approach*, IJCISIM, 5, PP. 532 - 541, 2013.

[26] K. Stoffel, P. Cotofrei, D. Han, *Fuzzy clustering based methodology for multidimensional data analysis in computational forensic domain*, 4, pp. 400-410, 2012.

Author Biographies

Abhay Kumar Alok received the B.Tech. degree in computer science and engineering from U.P Technical University Lucknow, India, in 2008 and the M.Tech. degree in Information Technology specialized in HCI(Human Computer Interaction) from the Indian Institute of Information Technology, Allahabad, India, in 2010. He is currently working towards the Ph.D. degree at the Indian Institute of Technology , Patna, India. His research interests include multiobjective optimization, pattern recognition, evolutionary algorithms, and data mining, Image Processing.

Dr. Sriparna Saha received her Masters' and Ph.D. degrees in Computer Science from Indian Statistical Institute Kolkata, India, in 2005 and 2011, respectively. She is currently an Assistant Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India. She has authored or coauthored more than 60 papers. Her current research interests include pattern recognition, multiobjective optimization and biomedical information extraction.

She is the recipient of the Lt Rashi Roy Memorial Gold Medal from the Indian Statistical Institute for outstanding performance in MTech (computerscience). She is the recipient of the Google India Women in Engineering Award,2008. She received India4EU fellowship of the European Union to work as a Post-doctoral Research Fellow in the University of Trento, Italy from September 2010-January 2011. She is also the recipient of Erasmus Mundus Mobility with Asia (EMMA) fellowship of the European Union to work as a Post-doctoral Research Fellow in the Heidelberg University, Germany from September 2009-June 2010.

Dr. Asif Ekbal received his Masters' and Ph.D. degrees in Computer Science from Jadavpur University, Kolkata, India, in 2004 and 2009, respectively. He is currently an Assistant Professor in the Department of Computer Science and Engineering, Indian Institute of Technology Patna, Patna, India. He has authored or coauthored more than 70 papers. His current research interests include natural language processing, biomedical information extraction and machine learning.

Dr. Ekbal is the recipient of best Innovative Project Award from the "Indian National Academy of Engineering (INAE)" in 2000. He received India4EU fellowship of the European Union to work as a Post-doctoral Research Fellow in the University of Trento, Italy from September 2010-January 2011 and from May, 2011-July, 2011. He is also the recipient of Erasmus Mundus Mobility with Asia (EMMA) fellowship of the European Union to work as a Post-doctoral Research Fellow in the Heidelberg University, Germany from September 2009-June 2010.