A New Clusterwise Similarity for Partitions based on Quantitative Disagreement

P Rajasekhara
CA Technologies Private Ltd.
115, IT Park Area,Nanakramguda, Gachibowli,
Hyderabad-500 019

Arun K Pujari*
Sambalpur University
Institute of Information Technology
Sambalpur, 768019

ABSTRACT
Combining the results of different clustering to get a consensus clustering has attracted the attention of data mining researchers. In this context, it becomes necessary to measure diversity (or similarity) of a pair of partitions. Several diversity indices exist and these are based either on pairwise agreement or on clusterwise agreement. In pairwise agreement approach, similarity of two clusters is the number of common pairs of data elements. However, it is equally important to measure the level of disagreement rather than counting the frequency of disagreed pairs. We formulate this problem as a Transportation Problem and use Northwest Corner rule to compute feasible significance measures. We use this idea to propose a new index which differs from the existing measures in evaluating the extent of agreement by measuring the disagreement of data-pairs in terms of distance between cluster-pair of the disagreed data. We show experimentally that this yields a far better diversity index.

Keywords
Clustering, Diversity Index, Machine Learning

1. INTRODUCTION
Data mining has emerged as one of the important areas of research in last decade and in this process, many clustering algorithms are developed in recent years. Clustering is essentially to partition the set of points so that similar points are together and dissimilar points are separated. Different clustering algorithms [10] generate different partitions based on their own strategy of joining or of separation. Since no single clustering strategy is universally suitable, combining the results of several clustering strategies has become one prime area of investigations [3, 4]. The aim is to determine a consensus partition by analyzing multiple partitions obtained from different clustering sessions in order to improve the quality and robustness. In this context, one of the important problems of investigation is to measure the diversity (or equivalently, similarity) of partitions. This is significant because a measure of agreement is required to compare the result of a clustering algorithm with the ground truth partition and to compare results of two different clustering. There are two main families of methods for comparing partitions- one evaluating the pairwise agreement, the other searching for the clusterwise agreement. Clusterwise methods use the *contingency table* that contains the dual classification of each individual data point in two partitions. On the other hand, pairwise methods are computed from a *mismatch matrix* derivable from the contingency table.

In the present work, we propose a new and improved method in pairwise agreement category. The well-known measures in this category are Rand index [12], Jaccard index [1], Adjusted Rand index [5], and Wallace index [15]. There has also been proposal to use mutual information. Although there are several measures of comparing partitions, there is no consensus on choice of a method. Denoeud et al [2] report a comparative study of these indices based on transfer distance between partitions. This analysis is one of the motivating factors of the present research.

The motivation to develop a new method stemmed from the observation that all the previous approaches attempt to measure the agreement (or, equivalently disagreement) based on agreement of pairs of data between partitions. In a naïve term, two clusters are similar if they contain same pairs of data elements. However, it is more appropriate to determine similarity based on agreement of individual data elements. That is, two clusters are to be similar when they contain same set of data elements. Though this looks obvious and natural, it is difficult to measure the diversity of the whole partition if we consider individual data elements. Hence majority of the earlier methods are based on either data-pairwise similarity or cluster-pairwise similarity. The proposed index, *IRM-index*, differs from the existing measures in evaluating the extent of agreement between any two groupings, taking into account intercluster similarities. It differs from the other known methods by measuring the disagreement of data pair in terms of distance between cluster-pair of the disagreed data. The similarity measures are compared with ground truth partition. The main idea is to measure a significance of agreement between clusters and this can be formulated as a

---

(*)Corresponding author

ICVGIP '10, December 12-15, 2010, Chennai, India

set of inequalities. Incidentally, the system of inequalities so generated has the similar structure as the well-known Transportation Problem and hence its solution can be found by any of the well-known methods such as Matrix Minimum method, North-West Corner rule etc. We use here north-west corner rule to get a solution of the system inequalities.

A brief account of existing diversity measures is given in section 2. We introduce the new index, called as IRM-index, in section 3. We report our experiments in Section 4.

2. SIMILARITY MEASURES

2.1 Notations and Definitions

Let S be the set of data points. $\mathbf{S} = \{\mathbf{s_1}, \mathbf{s_2}, \ldots, \mathbf{s_n}\}$. A partition **P** on **S** is defined as $\mathbf{P} = \{\mathbf{C_1}, \mathbf{C_2}, \ldots, \mathbf{C_p}\}$ such that $\mathbf{C_i} \subseteq \mathbf{S}$, $\mathbf{C_i} \cap \mathbf{C_j} = \varnothing$ and $\cup \mathbf{C_i} = \mathbf{S}$. Let $\mathbf{P} = \{\mathbf{C_1}, \mathbf{C_2}, \ldots, C_p\}$ and $\mathbf{Q} = \{\mathbf{C'_1}, \mathbf{C'_2}, \ldots, \mathbf{C'_q}\}$ be two partitions on **S**. **P** and **Q** are said to agree on a pair of data points $(\mathbf{s_i}, \mathbf{s_j})$ if there exists indices **k** and **m** such that $\mathbf{s_i}, \mathbf{s_j} \in \mathbf{C_k}$ and also, $\mathbf{s_i}, \mathbf{s_j} \in \mathbf{C'_m}$. For two partitions **P** and **Q**, let $\boldsymbol{\nu_{11}}$ denote the number of pairs of data points that are in the same cluster in partition **P** as well as in **Q**. Similarly, we can define three other quantities, $\boldsymbol{\nu_{00}}$ is the number of pairs of data points that are in different clusters in **P** as well as in **Q**; $\boldsymbol{\nu_{10}}$ is the number of pairs of data points that in same cluster in **P** but in different clusters in **Q** and $\boldsymbol{\nu_{01}}$ is the number of pairs that are in different clusters in **P** but in same clusters in **Q**. We can construct the contingency table **N** where the ij[th] entry $\mathbf{N_{ij}}$ is the number of data items in $\mathbf{C_i} \cap \mathbf{C'_j}$. Let n_{i*} and n'_{*j} be number of items present in $\mathbf{C_i}$ and $\mathbf{C'_j}$, respectively.

$$\sum_{j=1}^{q} N_{ij} = n_{i*}, \quad \sum_{i=1}^{p} N_{ij} = n'_{*j} \quad n = \sum_{i=1}^{p} n_{i*} = \sum_{j=1}^{q} n'_{*j}$$

2.2 Known Diversity Indices

This section describes several existing pairwise similarity measures of comparing partitions.

The Rand Index [12]

Rand index is a measure of agreement between partitions with values in [0, 1]. Rand index is defined as follows.

$$Rand(P, Q) = \frac{\nu_{00} + \nu_{11}}{\nu_{00} + \nu_{01} + \nu_{10} + \nu_{11}}$$

The Jaccard Index [1]

Unlike the Rand index where the pairs simultaneously joined or separated are treated similarly, the Jaccard index ignores the pairs of data points that are separated in both partitions. The Jaccard index is defined as follows.

$$J(P, Q) = \frac{\nu_{11}}{\nu_{00} + \nu_{01} + \nu_{10}}$$

The Adjusted Rand Index [5, 13]

There are many variants of the Rand index. An important variant, Adjusted Rand index corrects for the lack of a constant value of the Rand index when partitions are selected at random [5]. Let us first define the following quantities.

$$t_0 = \sum_{i,j} \binom{N_{ij}}{2}, \; t_1 = \sum_{i=1}^{p} \binom{n_{i*}}{2}, \; t_2 = \sum_{j=1}^{q} \binom{n'_{*j}}{2}, \; t_3 = \frac{t_1 \times t_2}{\binom{n}{2}}$$

The Adjusted Rand index is defined as [13]

$$AR(P,Q) = \frac{t_0 - t_3}{\frac{1}{2}(t_1 + t_2) - t_3}$$

The Wallace Index [15]

Wallace index is the ratio of number of joined pairs common to **P** and **Q** to the number of possible pairs. This denominator depends on the partition of reference and, if we do not want to favour neither **P** nor **Q**, the geometrical average is used.

$$W(P,Q) = \frac{\nu_{11}}{\sqrt{t_1 t_2}}$$

The Normalized Mutual Information [4, 14]

From contingency table the value of mutual information between P and Q is.

$$MI(P,Q) = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}}{n} \log\left(\frac{n_{ij} n}{n_{i*} n_{*j}}\right)$$

Measure of similarity between partitions is the Normalized Mutual Information given as

$$NMI(P,Q) = \frac{-\sum_{i=1}^{p} \sum_{j=1}^{q} \frac{n_{ij}}{n} \log\left(\frac{n_{ij} n}{n_{i*} n_{*j}}\right)}{\sum_{i=1}^{p} n_{i*} \log\left(\frac{n_{i*}}{n}\right) + \sum_{j=1}^{q} n_{*j} \log\left(\frac{n_{*j}}{n}\right)}$$

3. PROPOSED METHOD

Methods that compare partitions based on pairwise agreement consider whether pairs are in same clusters in different partitions or not. We observe that when a pair is not in agreement it is important to quantify the level of disagreement. This can be done by determining the closeness of the cluster pairs to which the disagreed data belong. Thus the pairwise agreement methods can utilize this information to measure the similarity among partitions. Our motivation of a new index stems from this observation. All known methods are, in a sense, binary as they determine the pairs in agreement or not. Our method tries to differentiate different levels of disagreement. IRM-**index** considers matching of clusters in **P** with all the other

clusters in **Q**. We denote d_{ij} as the distance between clusters $\mathbf{C_i}$ of **P** and $\mathbf{C'_j}$ of **Q** and define it as follows.

$$d_{ij} = \frac{|C_i \cap C'_j|}{|C_i \cup C'_j|}$$

Thus the similarity (or diversity) measure between P and Q should ideally be the sum total of all d_{ij}. We introduce a significance based weighting and the significance values are computed based on a sort of matching. A matching between clusters $\mathbf{C_i}$ and $\mathbf{C'_j}$ assigns a significance credit $\boldsymbol{\sigma_{ij}}$ where $\boldsymbol{\sigma_{ij}} \geq 0$. Thus the proposed diversity index, IRM-index, between two partitions is defined as follows.

$$IRMindex(P,Q) = \frac{1}{n}\sum_i \sum_j d_{ij}\,\sigma_{ij}$$

Thus significance credits $\boldsymbol{\sigma_{ij}}$ can be determined by the following set of inequalities.

$$\sum_{j=1}^{q} \sigma_{ij} = n_{i*}, \quad 1 \leq i \leq p, \quad \sum_{i=1}^{p} \sigma_{ij} = n_{*j}, \quad 1 \leq j \leq q \qquad \sigma_{ij} \geq 0$$

This becomes similar to the "Transportation Problem" constraints and any feasible solution of transportation problem will determine the significant values. We use Northwest Corner rule to get a feasible solution (the Initial Basic Feasible Solution in the context of the Transportation Problem). The following algorithm describes the process.

compute_significance

1. **initialize** ∀ i and j, labeled(i,j) = 0 and $\sigma_{ij} = 0$
2. **do while** there is no (i, j) with labeled(i,j) = 0
3. select the unlabeled index-pair (i,j) corresponding to the largest $\mathbf{d_{ij}}$
4. set $\sigma_{ij} \leftarrow$ min ($\mathbf{n_{i*}}$, $\mathbf{n_{*j}}$)
5. update labeled(i,j) ←1
6. update $\mathbf{n_{i*}} \leftarrow \mathbf{n_{i*}} - \sigma_{ij}$
7. update $\mathbf{n_{*j}} \leftarrow \mathbf{n_{*j}} - \sigma_{ij}$
8. if $\mathbf{n_{i*}} = 0$
9. update labeled(i, u) ←1, ∀ u
10. if $\mathbf{n_{*j}} = 0$
11. update labeled(v, j) ←1, ∀ v
12. **end dowhile**

The proposed method is similar to Integrated Region Matching (IRM) technique [9] which is a similarity measure for pairs of segmented images. IRM overcomes the difficulty of inaccurate segmentation in images by allowing one region of image to be matched with several other regions of other image. This philosophy can be applied in the present context as there is always the problem of inaccurate clustering. The algorithm *compute_significance* outlines the steps to compute the significance value satisfying the above set of inequalities.

Illustrative Example

In order to see the desirable properties of the proposed diversity measure, let us consider an example given in Figure 1. There are three different partitions P_1, P_2 and P_3 for eight data items. These partitions are
P_1 = {1,2}|{3}|{4}|{5}|{6}|{7,8};
P_2= {1,2}|{3}|{4}|{5}|{6}|{7}|{8}; and
P_3 = {1,2}|{3,6}|{4,5}|{7,8}.
We can see that P_1 is more similar to P_2 than it is to P_3. P_1 and P_2 differ only in one cluster whereas there are several other clusters in P_3. Thus it is desirable that any measure of similarity between pair of partitions should highlight this aspect. The values of measures for these pairs of partitions are summarized in Table 1.

Table 1: Distance indices between P_1, P_2 and P_1, P_3

Index	P_1, P_2	P_1, P_3
Jaccard	0.50	0.50
Wallace	0.70721	0.70721
AR	0.65	0.63157
Rand index	0.9642	0.8571
IRM-index	**0.875**	**0.75**
NMI	0.0757575	0.05555

It is observed that the Jaccard index, the Wallace index and the Adjusted Rand index determine both P_2 and P_3 equidistant from P_1. These measures do not take into account singleton clusters while computing similarities. In a situation where large number of clusters differ among themselves by single elements, these indices (except the Rand index) will not yield desired result. And the Rand index fails when there are many of nonintersecting pairs in both the clusters because it gives equal importance to pairs in same clusters, and pairs in different cluster. The Adjusted Rand index and the Normalized Mutual Information can take negative values and the behavior of NMI is unpredictable which can be observed from the experimental results.

4. EXPERIMENTAL RESULTS

In order to evaluate the discriminatory ability of the proposed index, we carried out experiments using data with known partitions from UCI/ML repository. We use k-means clustering method to obtain candidate partition. We determine the distance of the computed partition from the known partition by using all known diversity indices. A major design issue in the present context is the determination of reference partition in the absence of any "ground truth". In such cases normally one chooses one of the known method as reference and tries to compare other techniques with reference to this chosen one. We experimented with several known technique as references. We observed that all other indices exhibit consistent behavior with respect to Hungarian method. Thus reference distance is determined by computing the distance of two partitions using Hungarian method [6, 7].

This distance is normally used to measure the accuracy of the clustering technique. Our experiment is to determine

the index that is close to the accuracy. Thus for a given partition, obtained from one clustering session, we compute the accuracy by Hungarian technique and then compute the similarity measure using indices described earlier and then find the difference of accuracy and similarity. This quantity, accuracy – similarity, is the performance measure of an index for a given partition. A desirable property of an ideal index is to have smaller deviation from the accuracy. We use k-means algorithm for different values of k and for each k, we take the average of the deviation for 50 runs with random seed points. Thus for a fixed k, we take 50 different partitions generated by 50 runs of k-means and compute all the diversity indices for each of these partitions. Figures 2-5 show the graphs with k as the abscissa and deviation from accuracy value on the ordinate.

The Australia Dataset

The dataset consists of 690 data items in 2 clusters. We experimented for k from 2 to 50. The deviation of the Rand index is very low when k is near 2 and jumps to high values as k increases. The deviation of NM index is very high when k=2 and gradually reduces as k increases. IRM-index and the Wallace index show much balanced behaviour. It is observed that IRM-index has less deviation. Adjusted Rand index and the NMI are almost similar to each other. Figure 2 gives the performance of different indices for this data set.

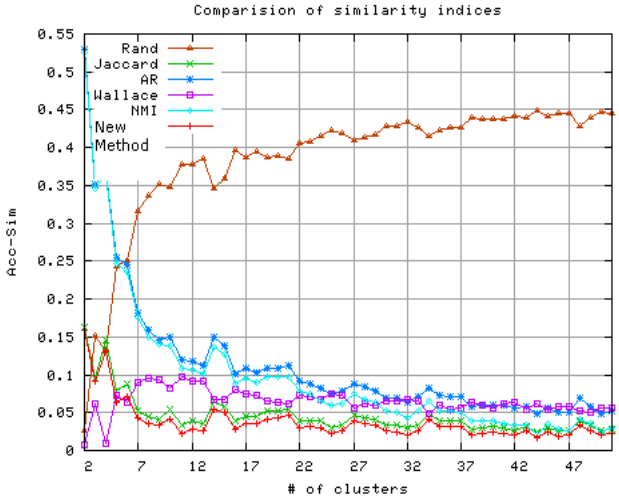


Figure 2. Experimentation and comparision for Australia Dataset

The Iris Dataset

The Iris dataset is of 150 data items with 3 clusters of which two of them are very similar. We take k in [3, 50]. The Rand index has larger accuracy whereas the Adjusted Rand index gives better results for initial values of k. The IRM-index demonstrates the distances in a balanced manner. It can be observed that the proposed method always supersedes Jaccard index which considers number of intersection by union of data pairs. Figure 3 summarizes the experiments for this data set.

The Spect Dataset

The Spect dataset consists of 267 data items with two clusters of which 187 are training and 80 are testing. The IRM-index is closer to real similarity. For this data set the Adjusted Rand index and NMI demonstrate very undesirable behaviour; the diversity is very high for correct value of k and decreases for higher k values. The Jaccard, Wallace and the IRM-index are acceptable in this case with IRM-index giving smallest deviation throughout. The detail result is depicted in Figure 4.

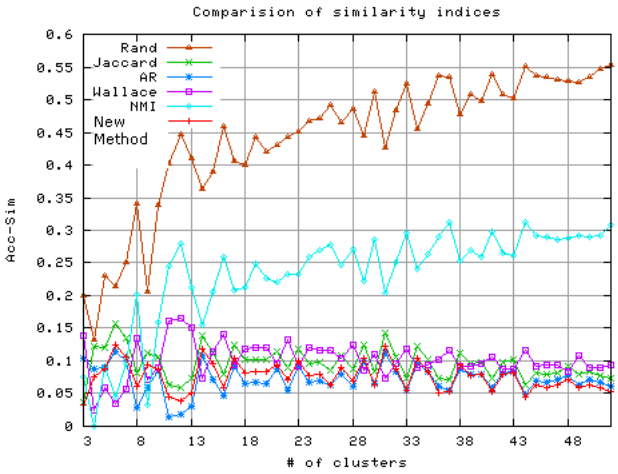


Figure 3: Comparison of similarity indices for Iris dataset.

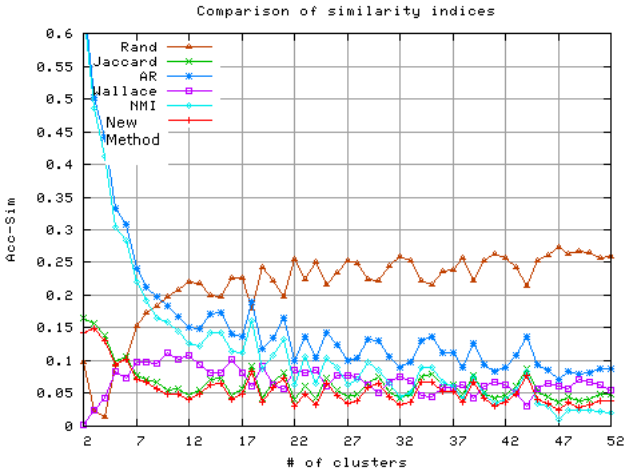


Figure 4: Comparison of similarity indices for Spect dataset.

The Liver disorders Dataset

This dataset consists of 345 data items separated into two clusters. It can be observed that with increasing cluster numbers IRM-index was much closer to accuracy values (Figure 5).

Results for different datasets are formulated in Table 2 below where k is the true number of clusters in the dataset and k_1 is the cluster number for which the IRM-index is better. For every dataset 50 partitions were created with number of clusters starting with k_1 and incrementing by one cluster and results were compared with the existing measures. We take the sum of deviation (= accuracy - similarity) values starting from k_1 clusters to 50 clusters. We observe that the IRM-index has smallest deviation. As the number of clusters increases the deviations of Jaccard, Adjusted Rand, Wallace and IRM indices tend to be closer to zero. Although for few datasets initially other measures were better, there is no consensus on choice of measure that best suites any given data. IRM measure overcomes this difficulty and can be considered to work for different kinds of data. Experimental results prove that IRM-index is even better choice when the number of clusters is large.

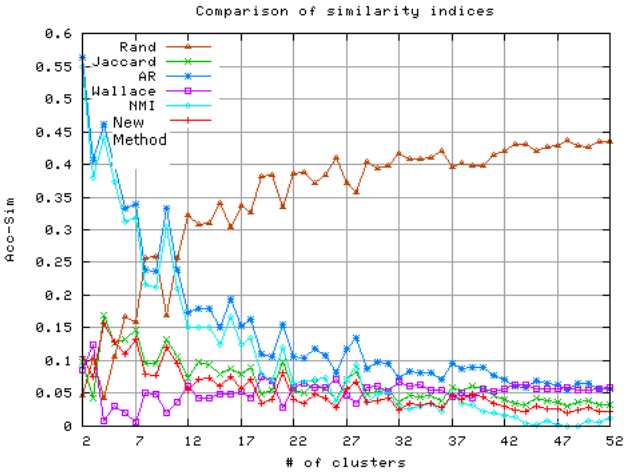


Figure 5. Comparison of similarity indices for Liver disorders dataset.

5. CONCLUSION

We, in this paper, develop a new diversity measure to compare partitions. The diversity measure takes into account the membership of individual data element to same or different clusters in a pair of partitions. This is different and more appropriate than the known methods which take into account the pairs of data element sharing same clusters. We show through elaborate experiments that our diversity measure is close to the accuracy. Thus the proposed measure will be very useful for consensus clustering. We propose to investigate this aspect in future. In the same line as reported in [11], one can even consider an additional parameter of distance between clusters when the two data elements disagree. We shall be investigating this in future.

6. REFERENCES

[1] Ben-Hur, A., Elisseeff, A. and Guyon, A. (2002). A stability based method for discovering structure in clustered data. In Proc. Pacific Symposium on Biocomputing, pages 6–17.

[2] Denoeud, L., Garreta, H., Guénoche, A., (2005). Comparison of distance indices between partitions, in: Procs of the 11th Conf of the Applied Stochastic Models and Data Analysis (ASMDA), Brest/France.

[3] Fern, X. Z. and Brodley, C. E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In Proc. 20th ICML, pa186–193, Washington DC.

[4] Fred, A.L.N. and Jain, A.K. (2003). Robust data clustering. In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, USA.

[5] Hubert, L. and Arabie, P. (1985). Comparing partitions**.** *Journal of Classification*, **2:**193-218.

[6] Kuhn, H.W. (1956). Variants on the Hungarian method for the assignment problem. Naval Res. Logistic. Quart., pages 253–258.

[7] Kuhn, H.W. (1955). The Hungarian method for the assignment problem. Naval Res. Logistic. Quart., pages 83–97.

[8] Kuncheva, L.I and Hadjitodorov, S.T. (2004). Using diversity in cluster ensembles. in: Proceedings of IEEE Int. Conf. on Systems, Man and Cybernetics, The Hague, The Netherlands, pp. 1214–1219

[9] Li, J., Wang, J. Z., Wiederhold, G. (2000) IRM: Integrated region matching for image retrieval. ACM Multimedia, pages 147-156.

[10] Milligan, G.W, Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21:441-458.

[11] Pinto F.R, Carrico, J.A., M Ramirez1 and J S Almeida, (2007). Ranked adjusted Rand: integrating distance and partition information in a measure of clustering agreement. *BMC Bioinformatics*, 8:44.

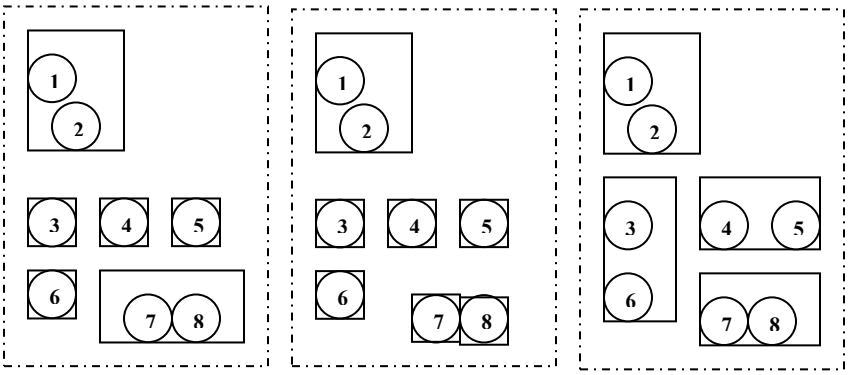
[12] Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association* **66**, pp. 846–850.

[13] Steinley, D. (2004). Properties of the Hubert-Arabic adjusted Rand index**.** *Psychol Methods*, **9:**386-396.

[14] Strehl, A and Ghosh, J. (2002). Cluster ensembles. A knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3:583–618.

[15] Wallace, DL. (1983). Comment**.** Journal of the American Statistical Association, **78:**569-576.

Figure 1(below): Partition P_1 has 6 clusters which can be identified by rectangle boxes. Similarly P_2 has 7 and P_3 has 4 clusters. It can be noticed clearly from the picture that P_1 is similar to P_2 than to P_3.



Dataset	N	k	k_1	IRM	Rand	Jaccard	AR	Wallace	NMI
Australian	690	2	2	**1.337249**	10.125183	1.597784	4.358551	2.044201	4.027432
German	1000	2	15	**1.048532**	4.645780	1.225258	2.174214	1.090978	1.565228
Glass	214	7	7	**1.215728**	15.853014	1.272213	3.865350	1.513213	2.288114
Heart	270	2	11	**0.967244**	7.024978	1.297638	2.455763	0.972834	1.803880
Iris	150	3	33	**0.452911**	10.527431	2.835106	0.493838	2.767596	5.541363
Lenses	24	3	4	**2.382998**	7.329707	5.475070	7.945702	2.410549	3.225222
Liver	345	2	14	**0.880220**	6.211665	1.128411	2.150390	0.933740	1.504367
Spect	267	2	2	**1.896824**	5.477993	2.144411	5.882934	2.235649	4.993913
Thyroid	215	3	29	**0.086105**	0.723668	0.144460	0.212707	0.116261	0.140647
Wine	178	3	26	**0.177859**	1.533105	0.246755	0.282415	0.182751	0.431561

Table 2 : Aggregated (Accuracy – Similarity) values from K_1 to 50 partitions.