A Weighted Consensus Function Based on Information-theoretic Principles to Combine "Soft" Clusterings

Yan Gao[1], Shiwen Gu[1], Jianhua Li[1], Zhining Liao[2]
[1]*School of Information Science and Engineering, Central South University University*
410083, Hunan, P.R.China China
{gaoyan,swgu,jhli}@csu.edu.cn
[2]*Department of Computer Science, Loughborough University,*
LE113TU, UK
liaozn@yahoo.com

Abstract

How to combine multiple clusterings into a single clustering solution of better quality is a critical problem in cluster ensemble. In this paper, we extend strehl's consensus function based on information-theoretic principles and propose a novel weighted consensus function to combine multiple "soft" clusterings. In our consensus function, we use mutual information to measure the sharing information between two "soft" clusterings and emphasize the clustering which is much different from the others. We use the algorithm similar to sequential k-means to obtain the solution of this consensus function and conduct experiments on four real-world datasets to compare our algorithm with other four consensus function, including CSPA, HGPA, MCLA, QMI. The results indicate that our consensus function provides solutions of better quality than CSPA, HGPA, MCLA, QMI and when the distribution of diversity in cluster ensembles is uneven, considering the influence of diversity can improve the quality of clustering ensemble.

1. Introduction

Combining the results of several clustering methods has recently appeared as one of the branches of multiple classifier systems. The aim of combining clustering results is to improve the quality and robustness of the results. The current clustering ensemble literature has mainly focused on two main problems: how to combine multiple clusterings (commonly referred to as the problem of consensus function); how to produce multiple clustering results and which diversity they have.

There are several efficient consensus functions derived from statistical, graph-based and information-theoretic principles etc. Fred & Jain[1], Fern & Brodley[2], Monti et al [3] established the co-association matrix based on similarities between different clustering solutions, and then use agglomerative hierarchical clustering. Topchy et al. [4][5] proposed a mixture model in order to obtain a consensus function. They also established a quadratic mutual information criterion for clustering ensemble and the approximate results for this criterion can be obtained by running k-means[5]. W. Tang and Z.H.Zhou[6] proposed bagging-based selective cluster ensemble algorithm in which the normalized mutual information between the clustering result and other results can be regarded as the weight of clustering result in bagging. Frossyniotis[7] applied boosting to clustering ensemble. Strehl & Ghosh[8] proposed three different approaches to generat consensus functions which are all based on hypergraph partitioning. They also point out that clustering ensemble can be regarded as the optimal problem based on mutual information, but not point out how to solve it and it is only to combine "hard" clustering.

Diversity plays the same significant role in cluster ensembles as classifier ensembles. Fern and Brodley [9] noted that more diverse ensembles offered larger improvement on the individual accuracy than less diverse ensembles. Kuncheva[10] also studied the diversity within such cluster ensembles and proposed a variant of the generic pairwise cluster ensemble approach which enforces diversity in the ensemble.

In this paper, we extend Strehl's consensus function based on information-theoretic principles and propose a weighted consensus function to combine "soft" and "hard" clusterings. In our consensus function, we use mutual information to measure the sharing information

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.156

between two “soft” clusterings and emphasize the clustering which is much different from the others.

The paper is organized as follows. Section 2 discusses related work: Strehl’s objective function for clustering ensemble and the measurement for diversity in cluster ensembles. Section 3 discusses a weighted consensus function based on information-theoretical principles to combine “soft” clusterings and the algorithm to implement consensus function. We conduct experiments in Section 4 and conclude with Section 5.

2.Related work

2.1 Consensus Function Based on Mutual Information

Strehl and Ghosh thought that combined clustering should share the most information with the original clusterings: Π={$C_{1,}$.... C_r}. But how do we measure shared information between clusterings? Strehl and Ghosh used normalized mutual information to measure it.

Suppose there are two clusterings: c_a and c_b. Let n_h is the number of instances which label is c_l in clustering C_b, n_l is the number of instance which label is c_h in clustering C_a, n_l^h is the number of instance which label not only is c_l in clustering C_b but also is c_h in clustering C_a. n is the total number of instances in data set. The normalize mutual information between C_a and C_b is:

$$\phi^{NMI}(C_a, C_b) = \frac{2}{n}\sum_{k=1}^{K}\sum_{k=1}^{K} n_l^h \log_{k^2}\left(\frac{n_l^h n}{n^h n^l}\right) \quad (1)$$

Strehl’s objective function is defined as:

$$C^{opt} = \arg\max_{C} \sum_{q=1}^{r} \phi^{NMI}(C, C_q) \quad (2)$$

Although Strehl&Ghosh proposed the objective function based on mutual information for clustering ensemble, they pointed out that for finite populations, the trivial solution is to exhaustively search through all possible clusterings with k labels (approximately kn/k! for n>> k) for the one with the maximum ANMI which is computationally prohibitive. And this objective function is only applied to combine “hard” clusterings.

A.Topchy[5] also established the objective function based on mutual information similar to Strehl&Ghosh’s. But they used quadratic mutual information. They proved that the approximate solution for their consensus function could be obtained by using k-means on the new feature space that is built on the original clusterings. And in Strehl’s objective function, the normalized mutual information is used to measure shared information between two “hard” clusterings, So their algorithm is also only applied to combine “hard” clusterings.

2.2 Diversity Between a Pair of Clusterings

Diversity within an ensemble is of vital importance for its success. There are many methods to measure the diversity between a pair of clusterings: (1) pair counting[11][12][13][14], (2) set matching [15] and (3) variation of information (VI)[16][17]. The pair counting method evaluates the similarity between two clustering algorithms by examining how likely they are to group a pair of objects together, or separate them in different clusters. All pair counting methods are restricted to handling hard clusterings. Other drawbacks of pair counting methods are also discussed in [12]. The set matching method seeks for a match between clusters, that is, the sets of objects grouped together in two clusterings respectively. Existing set matching approaches perform matching in a stepwise manner without a global optimization objective.

VI is the criterion based on information-theoretic principles for comparing two clusterings on the same dataset. The criterion makes no assumptions about how the clusterings were generated and can be applied to compare both soft and hard clusterings. Supposing C and C’ are two clusterings, the value of VI between clustering C and C’ is computed as follows.

$$\begin{aligned} VI(C, C') &= H(C) + H(C') - 2I(C, C') \\ &= H(C \mid C') + H(C' \mid C) \end{aligned} \quad (3)$$

Where H(C|C’) ,H(C’|C) are both the conditional entropies. H(C|C’) measures the amount of information about C that we loose, while H(C’|C) measures the amount of information about C’ that we still have to gain, when going from clustering C to clustering C’. If C=C’, VI (C, C’) =0. If VI (C, C’)>VI(C,C’’), the C is more different form C’ than C’’.

Before computing VI of two clusterings, we should know the joint distribution of two clusterings. If C and C’ are both hard clusterings, c is the cluster belonging to C, c’ is the cluster belonging to C’, n is the number of instance on the dataset X, then the joint distribution p(c, c’) used in VI is computed as follows:

$$p(c', c) = \frac{|c \cap c'|}{n} \quad (4)$$

If C and C’ are both soft clusterings, c is the cluster belonging to C, c’ is the cluster belonging to C’, supposing the knowing cluster c is irrelevant to c’ if the chosen instance x is known, i.e. p(c'|x,c)=p(c'|x), the joint distribution P(c, c') is computed as follows:

$$p(c',c)=\sum_{x} p(c'|x)p(c|x)p(x) \qquad (5)$$

3. The weighted consensus function

3.1 The weighted objective function based on mutual information

We also think that cluster ensemble should extract the combined clustering sharing the most information with the original clusterings. But in this paper, considering "soft" clusterings and the influence of diversity in cluster ensembles, we propose a weighted objective function for cluster ensemble.

First, instead of normalized mutual information, we use the Shannon mutual information to measure shared information between two "soft" clusterings.

Suppose there are a set of clusterings on dataset X: Π={C1,…Cr}, where C_q is the "soft" clustering, C_q={c_q^1,…c_q^k }, C is the combined clustering. Supposing C_q be any clustering in Π, the information of the cluster c_q^j in the clustering C_q should be transferred to cluster c_i in clustering C (c_q^j ->x-> c_i). So the joint probability p(c_q^j, c_i) can be assumed as follows:

$$p(c',c)=\sum_{x} p(c'|x)p(c|x)p(x) \qquad (6)$$

The marginal probability p(c_q^j) is :

$$p(c_q^j)=\sum\nolimits_{x} p(c_q^j|x)p(x) \qquad (7)$$

The marginal probability p(c_i) is:

$$p(c_i)=\sum\nolimits_{x} p(c_i|x)p(x) \qquad (8)$$

Consequently the Shannon mutual information between "soft" clustering C_q and the combined clustering C can be computed as follows:

$$I(C_q;C)=\sum_{i=1}^{k}\sum_{j=1}^{k} p(c_q^j,c_i)\log\frac{p(c_q^j,c_i)}{p(c_q^j)p(c_i)} \qquad (9)$$

Second, we consider the influence of diversity within an ensemble in our objective function. We use VI to measure the difference between two "soft" or "hard" clusterings. The diversity of every clustering is defined as the average of VI between itself and other clusterings. So the diversity can be computed as follows:

$$div\,(\mathrm{i})=\frac{1}{(\mathrm{r}-1)}\sum_{\mathrm{j=1..r,\ j\neq i}}\mathrm{VI}(C_{\mathrm{i}},C_{\mathrm{j}}) \qquad (10)$$

The clustering with high value of diversity is more different from other clusterings. It contains much information which other clusterings don't have. We think this clustering is important in clustering ensemble.

Considering "soft" clusterings and the influence of diversity in cluster ensembles, the weighted objective function to combine "soft" clusterings can be defined as:

$$\mathrm{C}^{opt}=\arg\max_{\mathrm{C}}\sum_{q=1}^{r}\mathrm{div}\,(q)I(\mathrm{C},\mathrm{C}_q) \qquad (11)$$

3.2 Algorithm implementation

In this section we propose an algorithm to solve the weighted objective function mentioned in section 3.1.

First we define F=$\sum_{q=1..r}$ div(q)I(C_q; C). F is decomposable, i.e. if the combined clustering is C={c_1,……, c_k}, then F=$\sum_i$ F(c_i), F(c_i)=$\sum_{q=1..r}$ div(q) I(c_i; C_q). So when the instance x is merged into the cluster c_i, the value of F is decreased and the change is computed as follows:

$$d_F(c_i,x)=F(\{c_i\})+F(\{x\})-F(\{c_i,x\}) \qquad (12)$$

$$=\sum_{q=1}^{r}div(q)I(c_i;\mathrm{C}_q)+\sum_{q=1}^{r}div(q)I(x;\mathrm{C}_q)-\sum_{q=1}^{r}div(q)I(\{c_i,x\};\mathrm{C}_q)$$

$$=\sum_{q=1}^{r}div(q)\sum_{j=1}^{k}(p(x)+p(c_i))JS(p(c_q^j|x),p(c_q^j|c_i))\geq 0$$

Where C_q is "soft" clustering, c_q^j is the cluster in C_q. JS is the Jensen-Shannon divergence defined as:

$$JS_{\Pi}[p(c_q^j|x),p(c_q^j|c_i)]=\pi_1 D_{KL}[p(c_q^j|x)\|\bar{p}]+\pi_2 D_{KL}[p(c_q^j|c_i)\|\bar{p}] \qquad (13)$$

$$\{\pi_1,\pi_2\}=\{\frac{p(x)}{p(x)+p(c_i)},\frac{p(c_i)}{p(x)+p(c_i)}\},\ \bar{p}=\pi_1 p(c_q^j|x)+\pi_2 p(c_q^j|c_i)$$

If x is removed from its current cluster c_j to c_i, the change of F is computed as follows:

$$\Delta F=F(\{c_i,x\})-F(\{c_j\}) \qquad (14)$$
$$=d_F(c_j',x)-d_F(c_i,x)$$

Where c_j' is the cluster which is formed by removing x from its current cluster c_j.

According to (14), we know that the value of F is increased if x is merged into the cluster c which satisfies the following codition:

$$c=\arg\min_{c_i} d_F(c_i,x) \qquad (15)$$

So here we can use the algorithm like sequential k-means to solve our objective function. We start from the initial random clustering of dataset X. For every instance x∈ X, we remove x from its current cluster c_j, and merge x into the cluster c which has the minimum of d_F (c, x). We repeat the process until no instance change its cluster.

When x is removed from its current c_i, we should update $p(c_i)$ and $p(c_q^j|c_i)$ ($l \leq j \leq k$). When x is merged into c, we should update $p(c_j)$ and $p(c_q^l|c)$. The update equation is defined as:

$$p(c_q^j \mid c) = \frac{1}{p(c)} \sum_{l=1}^{|c|} p(x_l, c_q^j) \qquad \forall c_q^j \in C_q \tag{16}$$

$$p(c) = \sum_{l=1}^{|c|} p(x_l)$$

4. Experiments

We experiment on four datasets from UCI[18]. The attributes in four datasets are numerical. The details of four datasets are described in Table I.

TABLE 1 The details of four datasets

Data Set	Num. of Features	Num. of Classes	Num. of Instances	Fuzzy K-means (Mean Error Rate)
Wine	13	3	178	0.314
Glass	9	6	241	0.509
Ionos-phere	34	2	351	0.291
Spam -base	57	2	4601	0.357

In order to produce diverse clusterings, we use random subspace method[19][20], where each base clustering is generated on a randomly selected subset of the original dimensions. Fuzzy k-means is used on new subspace to produce a "soft" clustering. The dimension of subspace is $\lceil d/4 \rceil$. The maximum iterative time in fuzzy k-means is 100.

In experiments, we define f-MI when our algorithm is used to combine "soft" clusterings without considering the influence of diversity. We define w-MI when our algorithm is used to combine "soft" clusterings with the influence of diversity. We define h-MI when our algorithm is used to combine "hard" clusterings without considering the influence of diversity. MCLA, HGPA, CSPA are three ensemble algorithm proposed by Strehl. They are all based on hypergraph partitioning. QMI proposed by Topchy is based on quadratic mutual information criterion and use k-means to obtain the approximately combined clustering. We compare w-MI, f-MI, h-MI with MCLA, HGPA, CSPA and QMI. The CSPA, MCLA, HGPA code is available in [21].

Because MCLA, HGPA, CSPA and QMI are used to combine "hard" clusterings, they cannot directly combine "soft" clustering produced by fuzzy k-means. Hard clustering used by MCLA, HGPA, CSPA and QMI can be produced by assigning every instance to the cluster in which its conditional probability is at maximum.

Our algorithm is susceptible to the presence of local minima of the objective functions. To reduce the risk of convergence to a lower quality solution, the final clustering was picked from the results of three runs (with random initializations) according to the value of objective function. The highest value of objective function (10) served as a criterion for our algorithm.

We randomly choose five values [10, 15, 20, 30, 40] for the size of ensemble.

The mean error rate of 10 clustering ensemble procedures is used to measure the performance of clustering ensemble. Let C_{true} represent the true (given) clustering and C the ensemble clustering, "Confusion" be the confusion matrix of two clusterings: (Confusion(k_{true}; k) = (C_{ktrue}, C_k) i.e. number of points x that are cluster k_{true} in true clustering and cluster k in the clustering produced. Clustering error rate is defined as follows:

$$error\_rate = \frac{1}{n}(\sum_{k_{true}} \sum_{k \neq k_{true}} Confusion(k_{true}, k)) / n \tag{17}$$

Where n is the total number of objects. The low value of error _rate indicates good quality of clustering.

Figure 1,2,3,4 describe shows the mean error rate on four data sets.

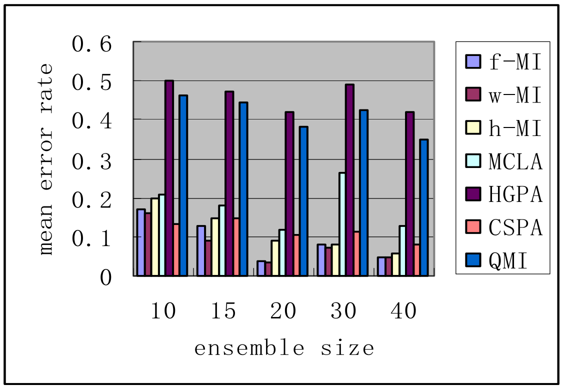

Fig. 1. The mean error rate on "wine"

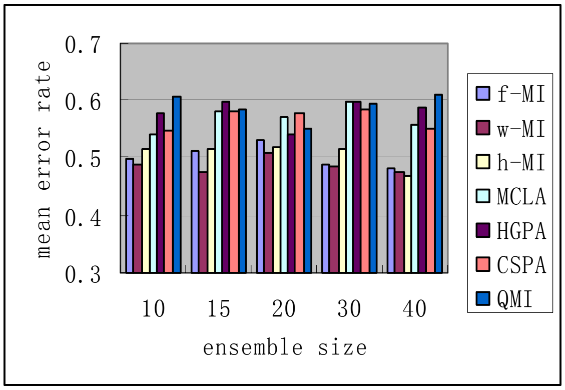

Fig.2. The mean error rate on "glass"

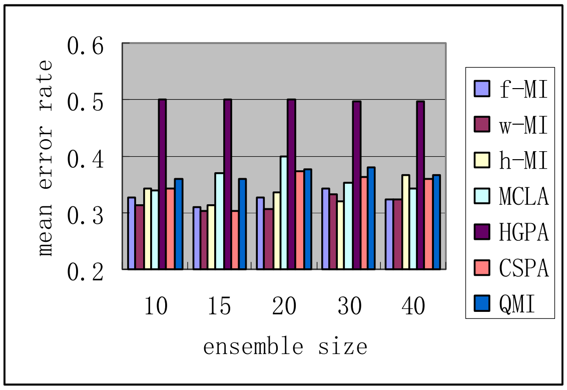

Fig. 3. The mean error rate on "spambase"

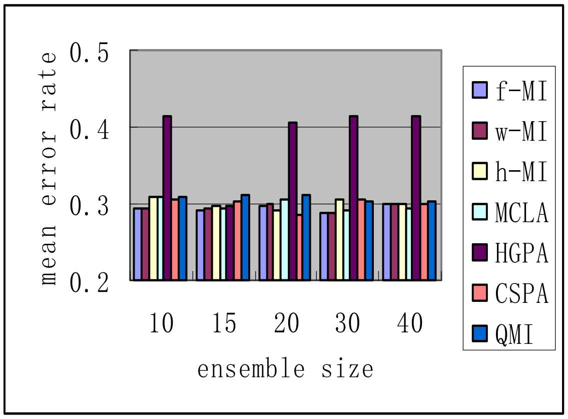

Fig. 4. The mean error rate on "ionosphere"

From Figure 1,2,3,4, we find that there is no algorithm which performance is best for all four datasets with different ensemble size. But on the whole, the performance of w-MI is best among seven algorithms. The second is f-MI. Especially when the size of ensemble is small, the mean error rate of f-MI and w-MI is lower than that of other algorithms, because original "soft" partitions contain much information than "hard" partitions. Although h-MI and QMI are both based on mutual information criterion for "hard" clustering ensemble, h-MI provides clusters with better quality than QMI.

Table 2. The distribution of diversity of clusterings on "glass".

Times No.	Mean	Standard Deviation	f-MI	w-MI
1	0.7433	0.0395	0.4766	0.4673
2	0.7450	0.0329	0.4346	0.4299
3	0.7827	0.0471	0.5280	0.5280
4	0.667	0.065	0.4766	0.4673
5	0.6789	0.0762	0.5421	0.5374
6	0.7288	0.0550	0.5234	0.5047
7	0.7659	0.0408	0.5467	0.5327
8	0.7450	0.0329	0.4346	0.4299
9	0.7241	0.040	0.4673	0.4555
10	0.7169	0.0516	0.4860	0.4776

The ensemble size is 10. The repeated times of the ensemble process is 10.

Table 3.The distribution of diversity of clusterings on "ionosphere".

Times No.	Mean	Standard Deviation	f-MI	w-MI
1	0.918	0.0243	0.2821	0.2821
2	0.9045	0.012	0.3048	0.3048
3	0.9216	0.0221	0.3077	0.3077
4	0.942	0.0174	0.2906	0.2908
5	0.9086	0.0256	0.2849	0.2849
6	0.9008	0.0151	0.2877	0.2877
7	0.9343	0.0112	0.2934	0.2934
8	0.9059	0.0126	0.2764	0.2764
9	0.9148	0.0163	0.2991	0.2991
10	0.9147	0.0121	0.3105	0.3105

The ensemble size is 10. The repeated times of the ensemble process is 10.

From figure 1, 2, 3, 4, we also find that the mean error of w-MI is lower than that of h-MI on most datasets except "ionosphere". So on most datasets except "ionosphere", considering the influence of diversity can improve the quality of clustering ensemble. Why does w-MI not improve the performance of clustering ensemble on "ionosphere"? We list the mean and standard deviation of diversity of clusterings on "glass" and "ionosphere" in Table 2 and Table 3. The large standard deviation of diversity means the distribution of diversity is uneven. Table 2 shows that the standard deviation of diversity on "glass" is large and the mean error rate of w-MI is lower than that of f-MI. Table 3. shows that the standard deviation of diversity on "ionosphere" is small but the mean error rate of w-MI is not lower than that of f-MI. So from Table 2 and Table 3, we know that when the distribution of diversity of clustering is uneven, the quality of clustering can be improved by considering the influence of diversity. The distribution of diversity on “ionosphere” is even, so w-MI can not improve the performance of clustering ensemble.

5 Conclusion

In this paper, we extend Strehl’s work and propose a weighted consensus function to combine “soft” clusterings. In this consensus function, we use mutual information to measure the sharing information between two “soft” clusterings and emphasize the clustering which is much different from the others. We use the algorithm similar to sequential k-means to obtain the solution of this consensus function. Experiments on four real data sets indicate that our consensus function is an effective way to combine “soft” clusterings and when the diversity of clusterings is uneven, the quality of clustering can be improved by considering the influence of diversity.

6. References

[1] A.L.N. Fred and A.K. Jain, “Data Clustering Using Evidence Accumulation”, *In Proc. of the 16th International Conference on Pattern Recognition, ICPR 2002*, Quebec City, pp.276– 280.

[2] Fern, X. Z., & Brodley, C. E. “Random projection for high dimensional data clustering: A cluster ensemble approach”. *ICML 2003*, Menlo Park, pp.186-193.

[3] Monti, S. Tamayo, P, Mesirov, J, & Golub, T. “Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data”. *Machine Learning* 2003, 52, pp.91–118.

[4] A. Topchy, A. Jain, and W. Punch. “A mixture model for clustering ensembles”. *In Proc. SIAM Data Mining*, 2004, pp.379-390.

[5] A. Topchy, A. Jain, and W. Punch. “Combining multiple weak clusterings”. *In Proc. Third IEEE International Conference on Data Mining (ICDM'03)*, November 2003.

[6] Wei Tang, ZhiHua Zhou, “Bagging-Based Selective Clusterer Ensemble”. *Journal of software*, 2005, Vol.16, No.4, pp.496-502.

[7] D. Frossyniotis, A. Likas, A. Stafylopatis, “A clustering method based on boosting”, *Pattern Recognition Letters* 25 (2004), 641–654.

[8] A. Strehl and J. Ghosh. “Cluster ensembles - a knowledge reuse framework for combining partitionings”, *In Proc.Conference on Artificial Intelligence (AAAI 2002)*, Edmonton, pp.93–98..

[9] A. Fred and A.K. Jain. “Data clustering using evidence accumulation”. *In Proc. 16th International Conference on Pattern Recognition, ICPR*, , Canada, 2002, pp. 276-280.

[10] L.I. Kuncheva, S.T. Hadjitodorov. “Using Diversity in Cluster Ensem- bles”. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, 2004, pp.1214-1219.

[11] A. Ben-Hur, A. Elisseeff and I. Guyon, “A stability based method for discovering structure in clustered data”, In Pacific Symposium on Biocomputing, 2002, pp.6-17.

[12] E.B.Fowlkes and C.L.Mallows, “A method for comparing two hierarchical clusterings”, *Journal of the American Statistical Association*, 1983, 78(383), pp.553-569.

[13] L. Hubert and P. Arabie, “Comparing clusterings”, *Journal of Classification*, 1985, 2, pp.193-218.

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

[15] S. Dongen, “Performance criteria for graph clustering and Markov cluster experiments”, *Technical Report*, Amsterdam, 2000.

[16] M. Meila. “Comparing clusterings by the variation of information”. *In Conference on Learning Theory(COLT)*, 2003, pp.173-187.

[17] Marina Meila. “Comparing clusterings: an axiomatic view”. *In International Conference of Machine Learning (ICML)*, 2005: 577-584.

[18] Blake C, Keogh E, Merz CJ. UCI Repository of machine learning databases. Irvine: Department of Information and Computer Science, University of California, [Online]. Available: http://www.ics. uci.edu /~mlearn/MLRepository.html.

[19] D. Greene, A. Tsymbal, N. Bolshakova, P. Cunningham, “Ensemble clustering in medical diagnostics”, *In Proc. 17th IEEE Symp. on Computer-Based Medical Systems CBMS 2004, Bethesda, MD, National Library of Medicine/National Institutes of Health*, IEEE CS Press, 2004, pp.576–581.

[20] T. K. Ho. “The random subspace method for constructing decision forests”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(8), pp.832–844.

[21] CSPA, MCLA, HGPA code[Online]. Available: http://strehl.com/75-280, 2001.