A NEOTERIC APPROACH FOR CLUSTERING DNA AMOEBEAN BASED ON JIBE INDEX

Mrs. S.Sarumathi
Professor,
Department of B.Tech IT
K.S.Rangasamy College of Technology ,
Tiruchengode -637215,
rishi_saru20@rediffmail.com

Mr.S.Sakthivel
Assistant Professor,
Department of B.Tech IT
K.S.Rangasamy College of Technology,
Tiruchengode -637215,
mcssakthivel@gmail.com

Mrs.G.Malathi
Assistant Professor,
Department of B.E. CSE
Vellalar College of Engineering and Technology, Erode
malathi_gurunathan@rediffmail.com

ABSTRACT

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects. In cluster analysis, the course of history challenges the process of estimating the true number of clusters in a certain domain is high importance in a data set. For example, in medical research detecting the true number of groups and subgroups of cancer would be of at most importance for their effective treatment. In this paper we propose a neoteric method on jibe clustering to estimate the number of clusters in a gene chip data. It provides valuable information about the appropriate number of clusters which is robust and high quality cluster structure in a data set. When the specified number of clusters coincides with the true number of clusters it tends to be less diverse. To quantify this diversity we develop a neoteric index, namely the Jibe Index (JI), which is built upon a suitable clustering technique known as CLANA clustering. Our experiments on gene chip data sets indicate that the JI is used to find the amoebean number of cluster using sequence clustering of an object that specifies how much the object truly belongs to the cluster.

Keywords **–** Clustering, CLANA Clustering, Jibi Index

1. INTRODUCTION

Determining the correct number of clusters present in a data set is a key problem in cluster analysis and has attracted considerable research attention. Back in the early days of cluster analysis, when hierarchical clustering was still the dominant clustering technique, Milligan and Cooper, in their extensive research [6], investigated up to 30 procedures for determining the number of clusters in a data set. Those procedures should be exercised with care, in conjunction with close examination of the clustering task. In this research, we are concerned with methods for determining DNA amoebean clustering using CLANA sequence approach. In an era where a huge number of clustering algorithm exits, but we implement a new clustering idea known as CLANA sequence clustering. It is not just another clustering algorithm: it rather provides a framework for unifying the knowledge obtained from the other algorithms. Jibe clustering employs the clustering algorithm(s) to generate a set of clustering solutions on either the original data set or its perturbed versions. From those clustering solutions, jibe clustering aims at choosing a robust and high quality representative clustering. Jibe clustering is particularly useful in the context of gene chip data clustering, since the unified clustering solution greatly facilitates biological interpretation. In this paper, we focus on another aspect of CLANA sequence clustering namely its potential for determining the appropriate number of clusters and to develop an index to realize sub potential. Although our approach, together with jibe clustering, can be considered as a general framework and can be applied to any type of data and in conjunction with any clustering algorithm, in the light of our analysis above, we stress that the components of this framework must be chosen carefully to fit the clustering task at hand. We implemented and tested the whole framework in the content of gene chip data analysis. The paper is organized as follows. In section II we review several works on cluster number estimation which have been successfully applied in micro array data cluster analysis. Section III details our approach. Some experimental results are given in section IV followed finally by some discussion and conclusions.

2. RELATED WORK

In this section, we review some previous approaches for clustering DNA amoebean, based on Jibe Index, that have been successfully applied to gene expression cluster analysis. To set place for our subsequent discussions, we first give some backgrounds and notations for jibe Clustering. Categorized the methods for generating multiple clustering in Consensus Clustering into five types: (i) using different algorithms, (ii) performing

multiple runs of a single algorithm, (iii) sub sampling, Re-sampling or adding noise to the original data, (iv) using selected subsets of features, (v) using different K values to generate different clustering solutions where K is the number of clusters. Though, in our opinion, the latter is used only when one is concerned with determining the appropriate number of clusters. Given a data set of N data points and a pre-specified value of K, using a single method or a combination of those methods, i.e. (i)-(iv), a set of B clustering solutions can be obtained. Associated with the u-th clustering solution is a connectivity matrix (or adjacency matrix) MuK of size N _ N where MuK (i; j) = 1 if the two points i and j are grouped into the same cluster and 0 otherwise. If a sub-sampling strategy were employed, then for each clustering solution there is also an associated indicator matrix Iu K of size N _ N, where Iu K (i; j) = 1 if the two points i and j are both chosen in the u-th sub-sample and 0 otherwise.

The aggregated knowledge about the clustering solutions corresponding to each value of K can be conveniently summarized in the so-called consensus matrixMK of which the entry Mu K(i; j) tells us how frequently the two data points i and j have been grouped in the same cluster. MK is determined by:

$$MK = \frac{1}{B}\sum_{u=1}^{B} MuK$$

When a sub-sampling strategy is employed, MK is determined by:

$$MK(i,j) = \frac{\sum_{u=1}^{B} MuK(i,j)}{\sum_{u=1}^{B} IuK(i,j)}$$

To determine the appropriate value for K, a set of clustering solutions for each value of K ranging from 2 to Kmax are generated. Using the sub-sampling approach in conjunction with the Hierarchical Clustering and Self Organizing Map algorithms, Monti et al. (2002) [7] proposed a procedure for determining the value for K based on observing the change in the area under the empirical cumulative distribution of the values in the consensus matrix when K changes. For a given histogram an empirical cumulative distribution (CDF) can be calculated as:

$$\sum MK(i,j) \le c$$

$$CDF(c) = \frac{i < j}{N(N-1)/2}$$

then the area under the CDF can be computed as:

$$A(K) = \sum_{i=1}^{m} [x_i - x_{i-1}] \quad CDF(x_i)$$

where {x1, x2,…., xm} is the set of sorted entries in the consensus matrix MK. Finally the relative increase in the CDF area as K increases is computed as:

$$\Delta(K) = \begin{cases} A(K) & \text{if } K = 2; \\ \dfrac{A(K+1) - A(K)}{A(K)} & \text{if } K > 2 \end{cases}$$

They notice that as K is increased the area under the CDF markedly increases as long as K is less than the true value Ktrue. However when Ktrue is reached any further increase in K does not lead to a corresponding marked increase in the CDF area. Based on this observation a rule for determining the value of K is built upon inspection of the CDFs and the 4(K)-vs-K graph. Since there is no area under the CDF for K = 1, an irregular value is assigned to 4(2) and the group suggest that inspection of the CDF will be needed to choose between 1 and 2 clusters. The method has been applied on 6 synthetic and 6 real micro array data sets with promising results. However the process of calculating the 4(K) is rather cumbersome and by looking at this statistic alone it is hard to extract any intuition about its meaning. Recently Yu et al. (2007) [14] have presented another consensus based approach for determining the number of clusters in micro array data. Their approach can be summarized as follows: given a set of N data points in a d-dimensional space (or d features) and the number of clusters K, using random subspace generation (randomly choosing 75% to 85% of the original features set) and a graph based clustering algorithm, they first generate a set of B clustering solutions with B corresponding adjacency matrices (M1 K;M2 K ; : : : ;MB K). By varying the number of clusters from 2 to Kmax. The aggregated consensus matrix R is defined by pooling the entire obtained consensus matrix together as:

$$R = \frac{\sum_{K=2}^{K\,max} MK}{B(K\,max - 1)}$$

Yu et al. further binarize the aggregated consensus matrix R to Rb as follows:

$$R(i,j) = \begin{cases} 1 & \text{if } R(i,j) >= 0.5, \\ 0 & \text{if } R(I,j) < 0.5 \end{cases}$$

By the same way, the consensus matrices MK are binarized to Mb K. The author commented that this index balances the degree of agreement between the two matrices Mb K and Rb against the term 1=K2, which penalizes a large set of clusters. This criterion has been applied on several synthetic and real micro array data sets and to successfully discover the true number of clusters. Nevertheless the method for determining the optimal value of K is heuristic without a strong supportive theoretical background or clear motivation. More explanation and justification need to be given to many points in the whole process as to why the consensus matrices need to be binarized, and why the penalty term takes the form 1=K2. In addition, the computation of the Modified Rand Index, and hence K_, implicitly involves Kmax, a weakly relevant parameter. Generally, Kmax indicates the range of K one would like to explore and should not appear directly in the computation of K_, although it might affect the result if Kmax is set to a lower value than Ktrue. Finally, for this criterion, no guideline was provided to distinguish between the case of 1 (no cluster structure) and 2-or-more clusters.

3. JIBE INDEX

In this paper we introduce a new framework for estimating the number of clusters based on the CLANA sequence clustering paradigm. We aim for clarity of motivation for the framework, that is, a criterion directly derived from an intuition concerning cluster agreement. We start by repeating the main idea of CLANA sequence clustering: it is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume n clusters) fixed a priori. The main idea is to define n centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as possible for away from each other.

The next step is to take each point belonging to a given dataset and associate it to the nearest centriod. When no point is pending, the first step is completed and an early group age is done. At this point, we need to re-calculate n new centroids as bar centers of the clusters resulting from the previous step.

After we have these n new centroids,a new binding has to be done between the same data points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the n centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Addendum Index(AI) Table				
u1 u2 . ..uR	v1 v2 . ..VC	w1 w2 . ..WD	x1 x2 . ..XF	Sum1 Sum2 . ..sumi
R $\sum_{i=1}$Ui	C $\sum_{j=1}$Vj	D $\sum_{k=1}$Wk	E $\sum_{e=1}$Xl	N $\sum_{z=1}$Sz

Finally, this algorithm aims at minimizing an objective function, in this case the objective function.

$$\mathbf{J = \sum_{j=1}^{n} \sum_{i=1}^{x} \left\| Xi - Cj \right\|^2}$$

where $\left\| Xi^{(j)} - Cj \right\|^2$ is a chosen distance measure between a data point $Xi^{(j)}$ and the cluster center Cj, is an indicator of the distance of the n data points from their respective cluster centers.

The algorithm is composed of the following steps:

- Place n points into the space represented by the objects that are being clustered. These points represent initial group centeroids.
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned; recalculate the positions of the n centroids.
- Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the CLANA sequence clustering algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The

CLANA sequence clustering algorithm can be run multiple times to reduce their effect.

We next turn to the choice of a suitable agreement measure for Jibe Index (JI). Such a measure should be first effective at discriminarity the differences/similarities between clustering.

For this purpose, **the addendum index (AI),** a similarity index based on pairs counting, which is widely employed in the clustering literature, seems to be good choice. Along with this **the accretion index (ACI), the effacing index (EI) and the veering index (VI)** is similar to the Aitches index is based on the clustering of similar sequences in different sequences. We give a more detailed review of each index.**A.**

A.The Addendum Index (AI): Given a dataset of N data points s={s1,s2,…..,sN} and its 4 clustering subgroups namely U={U1,U2,….,UR}with R clusters V={V1,V2,…,VC}with c clusters W={W1,W2,…WD} with D clusters X={X1,X2,..,XE} with E clusters,

$$\{\cap_{i=1}^{R} Ui=\cap_{j=1}^{C} Vj=\cap_{k=1}^{D} Wk=\cap_{l=1}^{E} Xl=\varphi\}$$

$$\{\bigcup_{i=1}^{R} Ui=\bigcup_{j=1}^{C} Vj=\bigcup_{k=1}^{D} Wk=\bigcup_{l=1}^{E} Xl=S\}$$

Addendum index form is that

The addendum index (AI)

$$=\sum_{i=1}^{R} Ui+\sum_{j=1}^{C} Vj+\sum_{k=1}^{D} Wk+\sum_{l=1}^{E} Xl+\left(\sum_{Z=1}^{N} Sz\right),$$

B.The accretion Index (ACI):

Given a dataset of S data points N={N1,N2,….Ns} and its subgroups clustering are R={R1,R2,…RU} with U clusters,C={C1,C2,…CV}with V clusters, D={D1,D2,…DW}with W clusters.

$$\{\cap_{i=1}^{U} Ri=\cap_{j=1}^{C} Cj=\cap_{k=1}^{W} Dk=\varphi\},$$

$$\{\bigcup_{i=1}^{U} Ri=\bigcup_{j=1}^{V} Cj=\bigcup_{k=1}^{W} Dk=N\}$$

The accretion index (ACI) =

$$\sum_{i=1}^{U} Ri+\sum_{j=1}^{V} Cj+\sum_{k=1}^{W} Dk+\left(\sum_{Z=1}^{S} Nz\right),$$

The accretion Index (ACI)Table			
R1 R2 . .RU	C1 C2 . .CV	D1 D2 . .DW	Sum1 Sum2 . .Sumi
U $\sum_{i=1}^{U} Ri$	V $\sum_{j=1}^{V} Cj$	W $\sum_{k=1}^{W} Dk$	S $\sum_{z=1}^{S} Nz$

C.The effacing Index (EI):

Given a dataset of M data points o={o1,o2,…..,oN} and its 3 clustering subgroups namely U= {U1,U2,….,UR} with R clusters V = {V1,V2,…,VC} with C clusters W = {W1,W2,…WD} with D clusters.

$$\{\cap_{i=1}^{R} Ui=\cap_{j=1}^{C} Vj=\cap_{k=1}^{D} Wk=\varphi\},$$

$$\{\bigcup_{i=1}^{R} Ui=\bigcup_{j=1}^{C} Vj=\bigcup_{k=1}^{D} Wk=O\}$$

The effacing index (EI) form is that

The effacing Index (EI)

$$=\sum_{i=1}^{R} Ui+\sum_{j=1}^{C} Vj+\sum_{k=1}^{D} Wk+\left(\sum_{Z=1}^{N} Sz\right),$$

The effacing Index (ACI)Table			
U1 U2 . ..UR	V1 V2 . .VC	W1 W2 . ..WD	Sum1 Sum2 . ..Sumi
R $\sum_{i=1}^{R} Ui$	C $\sum_{j=1}^{C} Vj$	D $\sum_{k=1}^{D} Wk$	M $\sum_{z=1}^{M} Oz$

D.The Veering Index (VI):

Given a dataset of P data points q={q1,q2,…..,qN} and its clustering subgroups namely R= {R1,R2,….,RU} with U clusters, C={C1,C2,…,CV} with V clusters .

$$\{\cap_{i=1}^{U} Ri=\cap_{j=1}^{V} Cj=\varphi\},$$

$$\{\bigcup_{i=1}^{U} Ri=\bigcup_{j=1}^{V} Cj=q\}$$

The Veering index is form that
The Veering Index (VI)

$$= \sum_{i=1}^{U} Ri + \sum_{j=1}^{V} Cj + (\sum_{Z=1}^{P} qz),$$

The Veering Index (ACI)Table		
R1 R2 .. RU	C1 C2 ... CV	Sum1 Sum2 . Sumi
U ∑Ri i=1	V ∑Cj j=1	P ∑qz z=1

In jibe index we calculate by using CLANA sequence clustering algorithm.

Jibe index (JI) =

$$\sum_{z=1}^{N} Sz + \sum_{z=1}^{S} Nz + \sum_{z=1}^{M} Oz + \sum_{z=1}^{P} qz$$

4. EXPERIMENTAL RESULTS

Method:

In this section we present our experimental results of our approach in the context of gene chip data clustering. In the context of micro array data clustering, we shall consider the following:

1) *Clustering algorithm*: We use the CLANA sequence clustering algorithm, it is similar to the K-means but the main difference is that in k-means, if the dataset is not similar to centroids then those dataset is not taken into account but in CLANA sequence clustering algorithm even though it is dissimilar it will taken into account and added into the index and finally it will produce reasonable output.

2) ***Clustering generation*:** To estimate the number of clusters, we generate clustering's with varied number of clusters k ranging from 1 to kmax normally set to 20. it is noted that the specific value of k max generally does not affect the jibe index and hence the value of k*, unless kmax is set to a value possibly lower than Ktrue. For each value of k, we generate B=500 different clustering solution by employing the sub-sampling approach [7]. (i.e.) performing CLANA sequence on 500 different subsets of the original dataset. To avoid the situation that CLANA sequence is trapped into a bad local optimum, 10 different initializations are used for each subset and the clustering with the highest objective value is retained. Thus for each value of k, the total number of times k-means is run is 500.for comparison we also test the two algorithms by Yu etal (Graph consensus clustering equipped with k-means – GCCkmeans,or correlation graph clustering –GCCcorr).The number of clustering solutions was set to the default values for the two Yu's algorithm, (i.e.)B=500,while Kmax is also set to 20.Experimental results for the two consensus clustering algorithms by Monti et al.(CCHC and CCSOM), whenever available either from [7] or [14], are also reproduced for reference (marked with *).

A*) Jibe Index*:

The jibe index is built upon the Addendum Index(AI), Accretion Index(ACI),The effacing Index(EI),The Veering Index(VI).since a sub-sampling strategy is employed, a subset of the original dataset might contain data points that are not present in another. The AI, ACI, EI, VI are therefore, calculated based upon the DNA sequence.

B) *Data sets*:

1) Simulated Data sets: For ease of comparison, we test our algorithm on several simulated data sets that have been used in previous studies [7],[14].in particular we use six simulated data sets in [7]which are publicy available from the authors. Also 2 datasets[13] are generated using the description in the paper and the source code is provided by the authors. Detailed description of the simulated data sets might be found in the respective references. The summary of the simulated data sets are given in Table I.

Table I. Summary of the Real Microarray Datasets

Dataset	Source	#Classes	#Samples	Genes
Leukemia	[7]	3	38	999
Novartis multi-tissue	[7]	4	103	1000
St. Jude Leukemia	[7]	6	248	985
Lung cancer	[7]	4+	197	1000
CNS tumors	[7]	5	42	1000
Normal tissues	[7]	13	90	1277
Cho's yeast data 1	[13]	5	17	384
Cho's yeast data 2	[13]	4	17	237

2) Real genetic chisel data sets:

We evaluate our algorithm on both same clustering and gene clustering. Sample clustering is performed on three real gene chip dataset used in [7] with all the datasets. Gene clustering is performed on NSHL (Nodular sclerosing Hodgkin lymphoma) Mixed cellularity Hodgkin Lymphoma (MCHL), Lymphocyte depleted Hodgkin Hodgkin Lymphoma (LDHL), Lymphocyte-rich classic Hodgkin Lymphoma (LRCHL), and Nodular Lymphocyte.

Table II. Summary of the Real Genetic Chisel Datasets

Dataset	Source	#Classes	#Samples	Genes
Leukemia	[7]	4	48	1000
St. Jude Leukemia	[7]	6	108	898
Lymphoma	[7]	7	110	999
NSHL	13]	5+	72	482
MCHL	13]	3	63	362
LDHL	13]	2	57	637
LRCHL	[13]	8	82	584
NLPHL	[13]	9	77	772

C) *Experimental Results on simulated data sets*: We first test the algorithms on Monti et al. [7] Clustering algorithm at Yeung KY [13]. It can be observed from table V that these data sets do not present any serious challenge for all algorithms. All produced perfect results with only the CCHC algorithm misidentifying the true number of clusters in one case.

For Monti's simulated data sets, we first examine the multiple-clusters cases. On the St.Jude Leukemia and CNS tumors all the algorithms achieve either the correct result or a close estimation. It is noted however that on the Cho's yeast data set, the GCC kmeans algorithm procedures notably unstable results. For the Lung cancer and CNS tumors, our approach and the two algorithms by Monti et al., namely the CCHC and CCSOM, successfully reveal the true cluster structure. On the other hand, the two algorithms proposed by Yu et al., namely the GCC Kmeans and GCCcorr, wrongly determine the appropriate number of clusters.

In figure 1 the Jibe index varies in the range [0.2, 0.4] while for the multiple-clusters data sets, JI ranges in [0.55, 1]. From our experiments the α value range is [0.6, 0.7].The modified index has overall high values after from 0.6 to 0.8 which is as high as in the other multiple-cluster data sets.

D. *Experimental Results on Real Genetic Chisel Data Sets:*

Experimental results for the real genetic chisel data sets are presented in table VIII in Figure 4. We have taken data sets namely the lymphoma, MCHL and LDHL the JI-based criteria give close estimation. For the other cases, the index has a very high value at K = 2, where it has wrongly determined the true number of clusters. However for these cases a local peak can still be identified near K = Ktrue When K = Ktrue a substructure is identified, thus the CI index rises again and gives a local peak.

Figure 1. The jibe indices on some simulated data sets

a) Leukemia
b) St. Jude Leukemia
c)Nor vatis
d)Normal tissues

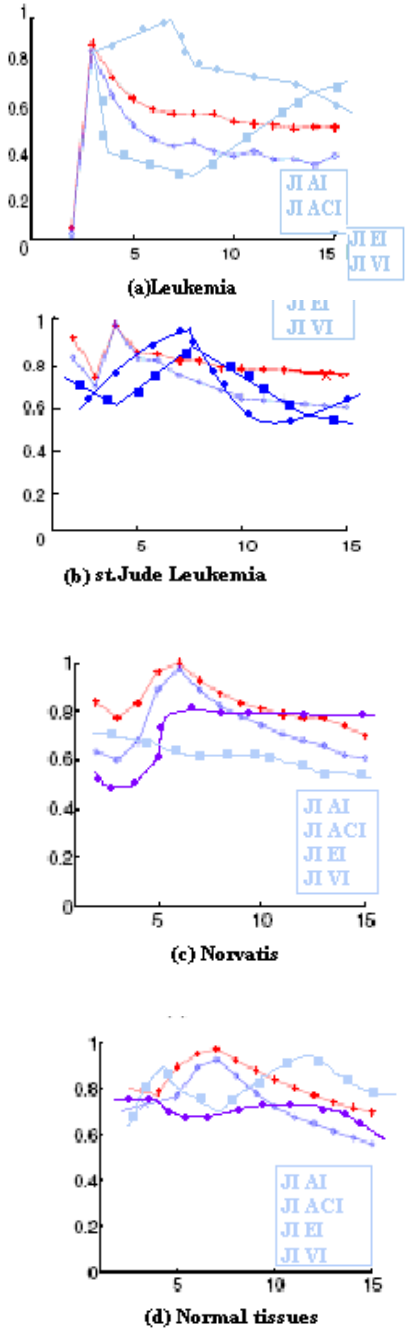
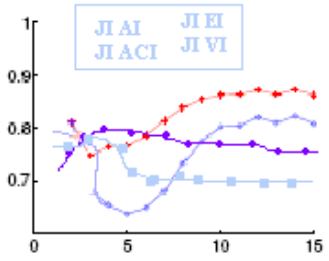
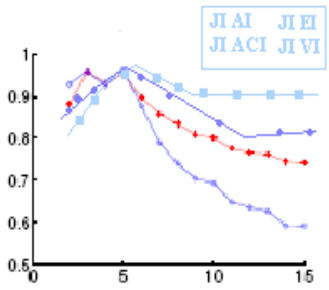


Figure 2. The jibe indices on some real data sets

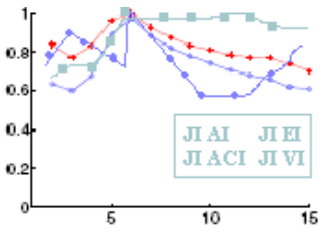
a)Leukemia
b)Lymphoma
c) MCHL
d)LDHL



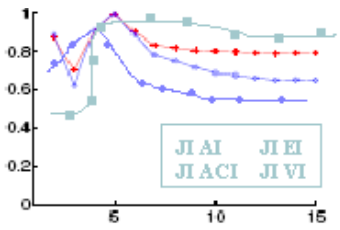
(a)Leukemia



(b) Lymphoma



(c) MCHL



(d) LDHL

5. CONCLUSION

In this paper we have presented a new framework for estimating the number of clusters in a dataset based on a neoteric index. The Jibe Index is built upon by the Addendum index (AI), the Accretion index (ACI), the Effacing index (EI) and the Veering index (VI).The Jibe index quantifies the agreement between the clustering solutions obtained by a CLANA sequence Clustering approach. The optimal value of number of clusters and their subgroups is chosen as the value that maximizes the jibe index, although the framework presented is general and can be theoretically applied to any type of data, we stress on the fact that all the components of the framework, e.g. the clustering algorithm, clustering generation and Jibe index. Extensive experiments on genetic chisel data clustering indicate the usefulness of the JI-based criterion for estimating the appropriate number of clusters. The JI measures namely the JIAI and JIACI, JIEI, JIVI tend to give quite concordant results, suggesting that AI, ACI, EI and VI are well suited for this purpose.

REFERENCES

[1] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, “Clustering on the unit hyper sphere using von mises-fisher distributions,” J. Mach. Learn. Res., vol. 6, pp. 1345–1382, 2005.

[2] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfs berg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis, “A genome-wide transcriptional analysis of the mitotic cell cycle,” Mol Cell, vol. 2, no. 1, pp. 65–73, 1998.

[3] L. Hubert and P. Arabie, “Comparing partitions,” Journal of Classification, pp. 193–218, 1985.

[4] H. Lancaster, “The chi-squared distribution.” New York: John Wiley, 1969.

[5] M. Meilˇa, “Comparing clustering’s: an axiomatic view,” in ICML ’05: Proceedings of the 22nd international conference on Machine learning. New York, NY, USA: ACM, 2005, pp. 577–584.

[6] G. Milligan and M. Cooper, “An examination of procedures for Determining the number of clusters in a data set,” Psychometrika, vol. 50, no. 2, pp. 159–179, June 1985.

[7] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, “Consensus clustering: A resampling-based method for class discovery and visualization of Gene expression microarray data,” Mach. Learn., vol. 52, no. 1-2, pp. 91–118, 2003.

[8] W. M. Rand, “Objective criteria for the evaluation of clustering Methods,” Journal of the American Statistical Association, vol. 66, No. 336, pp. 846–850, 1971.
[9] G. Schwarz, “Estimating the dimension of a model,” The Annals ofStatistics, vol. 6, no. 2, pp. 461–464, 1978.
[10] A. Strehl and J. Ghosh, “Cluster ensembles - knowledge reuse Framework for combining multiple partitions,” Journal of Machine Learning Research, vol. 3, pp. 583–617, 2002.
[11] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of Clusters in a dataset via the gap statistic,” Stanford University, Tech.
Rep., 2000.
[12] N. X. Vinh and J. Epps, “Information theoretic measures for clusteringComparison: Is a correction for chance necessary?” 2008-2009, in preparation.
[13] K. Y. Yeung, “Cluster analysis of gene expression data,” Ph.D.Dissertation, University of Washington, Seattle, WA, 2001.
[14] Z. Yu, H.-S. Wong, and H. Wang, “Graph-based consensus clustering for class discovery from gene expression data,” Bioinformatics, vol. 23, no. 21, pp. 2888–2896, 2007.
[15]Nguyen xuan vinh,Julien Epps,” A Novel Approach for Automatic Number of clusters Detection in micro array Data based on Consensus clustering”,2009.

BIOGRAPHY

Sarumathi.S received the B.E degree in Electronics and Communication Engineering from Madras University, in 1994 and the M.E degree in Computer Science and Engineering from Anna University, Chennai in 2007. She is working as a Professor in the Department of B.Tech IT at K.S.Rangasamy College of Technology, Tiruchengode. Her area of interests includes database systems, Data Warehousing and Mining, Computer Networks, Parallel databases and Multimedia systems. She is a member of ISTE.

Malathi.G received the B.E degree in Computer Science and Engineering from Bharathiyar University, in 1999 and the M.E degree in Computer Science and Engineering from Anna University, Chennai in 2007. She is working as an Assistant Professor in the Department of B.E. CSE at Vellalar College of Engineering and Technology, Erode. Her area of interests includes database systems, Data Warehousing and Mining, Computer Networks, Compiler Design and Grid Computing.

Sakthivel.S received the B.Sc degree in Computer Science from Periyar University, in 2002 and the M.Sc degree in Computer Science from Periyar University, in 2004. M.E degree in Computer Science and Engineering from Anna University of Technology, Coimbatore in 2009. He is working as an Assistant Professor in the Department of B.Tech (IT) at K.S.Rangasamy College of Technology, Tiruchengode. His area of interests includes database systems, Data Warehousing and Mining, Software Engineering.