2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering

A Novel Approach for Automatic Number of Clusters Detection in Microarray Data based on Consensus Clustering

Nguyen Xuan Vinh, *student member, IEEE* and Julien Epps, *member, IEEE*

Abstract— **Estimating the true number of clusters in a data set is one of the major challenges in cluster analysis. Yet in certain domains, knowing the true number of clusters is of high importance. For example, in medical research, detecting the true number of groups and sub-groups of cancer would be of utmost importance for their effective treatment. In this paper we propose a novel method to estimate the number of clusters in a microarray data set based on the consensus clustering approach. Although the main objective of consensus clustering is to discover a robust and high quality cluster structure in a data set, closer inspection of the set of clusterings obtained can often give valuable information about the appropriate number of clusters present. More specifically, the set of clusterings obtained when the specified number of clusters coincides with the true number of clusters tends to be less diverse. To quantify this diversity we develop a novel index, namely the Consensus Index (CI), which is built upon a suitable clustering similarity measure such as the well known Adjusted Rand Index (ARI) or our recently developed, information theoretic based index, namely the Adjusted Mutual Information (AMI). Our experiments on both synthetic and real microarray data sets indicate that the CI is a useful indicator for determining the appropriate number of clusters.**

I. 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 are often referred to as the stopping rules because they give guidance as to when to stop moving downward and to cut the dendrogram at that point. As more clustering techniques became available, more rules for determining the number of clusters were proposed; some are general while others are specifically designed for a certain clustering methods. For example, when model-based clustering is employed (*e.g.* Gaussian mixture model)a popular criterion is the Bayesian Information Criterion (BIC) [9]. BIC's strongest point is its solid theoretical support. However, as a great number of clustering methods are not model-based, its applicability is therefore limited. There are also more general criteria, such as the Gap statistic [11], which is applicable to virtually any clustering method.

Although a large number of criteria exist and have been shown to work well on a number of data sets, it is arguable that a universally satisfactory solution, if any exists, is still far from reach. Due to the ill-defined nature of the clustering problem, determining the "true" number of clusters is also an ill-defined problem in general. The process of determining an appropriate number of clusters should be considered closely with the clustering process itself, also bearing in mind the particular characteristics of the data set at hand. For example, for data sets with irregular shaped clusters, *e.g.* spiral or elongated, the application of clustering methods which rely on the assumption that clusters are of elliptical shape (such as Gaussian mixture model), will likely fail. Therefore, criteria built upon such clustering methods such as BIC will also likely fail in these instances. In addition, it can be seen that a large class of criteria based upon the homogeneity-separation intuition will also produce spurious results when applied to such an irregular shaped cluster scenario, since they only work best when the clusters are well separated and take on convex shapes. This analysis suggests that the application of any criteria should be exercised with care, in conjunction with close examination of the clustering task.

In this research, we are concerned with methods for determining the number of clusters using a Consensus Clustering approach. In an era where a huge number of clustering algorithms exist, the Consensus Clustering idea [7], [10], [14] has recently received increasing interest. Consensus Clustering is not just another clustering algorithm: it rather provides a framework for unifying the knowledge obtained from the other algorithms. Given a data set and a single or a set of clustering algorithms, Consensus 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, Consensus Clustering aims at choosing a robust and high quality representative clustering. Consensus Clustering is particularly useful in the context of microarray data clustering, since the unified clustering solution greatly facilitates biological interpretation. In this paper, we focus on another aspect of Consensus Clustering namely its potential for determining the appropriate number of clusters and develop an index to realize such potential. Although our approach, together with Consensus 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 context of microarray data analysis and compared our approach with several other previous works in this area.

The authors are with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia. {n.x.vinh,j.epps}@unsw.edu.au

978-0-7695-3656-9/09 $25.00 © 2009 IEEE
DOI 10.1109/BIBE.2009.19

The paper is organized as follows. In section II we review several works on cluster number estimation which have been successfully applied in microarray data cluster analysis. Section III details our approach. Some experimental results are given in section IV followed finally by some discussion and conclusions.

II. Related work

In this section, we review some previous approaches for automatic cluster number detection, based on the Consensus Clustering paradigm, 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 Consensus Clustering. Yu *et al.* [14] categorized the methods for generating multiple clusterings 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) M_K^u of size $N \times N$ where $M_K^u(i,j) = 1$ 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 I_K^u of size $N \times N$, where $I_K^u(i,j) = 1$ 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 matrix $\mathcal{M}_K$ of which the entry $\mathcal{M}_K^u(i,j)$ tells us how frequently the two data points i and j have been grouped in the same cluster. $\mathcal{M}_K$ is determined by:

$$\mathcal{M}_K = \frac{1}{B}\sum_{u=1}^{B} M_K^u \tag{1}$$

When a sub-sampling strategy is employed, $\mathcal{M}_K$ is determined by:

$$\mathcal{M}_K(i,j) = \frac{\sum_{u=1}^{B} M_K^u(i,j)}{\sum_{u=1}^{B} I_K^u(i,j)} \tag{2}$$

To determine the appropriate value for K, a set of clustering solutions for each value of K ranging from 2 to K_{max} 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:

$$CDF(c) = \frac{\sum_{i<j} \mathcal{M}_K(i,j) \leq c}{N(N-1)/2} \tag{3}$$

then the area under the CDF can be computed as:

$$A(K) = \sum_{i=1}^{m} [x_i - x_{i-1}] CDF(x_i) \tag{4}$$

where $\{x_1, x_2, \ldots, x_m\}$ is the set of sorted entries in the consensus matrix $\mathcal{M}_K$. Finally the relative increase in the CDF area as K increases is computed as:

$$\triangle(K) = \begin{cases} A(K) & \text{if } K = 2; \\ \frac{A(K+1)-A(K)}{A(K)} & \text{if } K > 2. \end{cases} \tag{5}$$

They notice that as K is increased the area under the CDF markedly increases as long as K is less than the true value K_{true}. However when K_{true} 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 $\triangle(K)$-vs-K graph. Since there is no area under the CDF for $K = 1$, an irregular value is assigned to $\triangle(2)$ 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 microarray data sets with promising results. However the process of calculating the $\triangle(K)$ 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 microarray 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 $(M_K^1, M_K^2, \ldots, M_K^B)$. By varying the number of clusters from 2 to K_{max} (pre-specified by the user) one obtains $K_{max} - 1$ consensus matrices $\{\mathcal{M}_2, \mathcal{M}_3, \ldots, \mathcal{M}_{K_{max}}\}$. The aggregated consensus matrix R is defined by pooling all the obtained consensus matrix together as:

$$R = \frac{\sum_{K=2}^{K_{max}} \mathcal{M}_K}{B(K_{max}-1)} = \frac{1}{B(K_{max}-1)} \sum_{K=2}^{K_{max}} \sum_{u=1}^{B} M_K^u \tag{6}$$

Yu *et al.* further binarize the aggregated consensus matrix R to R^b as follows:

$$R^b(i,j) = \begin{cases} 1 & \text{if } R(i,j) \geq 0.5, \\ 0 & \text{if } R(i,j) < 0.5, \end{cases} \tag{7}$$

By the same way, the consensus matrices $\mathcal{M}_K$ are binarized to $\mathcal{M}_K^b$. Finally the optimal number of clusters K^* is determined as the value of K that maximize the so-called Modified Rand Index [14]:

$$\zeta(\mathcal{M}_K^b, R^b) = \frac{\sum_{i<j} 1\{M_K^b(i,j) = R^b(i,j)\}}{N(N-1)} + \frac{1}{K^2} \tag{8}$$

where $1\{\ldots\}$ denotes the indicator function and

$$K^* = \underset{K\in\{2,\ldots,K_{max}\}}{\arg\max}\ \zeta(M_K^b, R^b) \tag{9}$$

The author commented that this index balances the degree of agreement between the two matrices $\mathcal{M}_K^b$ and R^b against the term $1/K^2$, which penalizes a large set of clusters.

This criterion has been applied on several synthetic and real microarray 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 $1/K^2$. In addition, the computation of the Modified Rand Index, and hence K^*, implicitly involves K_{max}, a weakly relevant parameter. Generally, K_{max} 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 K_{max} is set to a lower value than K_{true}. Finally, for this criterion, no guideline was provided to distinguish between the case of 1 (no cluster structure) and 2-or-more clusters.

III. The Consensus Index

In this paper we introduce a new framework for estimating the number of clusters based on the Consensus 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 Consensus Clustering: by generating a diverse set of clustering solutions, a sustainable, robust structure in the data set can be discovered. We further pose the hypothesis that with the correct number of clusters K_{true} imposed on the data, the discrepancy between the clusterings obtained by different algorithms or different runs of a single algorithm should be minimized, meaning that the cluster structure discovered is robust.

Given a value of K suppose we have generated a set of B clustering solutions $\mathcal{U}_K = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_B\}$, each with K clusters. We define the Consensus Index (CI) of $\mathcal{U}_K$ as:

$$CI(\mathcal{U}_K) = \sum_{i<j} AM(\mathbf{U}_i, \mathbf{U}_j) \tag{10}$$

where the agreement measure AM is a suitable clusterings similarity index. Thus, the Consensus Index CI quantifies the average agreement between all pairs of clustering solutions in a clustering set $\mathcal{U}_K$. The optimal number of cluster K^* is chosen as the one that maximizes CI:

$$K^* = \underset{K=2\ldots K_{max}}{\arg\max}\ CI(\mathcal{U}_K) \tag{11}$$

For the case of 1 cluster, *i.e.* no multi-cluster signature is present in the dataset, it can be predicted that the Consensus Index will be generally low across all the values of K since the cluster structure discovered should be unstable. Thus a rule for detecting $K^* = 1$ might be: choose $K^* = 1$ when $\max_{K=2\ldots K_{max}} CI(\mathcal{U}_K) < \alpha$ where α is a chosen threshold. We shall revisit this rule in the experimental results section, where more empirical evidence is available to build it up.

We next turn to the choice of a suitable agreement measure for CI. Such a measure should be first effective at discriminating the differences/similarities between clusterings. For this purpose, the Adjusted Rand Index [3], a similarity index based on pairs counting, which is widely employed in the clustering literature, seems to be a good choice. Also, information theoretic oriented similarity indices such as the Mutual Information [10] or the Variation of Information [5], are also attractive for their strong theoretical background. Second, a suitable measure should be well bounded with a comparable value range across all values of K. In this respect, the ARI is bounded in [0,1] which is good for our purpose. Information theoretic based measures on the other hand are either not bounded, or the baseline is not comparable with different K values. We give a more detailed review of some agreement measures along with the possible problems and remedies in the next subsections.

A. The Adjusted Rand Index (ARI)

Given a data set of N data points $S = \{s_1, s_2, \ldots s_N\}$ and its two clusterings namely $\mathbf{U} = \{U_1, U_2, \ldots, U_R\}$ with R clusters and $\mathbf{V} = \{V_1, V_2, \ldots, V_C\}$ with C clusters ($\cap_{i=1}^R U_i = \cap_{j=1}^C V_j = \emptyset$, $\cup_{i=1}^R U_i = \cup_{j=1}^C V_j = S$). The information on cluster overlap between $\mathbf{U}$ and $\mathbf{V}$ can be summarized in the form of a $R \times C$ contingency table $T = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}$ as illustrated in Table I where n_{ij} denotes the number of objects that are common to cluster U_i and V_j. Based on this contingency table, various cluster similarity indices can be built. An important class of criteria for comparing clusterings is based upon counting the pairs of points on which two clusterings agree or disagree. Any pair of data points from the total of $\binom{N}{2}$ distinct pairs in S falls into one of the following 4 categories: (1) N_{11}: the number of pairs that are in the same cluster in both $\mathbf{U}$ and $\mathbf{V}$; (2) N_{00}: the number of pairs that are in different clusters in both $\mathbf{U}$ and $\mathbf{V}$; (3) N_{01}: the number of pairs that are in the same cluster in $\mathbf{U}$ but in the different clusters in $\mathbf{V}$; (4) N_{10}: the number of pairs that are in different clusters in $\mathbf{U}$ but in the same cluster in $\mathbf{V}$. Explicit formulae for calculating the number of the four types can be constructed using entries in the contingency table [3], *e.g.* $N_{11} = \frac{1}{2}\sum_{i=1}^R \sum_{j=1}^C n_{ij}(n_{ij}-1)$. Intuitively, N_{11} and N_{00} can be used as indicators of agreement between $\mathbf{U}$ and $\mathbf{V}$ while N_{01} and N_{10} can be used as disagreement indicators. The Rand Index [8] is defined straightforwardly

TABLE I
The Contingency Table

Clustering **U**/ Clustering **V**	V_1	V_2	$\ldots$	V_C	Sums
U_1	n_{11}	n_{12}	$\ldots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\ldots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\ldots$	n_{RC}	a_R
Sums	b_1	b_2	$\ldots$	b_C	N

as:

$$RI(\mathbf{U},\mathbf{V}) = \frac{N_{00}+N_{11}}{N_{00}+N_{11}+N_{01}+N_{10}} = \frac{N_{00}+N_{11}}{\binom{N}{2}} \tag{12}$$

The Rand Index lies between 0 and 1. It takes the value of 1 when the two clustering are identical and 0 when the two clusterings have no similarity, *i.e.* when one consists of a single cluster and the other only of clusters containing single points. However as can be seen, the unique case where $RI(\mathbf{U},\mathbf{V}) = 0$ is quite extreme and has little practical value. In fact, it is desirable for the similarity index between two random partitions to take values close to zero or at least, a constant value. The problem with the Rand index is that its expected value of two random partitions does not even takes a constant value. Hubert and Arabie [3], by taking the generalized hypergeometric distribution as the model of randomness, *i.e.* the two partitions are picked at random subject to having the original number of classes and objects in each, found the expected value for $(N_{00} + N_{11})$. They suggested using a corrected version of the Rand index of the form:

$$Adjusted\_Index = \frac{Index - Expected\_Index}{Maximum\_Index - Expected\_Index} \tag{13}$$

thus giving birth to the (Hubert and Arabie) Adjusted Rand Index (ARI):

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{N}{2}} \tag{14}$$

The ARI is bounded above by 1 and takes on the value 0 when the index equals its expected value (under the generalized hypergeometric distribution assumption for randomness).

B. Information Theoretic Indices

From the contingency table, it is also possible to directly define another type of agreement indices, which are information theoretic oriented. Let us first define the Entropy of a clustering $\mathbf{U}$. Suppose that we pick an object at random from S, then the probability that the object falls into cluster U_i is:

$$P(i) = \frac{|U_i|}{N} \tag{15}$$

We define the entropy associated with the clustering $\mathbf{U}$ as:

$$H(\mathbf{U}) = -\sum_{i=1}^{R} P(i)\log P(i) \tag{16}$$

$H(\mathbf{U})$ is non-negative and takes the value 0 only when there is no uncertainty determining an object's cluster membership, *i.e.* there is only one cluster. Now we arrive at defining the Mutual Information (MI) between two clusterings:

$$MI(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{R}\sum_{j=1}^{C} P(i,j)\log\frac{P(i,j)}{P(i)P(j)} \tag{17}$$

where $P(i,j)$ denotes the probability that a point belongs to cluster U_i in $\mathbf{U}$ and cluster V_j in $\mathbf{V}$:

$$P(i,j) = \frac{|U_i \cap V_j|}{N} \tag{18}$$

MI is a non-negative value upper bounded by the entropies $H(\mathbf{U})$ and $H(\mathbf{V})$. It quantifies the information shared by the two clusterings and thus can be employed as a clusterings similarity measure [1]. Meila [5] suggested using the so-called Variation of Information (VI):

$$VI(\mathbf{U},\mathbf{V}) = H(\mathbf{U}) + H(\mathbf{V}) - 2I(\mathbf{U},\mathbf{V}) \tag{19}$$

The Variation of Information has been proved to be a true metric on the space of clusterings and has several other interesting properties. The MI and VI however do not have a constant upper bound and hence, are not quite suitable for building the Consensus Index. It is preferable to employ a normalized version of the Mutual Information such as [10]:

$$NMI(\mathbf{U},\mathbf{V}) = \frac{I(\mathbf{U},\mathbf{V})}{\sqrt{H(\mathbf{U})H(\mathbf{V})}} \tag{20}$$

The Normalized Mutual Information (NMI) is bounded in $[0, 1]$, takes the value of 1 when the two clusterings are identical and 0 when the two clusterings are totally independent. In the latter case, the contingency table takes the form of the so-called "independence table" where $n_{ij} = |U_i||V_j|/N$ for all i, j. It can be seen that this scenario is less extreme than the one where the Rand Index takes on a zero value as described above. The NMI however have the same problem as the Rand Index does. Specifically, its baseline values under random partitions do not take on a constant value. To see how this would affect our method for estimating K, we first do a small experiment with the Normalized Mutual Information of the form in (20). The experiment is as follows: given N data points, randomly assign each data point into one of the K cluster with equal probability, check to assure that the final clustering contain exactly K clusters. Repeat this 100 times to create 100 clusterings of N data points and K clusters. The average values of NMI, RI and ARI between all 4950 pairs of clusterings corresponding to this particular value of N and K, *i.e.* averageNMI(N, K), *i.e.* averageRI(N, K) and averageARI(N, K) are recorded. A typical experimental result looks like the one presented in Figure 1.

It can be observed that with the same number of data points, the average value of the NMI and RI between random partitions tends to increase as the number of clusters increases, while the average value of the Adjusted Rand Index is always kept very close to zero. When the ratio of N/K is larger, the average value for NMI is reasonably close to zero, but grows as N/K becomes smaller. This is clearly an unwanted effect for our purpose, since the Consensus Index built upon the NMI would be biases in favour of a larger number of clusters. Thus, an adjusted version of the MI with correction for chance will be necessary for our purpose.

C. Proposed Index: the Adjusted Mutual Information (AMI)

Recently we have developed the adjusted versions for various information theoretic measures for clustering comparison

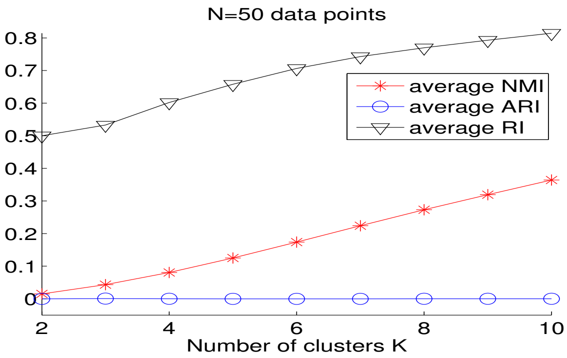


Fig. 1. Average Normalized Mutual Information (NMI), Rand Index (RI) and Adjusted Rand Index (ARI) under random partitions

[12]. To correct the NMI for randomness, it is necessary to specify a model according to that random partitions are generated. A common model for randomness is the “permutation model” [4, p. 214], in which the clusterings are generated randomly subject to having a fixed number of clusters and points in each cluster, *i.e.* $|U_i| = a_i, |V_j| = b_j, i = 1 \ldots R, j = 1 \ldots C$ and the two marginal sums vector $a = [a_i]$ and $b = [b_j]$ are constant, satisfying $\sum_{i=1}^{R} a_i = \sum_{j=1}^{C} b_j = N$. This model has been adopted by Hubert and Arabie when they derived the adjusted version of the Rand index [3]. We shall also adopt this model to derive the adjusted versions for various information theoretic based measures for comparing clusterings. Under the permutation model it can be shown that the probability of encountering a particular contingency table T from a random clustering formation subject to the fixed marginals condition is:

$$\mathcal{P}\{T = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}|a,b\} = \frac{\prod_{i=1}^{R} a_i! \prod_{j=1}^{C} b_j!}{N! \prod_{i=1}^{R} \prod_{j=1}^{C} n_{ij}!} \tag{21}$$

Let $\mathcal{T}$ be the set of all the feasible contingency tables T with marginals a and b. The probability distribution of T in $\mathcal{T}$ as specified by (21) is known as the Generalized Hypergeometric distribution [3], [4]. To correct the NMI for chance we will have to calculate the expected value of the Mutual Information between two random clusterings generated by the permutation model described above. The mutual information of such a pair of clusterings can be calculated from the associated contingency table. In fact, let $MI(T)$ denote the mutual information between (any) two clusterings associated with the contingency table T, clearly we have:

$$MI(T = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}|a,b) = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{N.n_{ij}}{a_i b_j} \tag{22}$$

Thus the average mutual information value between all possible pairs of clusterings is actually the expected value of $MI(T)$ over the set of the associated contingency tables $\mathcal{T}$. This value is given by:

$$E\{MI(T)|a,b\} = \sum_{T \in \mathcal{T}} MI(T) \mathcal{P}\{T|a,b\} \tag{23}$$

$$= \sum_{T \in \mathcal{T}} \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{n_{ij}}{N} \log \frac{N.n_{ij}}{a_i b_j} \mathcal{P}\{T = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}|a,b\}$$

Indeed it can be shown [12] that the value of $E\{MI(T)|a,b\}$ is given by the following expression:

$$\sum_{ij} \sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{\frac{n_{ij}}{N} \log(\frac{N.n_{ij}}{a_i b_j}) a_i! b_j! (N-a_i)!(N-b_j)!}{N! n_{ij}! (a_i - n_{ij})! (b_j - n_{ij})! (N - a_i - b_j + n_{ij})!} \tag{24}$$

with the usual conventions $0 \log 0 = 1, 0! = 1$, and $(a_i + b_j - N)^+$ denoting $\max(0, a_i + b_j - N)$. Having calculated the expectation, we propose an adjusted version of Mutual Information (AMI) following the general form in (13) as:

$$AMI(\mathbf{U}, \mathbf{V}) = \frac{MI(\mathbf{U}, \mathbf{V}) - E\{MI(T)|a,b\}}{\sqrt{H(\mathbf{U})H(\mathbf{V})} - E\{MI(T)|a,b\}} \tag{25}$$

Similar to the ARI, the AMI is bounded in [0,1], takes on the value of 1 when the two clusterings are identical, and the value of 0 when the index equals its expected value. To verify the validity of the approach, we repeat the above experiment. It can be observed from the results in Figure 2 that, just like the ARI, the average value of the AMI between random partitions is now kept close to zero.

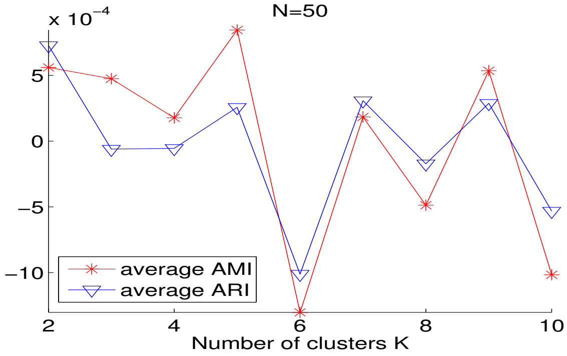


Fig. 2. Average Adjusted Mutual Information and Adjusted Rand Index under random partitions. Both indices show negligible variation with K, in comparison with Fig. 1

IV. Experimental results

A. Method

In this section we present our experimental results of our approach in the context of microarray data clustering. As we have remarked earlier, although our framework is general, its application in a particular domain requires carefully choosing the components of the framework to suit the clustering tasks in that domain, with the most important component being the clustering algorithm. It is reasonable to believe that any procedure for estimating the number of clusters will produce spurious results if the clustering algorithm itself doesn't perform well, and fail to discover a good cluster structure of the data set. In the context of microarray data clustering, we shall consider the following:

1) Clustering algorithm: we use the familiar K-means algorithm equipped with the Euclidean distance. The Euclidean K-means is certainly not a state-of-the-art algorithm for microarray data clustering, however when used correctly in conjunction with a proper data normalization procedure, it will produce reasonable output.

2) Data preprocessing and normalization: As per popular practice in microarray data clustering, we normalize each data point (sample profile or gene profile) to have unit norm and zero mean except for some simulated data sets (to be described in the next section) where there are only 2 genes. This effectively modifies the Euclidean K-means into an algorithm which uses a correlation type measure. Other issues, such as gene selection (for sample clustering), can be found in the respective original papers describing the data sets employed herein.

3) Clusterings generation: to estimate the number of clusters, we generate clusterings with varied number of clusters K ranging from 2 to K_{max} normally set to 15. It is noted that the specific value of K_{max} generally does not affect the Consensus Index and hence the value of K^*, unless K_{max} is set to a value possibly lower than K_{true}. For each value of K, we generate $B = 100$ different clustering solutions by employing the sub-sampling approach [7], *i.e.* performing K-means on 100 different subsets of the original data set. To avoid the situation that K-means is trapped in a bad local optimum, 5 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 *et al.* [14] (Graph Consensus Clustering equipped with K-means - GCC$_{Kmeans}$, or correlation graph clustering - GCC$_{corr}$). The number of clustering solutions was set to the default values for the two Yu's algorithms, *i.e.* $B = 500$, while K_{max} is also set to 15. Experimental results for the two Consensus Clustering algorithms by Monti *et al.* (CC$_{HC}$ and CC$_{SOM}$), whenever available either from [7] or [14], are also reproduced for reference (marked with *).

4) Consensus Index: the Consensus Index is built upon either the Adjusted Rand Index (ARI) or the Adjusted Mutual Information (AMI). Since a sub-sampling strategy is employed, a subset of the original data set might contain data points that are not present in another. The ARI and AMI are, therefore, calculated based on the overlapping portion of any two subsets.

5) Stability assessment: Due to the stochastic elements in the algorithms, the estimate of K^* obtained from different runs might be different. Hence to assess the stability of the algorithms, the whole process for estimating K^* is repeated a few times. Where an algorithm produced more than a single estimate of K^*, all estimates are reported.

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 publicly available from the authors. Also three data sets in [14] are generated using the description in the paper and the source code 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 II.

TABLE II
SUMMARY OF THE SIMULATED DATASETS

Dataset	Source	#Classes	#Samples	#Genes
Synthetic1	[14]	3	75	1000
Synthetic2	[14]	4	100	1000
Synthetic3	[14]	7	140	1000
Uniform1	[7]	1	60	600
Gaussian1	[7]	1	60	600
Gaussian3	[7]	3	60	600
Gaussian4	[7]	4	400	2
Gaussian5 ($\lambda = 3$)	[7]	5	500	2
Gaussian5 ($\lambda = 2$)	[7]	5	500	2
Simulated6	[7]	6-7	60	600
Simulated4	[7]	4	40	600

2) Real Microarray Data Sets: We evaluate our algorithm on both sample clustering and gene clustering. Sample clustering is performed on six real microarray data sets used in [7] with all the data sets details and preprocessing issues described carefully therein. Gene clustering is performed on two yeast cell cycle data sets. The yeast cell cycle data studied by Cho *et al.* [2] showed the fluctuation of expression level of more than 6000 genes over two cell cycles (17 time points). Following [13], we use two different subsets of this data with known class label for each gene: set 1 consists of 384 genes whose expression level peak at different time points corresponding to the five phases of cell cycle; set 2 consists of 237 genes corresponding to four categories in the MIPS database. The summary of the real data sets is provided in Table III.

TABLE III
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

C. Experimental Results on Simulated Data Sets

We first test the algorithms on Yu's three simulated data sets (Synthetic1-3). Experimental results for the two Consensus Clustering algorithms by Monti *et al.* [7] on similar data sets are also reproduced from [14] for reference (marked with *). It can be observed from table IV that these three data sets do not present any serious challenge for all algorithms. All produced perfect results with only the CC_{HC} algorithm misidentifying the true number of clusters in one case.

TABLE IV

Comparison of consensus based number of clusters detection techniques on simulated data sets

Dataset	K_{true}	CC$^*_{HC}$ [7]	CC$^*_{SOM}$ [7]	GCC$_{Kmeans}$ [14]	GCC$_{corr}$ [14]	CI$_{ARI}$	CI$_{AMI}$
Synthetic1	3	3	3	3	3	3	3
Synthetic2	4	4	4	4	4	4	4
Synthetic3	7	6	7	7	7	7	7
Uniform1	1	1	1	15	15	1	1
Gaussian1	1	1	1	15	15	1	1
Gaussian3	3	3	3	3	3	3	3
Gaussian4	4	4	4	8	2	4	4
Gaussian5 ($\lambda = 3$)	5	5	5	8	2	5	5
Gaussian5 ($\lambda = 2$)	5	4-5	4	8	2	4,5	4,5
Simulated6	6-7	7	6	6,7,9,10,13	6,7	6	6
Simulated4	4	4	4	4	4	4	4

TABLE V

Comparison of consensus based number of clusters detection techniques on real microarray data sets

Dataset	K_{true}	CC$^*_{HC}$ [7]	CC$^*_{SOM}$ [7]	GCC$_{Kmeans}$ [14]	GCC$_{corr}$ [14]	CI$_{ARI}$	CI$_{AMI}$
Leukemia	3	5	4	3	3	3	3
Novartis multi-tissue	4	4	4	5,6	6,7	3(5)	2(5)
St. Jude Leukemia	6	5	5/7	5,6,8	5-7	5	3,5
Lung cancer	4+	5	5	7,8	5-7	2(6)	2(6)
CNS tumors	5	5	5	6,7	6,8	2(4)	2(4)
Normal tissues	13	7	4/5	10,11	9,11,12	12-14	12,13
Cho's yeast data 1	5	N/A	N/A	7,8	4,5	4	4
Cho's yeast data 2	4	N/A	N/A	6,8-10	2	4	4

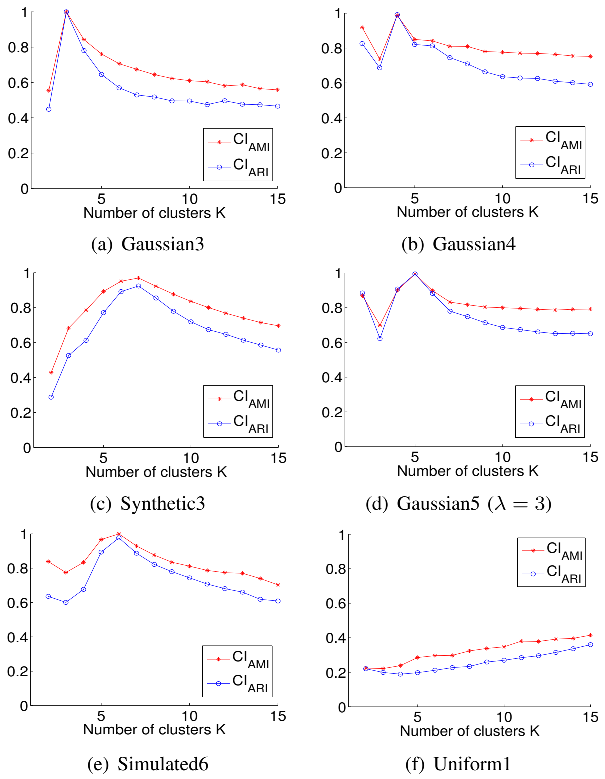
For Monti's simulated data sets, we first examine the multiple-clusters cases. On the Gaussian3, Simulated6 and Simulated4, all the algorithms achieve either the correct result or a close estimation. It is noted however that on the Simulated6 data set, the GCC$_{Kmeans}$ algorithm produces notably unstable results. Inspection of the MRI-vs-K graph (data not shown) showed that the graph is quite flat for the K values from 6 to 15 without a strong global peak. Thus the estimate of K^* obtained does not exhibit strong confidence. For the Gaussian 4, Gaussian5 ($\lambda = 3$) and Gaussian5 ($\lambda = 2$), our approach and the two algorithms by Monti *et al.*, namely the CC$_{HC}$ and CC$_{SOM}$, successfully reveal the true cluster structure. On the other hand, the two algorithms proposed by Yu *et al.* , namely the GCC$_{Kmeans}$ and GCC$_{corr}$, wrongly determine the appropriate number of clusters. This failure may be attributed to the small number of features in these datasets (2), meanwhile the GCC$_{Kmeans}$ and GCC$_{corr}$ use the random subspace clustering technique. With only 2 features, the random subspace clustering scheme thus becomes a random initialization scheme of the clustering algorithms.

It is interesting to observe the form of the CI-vs-K graph (Figure 3). For both the ARI-based and AMI-based Consensus Index, on all the simulated data sets, the CI-vs-K graph shows a strong peak at $K = K_{true}$ indicating that the estimate for K^* exhibits strong confidence.

Finally we examine the two 1-cluster data sets (Uniform1 and Gaussian1) and return to the question of how to differentiate between 1 and 2 or more clusters. In section III we proposed using a threshold α. We observe that for these two data sets the overall values of CI are quite low over the range of K, in particular CI varies in the range [0.2,0.4] while for the multiple-clusters data sets, CI ranges in [0.55,1]. From our experiments with various other 1-cluster data sets (data not shown) we recommend that a reasonable value for α would be in the range of $[0.4, 0.5]$, which indicates very weak agreement between the clustering solutions. For Yu's 2 algorithms, the Modified Rand Index has overall high values, often from 0.7 to 0.95, which is as high as in the other multiple-cluster data sets. This fact suggests that it is not easy to apply a similar thresholding mechanism for this index to cope with the 1-cluster data set cases. In [7], Monti *et al.* commented that manual inspection of the CDFs plot would be necessary for distinguishing between 1 and 2-or-more clusters, however, a concise working rule was not given.

D. Experimental Results on Real Microarray Data Sets

Experimental results for the real microarray data sets are presented in table V along with the $K - CI$ graph on Figure 4. For most of the data sets, namely the Leukemia, Normal tissue, St. Jude leukemia and Cho yeast's data 1 and 2, the CI-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 $K = K_{true}$ (value reported in parentheses). The phenomenon that CI tends to be higher for $K = 2$ has also been observed in simulated data sets (Figure 3). We haven't been able to find a clear explanation for this phenomenon however several hypothesis might be posed. First, at $K = 2$ there might not be as many diverse clustering solutions as at higher values of K, the CI thus has a natural higher value. Second, the data might have a hierarchical cluster structure with 2 clusters being the meta cluster structure. When $K = K_{true}$ a substructure is identified, thus the CI index rises again and gives a local peak.



(a) Gaussian3

(b) Gaussian4

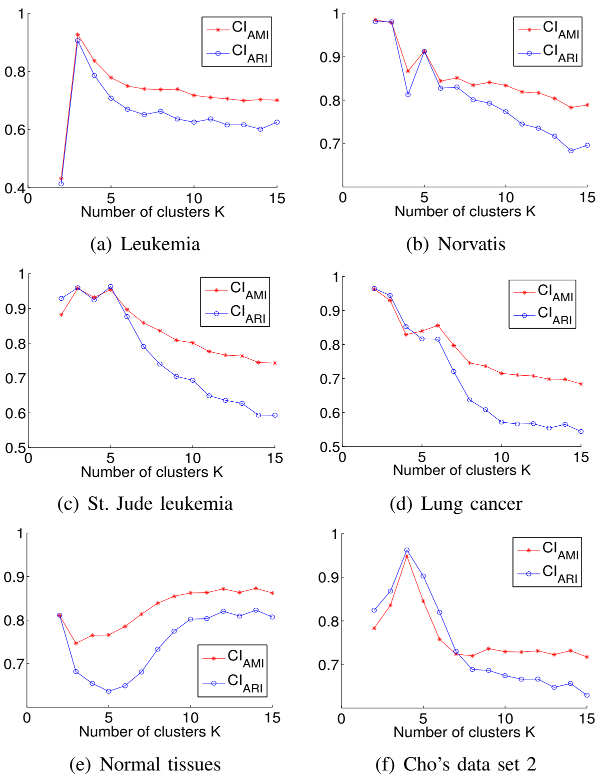
(c) Synthetic3

(d) Gaussian5 ($\lambda = 3$)

(e) Simulated6

(f) Uniform1

Fig. 3. The Consensus Indices on some simulated data sets



(a) Leukemia

(b) Norvatis

(c) St. Jude leukemia

(d) Lung cancer

(e) Normal tissues

(f) Cho's data set 2

Fig. 4. The Consensus Indices on some real data sets

V. Conclusion

In this paper we have presented a new framework for estimating the number of clusters in a dataset based on a novel index. The Consensus Index is built upon either the Adjusted Rand Index or our recently developed Adjusted Mutual Information. The Consensus Index quantifies the agreement between the clustering solutions obtained by a Consensus Clustering approach. The optimal value of number of clusters is chosen as the value that maximizes the Consensus Index, that is, when the cluster structure obtained is most stable in terms of cluster agreement. 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, the data preprocessing and normalization procedure, the similarity measure, the clusterings generation method, must be carefully chosen to fit a particular clustering task. Our extensive experiments on microarray data clustering indicate the usefulness of the CI-based criterion for estimating the appropriate number of clusters. The two CI measures namely the CI_{AMI} and CI_{ARI} tend to give quite concordant results, suggesting that both the ARI and the AMI are well suited for this purpose. Matlab code for computing the AMI is available from http://ee.unsw.edu.au/∼nguyenv/Software.htm

References

[1] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere 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. Wolfsberg, 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. Meilă, "Comparing clusterings: 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 of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[10] A. Strehl and J. Ghosh, "Cluster ensembles - a 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," Standford University, Tech. Rep., 2000.

[12] N. X. Vinh and J. Epps, "Information theoretic measures for clusterings comparison: 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.