On external indices for mixtures: validating mixtures of genes

Ivan G. Costa[1] and Alexander Schliep[1 2]

[1] Department of Computational Molecular Biology
Max-Planck-Institute for Molecular Genetics
Ihnestraße 73, D-14195 Berlin, Germany
[2] Institut für Mathematik-Informatik Martin-Luther-Universität
Halle-Wittenberg, 06099 Halle, Germany

Abstract. Mixture models represent results of gene expression cluster analysis in a more natural way than 'hard' partitions. This is also true for the representation of gene labels, such as functional annotations, where one gene is often assigned to more than one annotation term. Another important characteristic of functional annotations is their higher degree of detail in relation to groups of co-expressed genes. In other words, genes with similar function should be be grouped together, but the inverse does not holds. Both these facts, however, have been neglected by validation studies in the context of gene expression analysis presented so far. To overcome the first problem, we propose an external index extending the corrected Rand for comparison of two mixtures. To address the second and more challenging problem, we perform a clustering of terms from the functional annotation, in order to address the problem of difference in coarseness of two mixtures to be compared. We resort to simulated and biological data to show the usefulness of our proposals. The results show that we can only differentiate between distinct solutions after applying the component clustering

1 Introduction

Biology suggests that a single gene will often participate not in one, but in multiple metabolic pathways, regulatory networks or protein-complexes. As a result, mixture models represent the results of gene expression clustering analysis in a more natural way than 'hard' partitions (Schliep *et al.* (2005)). This is true not only for the clustering results, but also for the representations of gene labels. Biological sources of information, such as functional annotations, transcription binding sites or protein-protein interactions are formed by overlapping categories. However, this has been neglected so far by validation studies for gene expression analysis. A classical approach for comparing two partitions is the use of external indices (Jain and Dubes (1988)). Their basic definition only allows the comparison of 'hard' clusterings. To overcome this limitation, we propose extensions of external indices, such as the corrected Rand (`CR`), suitable for comparing mixtures or overlapping partitions (encoded as mixtures). In order to investigate the characteristics of the proposed index, we make use of experiments with simulated data sets.

Other important characteristics of most biological information are their complex structure, large size and specificity of information. Gene Ontology (G.O. Consortium (2000)), for example, is composed of a redundant directed acyclic graph with thousands of biological terms. The terms in Gene Ontology (GO) can either describe general concepts, such as 'development', which has more then 17.000 annotated genes, or very specific concepts, such as 'pupal cuticle biosynthesis', which has only one associated gene. The construction of a 'compact' and 'meaningful' mixture from such complex structure is non-trivial. Furthermore, one should not expect that the information contained in a single gene expression data set is as specific as the information contained in GO. Biologically speaking, co-regulated genes should share similar function, but clusters of co-regulated genes will be associated not with one, but with several biological functions. The use of `CR` to compare two mixtures (or partitions), where one of the mixture represents a more coarse representation of the data, yields too conservative `CR` values, given the high number of false positives. As a consequence, a procedure for clustering GO terms prior to the comparison of the mixtures – clustering of components – is necessary in order to achieve more general representations of GO. This compact representation of GO yields a better basis for comparison of distinct results. To evaluate the proposal, we perform analysis of gene expression time-courses from Yeast during sporulation (Chu *et al.* (1998)). The results with and without the component clustering are then compared with Yeast annotation from GO.

2 External Indices

External indices assess the agreement between two partitions, where one partition U represents the result of a clustering method, and the other partition V represents a priori knowledge of the clustered data. A number of external indices have been introduced in the literature, but the use of corrected Rand (`CR`) has been suggested given its favorable characteristics (Hubert and Arabie (1985)). Among others, `CR` has its values corrected for chance agreement, and is not dependent of the object distribution in U or V (Milligan and Cooper (1986)). This work proposes an extension of the corrected Rand, in order to access the agreement of partitions with overlap (encoded as mixtures) or mixture models, by comparing their posterior distributions for a fixed data set. The main idea of the extended corrected Rand (`ECR`) is to redefine the indicator functions, as defined in Jain and Dubes (1998), giving them a probabilistic interpretation.

To simplify the notation, we consider for a given mixture model $f(\cdot|\Theta) = \sum_{k=1}^{K} \alpha_k f_k(\cdot|\Theta_k)$ [1] the components $U = \{u_k\}_{1 \leq k \leq K}$; similarly $V = \{v_l\}_{1 \leq l \leq L}$ for a second mixture model. Let $O = \{o_n\}_{1 \leq n \leq N}$ be the set of objects to be clustered, U be the estimated mixture model (or clustering solution), and V

[1] Θ_k and α_k are the mixture model parameters (McLachlan and Peel (1996))

be the mixture defined by the *a-priori* classification. The posterior distribution defines the probability that a given object $o \in O$ belongs to a component u_k from U or v_l from V, $\{\mathbf{P}[u_k|o]\}_{1\leq k\leq K}$ and $\{\mathbf{P}[v_l|o]\}_{1\leq l\leq L}$. We denote the event that a pair of objects has been generated by the same component in model U, the co-occurrence event, as $o_i \equiv o_j$ given U. Assuming independence of the components in U, the probability of the co-occurrence of o_i and o_j given U for $1 \leq i \leq j \leq N$ is:

$$\mathbf{P}[o_i \equiv o_j \text{ given } U] = \sum_{k=1}^{K} \mathbf{P}[u_k|o_i]\mathbf{P}[u_k|o_j] \tag{1}$$

We use the above formula to redefine the variables a, b, c and d, used in the definition of `CR`, which are equivalent to the number of true positives, false positives, false negatives and true negatives respectively.

$$a = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{P}[o_i \equiv o_j \text{ given } U]\mathbf{P}[o_i \equiv o_j \text{ given } V] \tag{2}$$

$$b = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{P}[(o_i \equiv o_j \text{ given } U)^C]\mathbf{P}[o_i \equiv o_j \text{ given } V] \tag{3}$$

$$c = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{P}[o_i \equiv o_j \text{ given } U]\mathbf{P}[(o_i \equiv o_j \text{ given } V)^C] \tag{4}$$

$$d = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{P}[(o_i \equiv o_j \text{ given } U)^C]\mathbf{P}[(o_i \equiv o_j \text{ given } V)^C] \tag{5}$$

From these the extended corrected Rand (`ECR`) can be calculated by the original formula for the `CR`, as defined below.

$$ECR = \frac{(a+d) - ((a+b)(a+c)+(c+d)(b+d))p^{-1}}{p - ((a+b)(a+c)+(c+d)(b+d))p^{-1}} \tag{6}$$

where p is equal to the sum $a+b+c+d$ or the total number of object pairs.

`ECR` takes values from -1 to 1, where 1 represents perfect agreement while values of `ECR` near or below zero represent agreements occurred by chance. The original `CR`, proposed in Hubert and Arabie (1984), estimates the expected Rand value by assuming that the baseline distributions of the partitions are fixed. By definition, `ECR` is an extension of `CR`. It works exactly as the latter when hard partitions are given. In the used terminology, a 'hard' partition can be described by the following posterior.

$$\mathbf{P}[u_k|o] = \begin{cases} 1, \text{ if } o \in u_k \\ 0, \text{ otherwise} \end{cases} \tag{7}$$

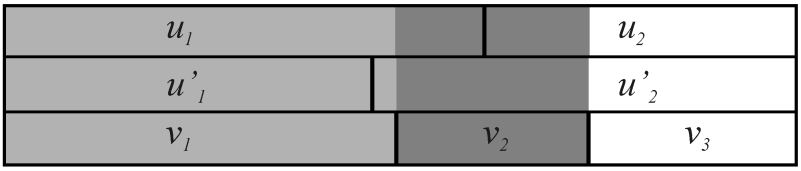

Fig. 1. We display three hypothetical partitions, U and U', which represent two distinct clustering results, and V, which represents the true labels (the objects in U and U' are depicted in the correspondent label color defined in V). Both clusterings failed to recover the three true components. U splits the objects from v_2 in half, while U' joined the objects of v_2 and v_3. Comparing the partitions with V, U has a `CR` value of 0.57 and U' a value of 0.53. Assuming, however, that the classes v_2 and v_3 can not be distinguished in the clustered data, and joining these two components, U would have a `CR` of 0.56 while U' a value of 0.78.

3 Component Clustering

The component clustering deals with the problem of difference in coarseness of two mixtures (or partitions). Given the two mixtures U and V, using the `ECR` (or `CR`) to compare the agreement will always result in low values when $\#U << \#V$, even when U is a more coarse representation of V. A simple example of this, in the context of partitions, can be seen in Fig 1. In some real world problems, as with the use of functional annotation of genes to validate co-regulation of genes, it is reasonable to assume that U is a coarser representation of V, and the clustering of components in V yields a better comparative basis for choosing between distinct solutions U, and hence between different methods.

More formally, given that the number of components of the model V is higher then the one in U, and assuming that model U is a more general description of V, we want to find a partition $P = \{p_k\}_{1 \leq k \leq K}$ of the components in V. This partitioning can be used to define a new model V', where each group of components in P is a single component in V' and V' is similar to U. A natural choice of a criterion for evaluating the 'similarity' of the two models is the mutual information.

$$I(X,Y) = \sum_{i=1}^{L} \sum_{j=1}^{J} \mathbf{P}[X = x_i, Y = y_j] \log \left(\frac{\mathbf{P}[X = x_i, Y = y_j]}{\mathbf{P}[X = x_i]\mathbf{P}[Y = y_j]} \right) \quad (8)$$

Given mixture models U and V, its posteriors on O and, assuming independence between them, we can define the joint probability $\mathbf{P}[U, V|O]$ and the probability distribution $\mathbf{P}[U|O]$ as:

$$\mathbf{P}[U = u_k, V = v_l|O] = \frac{1}{N} \sum_{i}^{N} \mathbf{P}[u_k|o_i]\mathbf{P}[v_l|o_i] \quad (9)$$

$$\mathbf{P}[U = u_k|O] = \frac{1}{N}\sum_{i}^{N} \mathbf{P}[u_k|o_i] \tag{10}$$

We accomplish the components clustering by applying a algorithm similar to hierarchical clustering. It joins a pair of groups of components at a time, until a certain number of clusters is reached. At each step, the partition in the set of candidate partitions (C) with higher mutual information is selected. Starting with the singleton partition, where $p_i = \{v_i\}$ for $1 \leq i \leq L$, the method works as follows:

1. while ($\#P > \#U$) do
2. $C = \emptyset$
3. for each pair (p_i,p_j), where $1 \leq i < j \leq \#P$ do
4. $P' = P \setminus p_j$
5. $p'_i = p_i \cup p_j$
6. $C = C \cup \{P'\}$
7. $P = argmax_{H \in C} I(U, merge(V, H))$

where $merge(V, P)$ defines a new model V' from V, where $\#V' = \#P$ and $P[v'_k|o] = \sum_{i \in p_k} \mathbf{P}[v_i|o]$.

4 Experiments

To evaluate the extended corrected Rand, we make use of simulated data from multivariate mixture of normals. We use a simple test data with two normal components to compare the characteristics of `ECR` and `CR` when distinct overlaps are present. Then, we make use of biological data in order to show the applicability of the proposal, in particular the component clustering method, to real data. The Estimation-Maximization algorithm (EM) is used to fit multivariate normal mixtures with unrestricted covariance matrices (McLachlan and Peel (1996)). For each data set, 15 repetitions of the EM algorithm with random initialization are performed, and the result with maximum likelihood is selected. In the simulated data experiments, 50 test data sets are generated for each proposed mixture.

4.1 Simulated Data

We perform experiments with a normal mixture with two equiprobable components to evaluate the proposed index characteristics in the presence of distinct overlaps. The components have means $\mu_1 = [0, 0]^T$, $\mu_2 = [d, 0]^T$, covariant matrices $C_1 = C_2 = I$, and $0.0 < d < 7.5$ (structured data) (Figueiredo and Jain (2002)). For each component we draw 200 samples (or objects), and the multivariate normal density of the mixtures are used to obtain the distributions $\mathbf{P}[V|o]$. We also display the value from the `CR`, by the following partition assignment of the objects of a given posterior distribution:

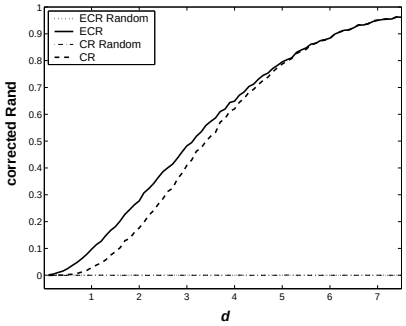

Fig. 2. We show results of the mixture estimation with the normal bivariates. The larger d, the lower is the overlap between the two components.

$$\mathbf{P}[u_i|o] = \begin{cases} 1, \text{ if } i = argmax_{1 \leq k \leq K}(\mathbf{P}[u_k|o]) \\ 0, \text{ otherwise} \end{cases} \tag{11}$$

Additionally, we generate random noise data to serve as a null case. This consists of data generated from a single normal component with $\mu = [d/2, 0]^T$ and $C = I$. A 'hypothetical solution' (V) with the same number of components and object distributions is calculated from the definition of the respective structured data. We carried out a non parametric equal-means hypothesis test based on bootstrap (Efron and Tibshriani (1997)) to compare the mean `ECR` (or `CR`) obtained with the structured ($\overline{s}$) and random data ($\overline{r}$).

$$H_0 : \overline{r} = \overline{s} \text{ and } H_1 : \overline{r} < \overline{s} \tag{12}$$

As displayed in Fig. 2, for data with high overlap, `ECR` has higher values then `CR`, while for data with low overlap both indices have similar values. With random data, both indices take on mean values near zero and low variance (< 0.001), which indicates that `ECR` is successful in the correction for randomness. In relation to the hypothesis test, H_0 is rejected in all d values with $\alpha = 0.001$ with the use of `ECR`, while for data with very high overlap ($d < 0.4$) the null hypothesis is not rejected ($\alpha = 0.001$) with the use of `CR`. From these we can conclude that `ECR` is able to show significant distinctions between the agreement of the random and structured data, even when the overlap is great, while `CR` fails.

4.2 Biological Data

We use gene expression data from Yeast (Chu *et al.* (1998)) in our evaluation. This data set contains gene expression measurements during sporulation for over 6400 genes of budding yeast. The measurements were taken at seven time points (0h, 0.5h, 2h, 5h, 7h, 9h and 11h). Clones with more than 20% of values missing were excluded. The data is pre-processed by extracting all those genes

with an absolute fold change of at least two in at least one time point. The resulting data set contains 1171 genes. We perform mixture estimation, as described in Sec. 4, and we use the Bayesian information criteria to determine the optimal number of components (10 for this data set).

Gene Ontology Gene Ontology (GO) describes genes in three distinct categories (G.O. Consortium (2000)): cellular component, molecular function and biological process. Such an ontology has the form of a directed acyclic graph (DAG), where the leaves are genes and the internal nodes are terms (or annotations) describing gene function, gene cellular localization or the biological processes genes take part in. Gene are associated not only with the terms which it is directed linked, but also to all parents of this term. Given this parent relation and the number of GO terms, a reasonable way to obtain a mixture from GO is to cut it at a fixed level m, where each GO term in level m represents one component from the mixture $T^m = \{t_p^m\}_{1 \leq p \leq P}$. For a given set of genes O, one could define a simple definition of a posterior distribution of a gene o given T^m by:

$$\mathbf{P}[t_p^m|o] = \begin{cases} 1/\#\{i|o \in t_i^m, i = 1, ..., P\}, & \text{if } o \in t_p^m \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

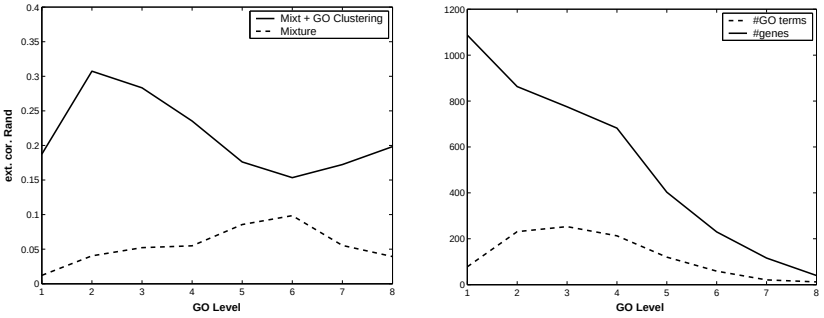

Fig. 3. In the left, we show the ECR values obtained for distinct levels of GO and in the right we show the number GO terms and annotated genes for distinct GO levels. The higher the level the lower the number of genes. The number of GO terms increases until level 3 reaching a peak of 234, and decreases afterwards.

The use of the component clustering posterior to the mixture estimation represented a considerable increase in the ECR values (Fig. 3), while the ECR values obtained only with the mixture estimation are not too far apart from zero (similar results are encountered with other gene expression data sets). The main reason for this difference is the reduction in the number of false positives obtained after the application of the clustering of components. In

relation to the use of GO, the choice of the level of cutting the DAG is a rather subjective task. Figure 3 shows that high levels of GO should be avoided, since there is a lower percentage of annotated genes. The levels two and three represent a better choice, since they obtained the highest ECR while they still maintain a reasonable number of genes. These characteristics, however, are dependent on the data set analyzed and on the GO annotation used.

5 Conclusions

The use of simulated data allow us to assess the characteristics of the extended corrected Rand. It displayed superior results in comparison to the original corrected Rand when high overlap is present and values near zero when the data is random. With the biological data, the results indicate that (1) there is a low agreement between the results of mixture analysis and GO and (2) this agreement is greatly enhanced by a clustering of components. We can conclude that the use of component clustering prior to ECR is important when structures with distinct level of coarseness are compared allowing to choose between different solutions which were previously not very distinguishable. Despite the importance of this problem, it has been neglected in the bioinformatics literature, where in several problems we are faced with the comparison of data with such distinctions in coarseness.

References

CHU S., *et al.* (1998), The Transcriptional Program of Sporulation in Budding Yeast, *Science*, 282, 5389, 699-705.

EFRON B. and TIBSHIRANI, R. (1993), An Introduction to the Bootstrap, Chapman & Hall, New York.

FIGUEIREDO M. and JAIN, A.K. (2002), Unsupervised learning of finite mixture models, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24, 3, 381-396.

HUBERT, L. J., ARABIE, P. (1985), Comparing partitions, *Journal of Classification*, 2, 63-76.

JAIN A.K., DUBES, R.C. (1988), *Algorithms for clustering data*. Prentice Hall, New York.

MCLACHLAN G. and PEEL D. (2000), *Finite Mixture Models*, Wiley, New York.

MILLIGAN G. W. and COOPER M. C. (1986), A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavorial Research*, 21, 441-458.

SCHLIEP, A., COSTA, I.G., STEINHOFF, C. and SCHÖNHUTH, A. (2005), Analyzing gene expression time-courses , *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3), 179-193.

T. G. O. CONSORTIUM (2000), Gene ontology: tool for the unification of biology, *Nature Genet*, 25, 25-29.