On external indices for mixtures: validating mixtures of genes

Ivan G. Costa[1] and Alexander Schliep[12]

[1] Department of Computational Molecular Biology
Max-Planck-Institute for Molecular Genetics
Ihnestraße 63–73, 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 clustering analysis in a more natural way than 'hard' partitions. This is also true for the representation of gene labels, such as functional annotation, 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 biological data to show the usefulness of our proposals. The results show that only after applying the component clustering we can differentiate between different solutions.

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 (McLachlan and Peel (1996)) represent the results of gene expression clustering analysis in a more natural way than 'hard' partitions. This is true not only for the clustering results, but also for the representation 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)). However, 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, suitable for comparing mixtures or overlapping partitions (encoded as mixtures).

Other important characteristics of most biological information are their complex structure, great 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 of annotated genes, or very specific concepts, as 'pupal cuticle biosynthesis', which has only one associated gene. The proposition 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 corrected Rand 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 performed 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 defined 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.

Let $O = \{o_n\}_{1 \leq n \leq N}$ be the set of objects, $U = \{u_k\}_{1 \leq k \leq K}$ be the mixture given as a clustering solution , and $V = \{v_l\}_{1 \leq l \leq L}$ 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 $\{\mathbf{P}[v_l|o]\}_{1 \leq l \leq L}$. Defining the event that a pair of objects has been generated by the same component in model U or the co-occurrence event as, $o_i \equiv o_j$ given U and 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 terms a, b, c and d, which are, in the original definition of CR, equivalent to the number of true positives, false positives, false negatives and true negatives:

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

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

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

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

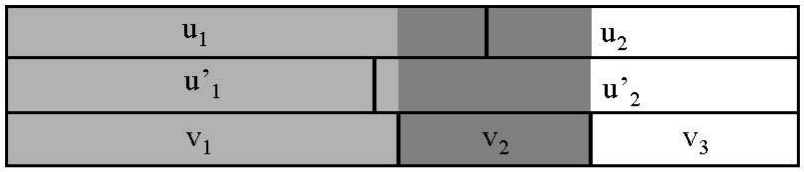

Fig. 1. We display three hypothetical partitions, U and U', which represent two distinct clustering results, and V, which represents the true labels (the objects in U and U' are depicted in the correspondent label color defined in V). Both clustering results fail to recover the three true components, while U splits the objects from v_2 in half, U' joins 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 detected in the clustered data, and joining these two components, U would have a `CR` of 0.56 while U' a value of 0.78.

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

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

`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.

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

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.

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

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:

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

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

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

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

We make use of biological data in order to show the applicability of the proposal, in special 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 with random initialization are performed, and the result with maximum likelihood is selected.

4.1 Gene Expression 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 performed mixture estimation, as described in Sec. 4, and we use the Bayesian information criteria (McLachlan and Peel (1996)) 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 annotations which it is directed linked, but also to all parents of the linked node. Given this parent node 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:

$$\mathbf{P}[t_p^m|o] = \begin{cases} 1/\#\{i|o \in t_i^m, i = 1, ..., P\}, \text{ if } o \in t_p^m \\ 0, \qquad\qquad\qquad\qquad \text{otherwise} \end{cases} \tag{11}$$

4.2 Results

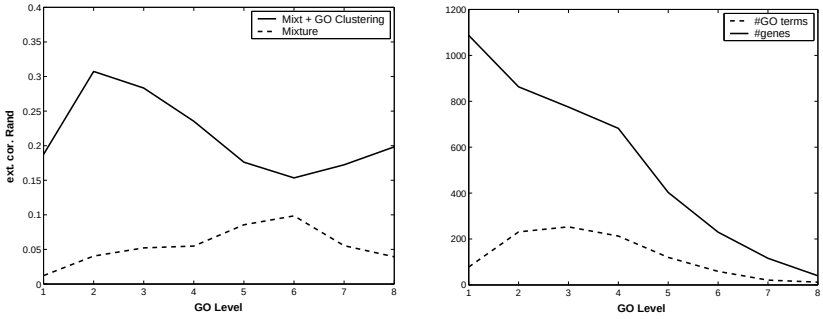

Fig. 2. In the left, we show the ECR values obtained for distinct levels of GO and in the right we shown the number GO terms and annotated genes for distinct GO levels. The higher the level the lower the number of genes. The number of GO terms increases until level 3 reaching a peak of 234, and decreases afterwards.

The use of the component clustering posterior to the mixture estimation represented a considerable increase in the ECR values (Fig. 2), 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 2 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 results indicates that (1) there is a 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 previous to ECR is important when structures with distinct level

of coarseness are compared allowing to choose between different solutions which were previously indistinguishable. 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. Among others, this seems to be the case of data from gene annotation, protein-protein interactions, metabolic pathways and transcription binding sites of genes.

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.

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