A Novel HMM-based Cluster Validity Index for Gene Expression Time-Course Data

Yujing Zeng[1], Javier Garcia-Frias[1]

Keywords: gene expression time-course data, clustering validation, temporal dependence, hidden Markov model (HMM)

1 Introduction.

Gene expression time-course data is usually obtained by performing microarray experiments at consecutive time points. The analysis of these data is helpful to reveal the mechanisms regulating different cellular processes. Different from other microarray data, the pattern represented by each profile is decided not only by the observations at different time-points, but also by the order of these observations. This sequential dependence is crucial for clustering processing and validation.

Our research addresses the problem of clustering validation for time-course data by proposing a novel hidden Markov model (HMM)-based clustering validity index. In the index definition, we use a specially designed HMM to model the data distribution under the constraints of the clustering result. The evaluation is calculated based on the likelihood of each time series given this HMM. The main novelty of the proposed index is its ability to take account of the temporal dependences in the sequential data. Contrary to other validity indexes, in the proposed model the observations at different time-points are not considered independently, and the dependences in each time-interval are modeled and used to evaluate the clusters quality explicitly. In other words, if these dependences change because of permutations among time-points, the validation result is able to reflect such a change accurately. The simulation discussed in next subsection was designed to test this ability.

2 Simulation.

In this experiment, we generated two datasets which have the same observations but in different order. Any object in both datasets has 9 attributes, which can be divided into two categories. The first category includes 4 attributes, whose values are i.i.d. samples from zero-mean Gaussian distributions. Since there is no useful information for clustering in these 4 features, we called them noninformative features. The rest 5 attributes, called informative features, are generated from three different basic patterns, as shown in Figure 1(a). To generate similar but distinct sequences, random variables with different distributions are added to each component of the basic patterns, which is shown in Figure 1(b, c). As shown in Figure 1(d), all the informative features are observed at the first 5 consecutive time-points in dataset I, which shows three distinctive patterns. Dataset II is generated by permuting the order of the attributes in dataset I so that each informative feature is separated by noninformative features, as shown in Figure 1(e). Since all the noninformative features are i.i.d. and unrelated with the basic patterns, the dependence between a noninformative and an informative feature is very weak, and, therefore, interleaving the two kinds of features reduces the sequential dependences in the whole data.

In order to corroborate the effectiveness of the proposed HMM-based index, the partitions with different numbers of clusters, from 2 to 10, were obtained using one of the most popular and simplest methods, K-means. As shown in Table 1, although the datasets share the same observations, the

[1] Department of Electrical and Computer Engineering, University of Delaware, Delaware, USA. E-mail: {zeng,jgarcia}@ee.udel.edu

proposed algorithm produces different evaluations for the two datasets, and suggests different optimal numbers of clusters for them.

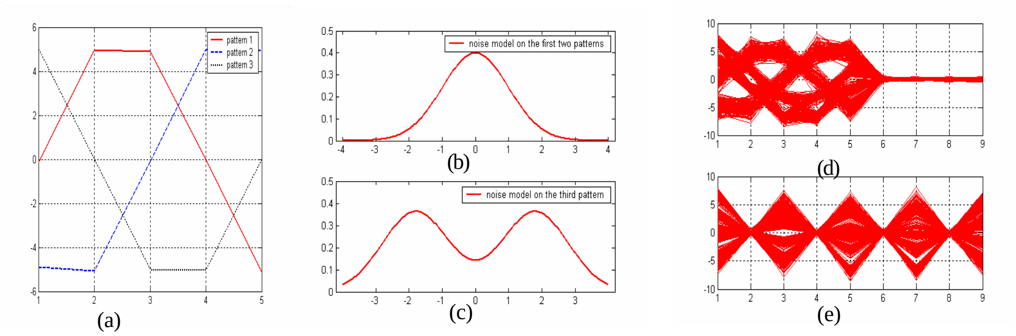

Figure1: Illustration of the generation of the simulation data. (a) three basic patterns for the informative features; (b) density function of the noise model added on the first two patterns; (c) density function of the noise model added on the third pattern; (d) simulation dataset I; (e) simulation dataset II.

According to the results of the proposed index, the optimum number of clusters for dataset I is estimated as 3, which agrees with the original model generating the data. Notice that because of the bimodal noise added to the third basic pattern, this cluster shows a more dispersive distribution than the other two and tends to be split into several groups. This is reflected in the curve of the proposed index, which displays a sub-optimum for the partition with 6 clusters. In dataset II, the sequential dependences are reduced by permutation, so that the noise model becomes overwhelming and gets emphasized in the clustering evaluation. As shown in Table 1, although a sub-optimum is shown for the partition with 3 clusters, the proposed index suggests that the optimal number of clusters for dataset II is 6.

Comparison has been performed with the results of the Silhouette [1] and Davis-Bouldin indexes [2]. Both indexes ignore the sequential dependences in the data and give the same evaluation on both datasets, which, as shown in Table 1, is similar to the result of the proposed index for dataset II.

Index	2	3	4	5	6	7	8	9	10
HMM-based index for Dataset I	0.5000	**0.7042**	0.4845	0.3590	***0.6542***	0.4793	0.3987	0.2058	0.2101
HMM-based index for Dataset II	0.5000	**0.6115**	0.4196	0.4960	***0.6156***	0.4473	0.3741	0.1915	0.1959
Davis-Bouldin index	0.6607	**0.4407**	0.8404	0.6252	***0.4306***	2.2809	2.2748	2.2252	2.2057
Silhouette index	0.5388	**0.7134**	0.5593	0.6688	***0.7260***	0.6536	0.4945	0.4133	0.3422

Table 1: Values of different indexes for the two simulation datasets

3 Conclusion.

We have illustrated the ability of the proposed index to capture the sequential dependence of time series data. This ability is reflected in the higher robustness of the HMM-based index to deal with data corrupted by noise models more general than the Gaussian one, which is a very attractive property in the context of gene expression data analysis.

References

[1] Davies, D.L and Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.1, pp. 224-227, 1979.

[2] Gordon, A. *Classification*. Chapman and Hall, 2nd edition, 1999.