

Class Discovery and Classification of Tumor Samples Using Mixture Modeling of Gene Expression Data - Supplementary Material

Analysis and Results of the Leukemia Dataset

No unknown class - UK_0

The top-left plot of figure S1 (or 1(a) in the paper) shows the estimated number of classes using BIC, for each of the 200 classifier sizes, and the corresponding prediction accuracy rate. We can see that the best prediction accuracy rate of 1.0 is achieved for the classifiers with 51, 54, and 55 genes. That is, there is no prediction error for a small number of classifier sizes, as indicated in the bottom half of table 1 (in paper), which records the best prediction results among the 200 classifier sizes. We also observe that, for datasets with a moderate number of genes, say $80 \leq G \leq 140$, BIC identifies the correct number of classes, and the class assignments of the test samples are fairly accurate, with only one mispredicted sample, 67, as indexed in Golub *et al.*, (1999), corresponding to an accuracy rate of .97. The second sample that we also misclassified for larger classifiers ($G > 140$) is 66. These two samples seem to be rather non-conforming, as they are frequently mispredicted by other authors as well (Golub *et al.*, 1999; Furey *et al.*, 2000; Dudoit *et al.*, 2002)

One unknown class - UK_1

There are three UK_1 data sets, each containing test samples from one unknown class. The results show that the number of classes is correctly identified for most classifiers with at least a moderate number of genes. Furthermore, the corresponding prediction accuracy rates are usually quite high. For example, for UK_1 -AML (entry (3,1) in figure S1 or figure 1(b) in paper), the prediction accuracy rate is 0.98 for all classifiers with at least 75 genes, with the only prediction error being sample 67, as shown in table 1.

Two unknown classes - UK_2

For the three UK_2 data sets, the problems are harder, and the results are more sensitive to the classifier size, for two of the datasets. For UK_2 -AML-ALLT (entry (3,2) in figure S1 or figure 2(c) in paper), for instance, the best prediction accuracy of 1.0 is achieved in several narrow ranges ($G \in \{57 - 60, 62 - 64, 68 - 73\}$). For most of the classifiers with size between 50-100, the correct number of classes are identified with at most two prediction errors. However, for larger classifiers ($G > 170$), usually the number of classes is underestimated. Similar pattern is observed for the dataset UK_2 -ALLB-ALLT, with the correct number of classes identified for most mid-size classifiers

($50 \leq G \leq 160$). On the other hand, for UK₂-AML-ALLB, the performance of BIC is rather insensitive to the classifier size. These results are not surprising. Since the class ALL-T has the smallest number of samples in it, they can easily be absorbed into one of the other two classes, if the number of genes used for the analysis is too large to mask the true discriminant powers of the other genes, unless ALL-T is the known class as in UK₂-AML-ALLB.

Three unknown classes - UK₃

In this case, all 72 leukemia samples are regarded as test samples. Despite being the most difficult problem due to the lack of training data, the results in entry (1,2) of figure S1 (figure 1(d) in paper) nonetheless compare reasonably well to those from the other variants. The best predictions are achieved with $G \in \{58 - 59, 63 - 65, 70, 72 - 75\}$, having just one error. Other than two small ranges of the classifier size, the number of classes are correctly identified for mid-size classifiers, with the prediction accuracy rates reaching 90% for many of the classifiers with at least 50 genes.

Analysis and Results of the Colon Cancer Dataset

The colon cancer data of Alon et al. (1999) consists of 40 tumor samples and 22 normal samples. First, we analyze Colon57, the data excluding the five frequently misclassified samples (thought to be likely contaminated by some authors, e.g., Alon *et al.* (1999) and McLachlan *et al.* (2002)), as in Soukup and Lee (2003). We analyze the data in four ways: (1) assuming the availability of training samples for both the cancer type and normal type (UK₀), (2)-(3) assuming the availability of training samples for either the cancer type (UK₁-N) or the normal type (UK₁-T) but not both, and (4) assuming that there is no training data available (UK₂). For each analysis, training samples, if they exist, are selected according to the best split obtained by Soukup and Lee (2003), which would give advantage to their method for comparison purpose. The data are first preprocessed as described in Soukup and Lee (2003), and genes are selected as done for the Leukemia data in our study. The maximal number of classes considered is 10, as in the Leukemia analysis. BIC is used as the criterion for model selection. The prediction accuracy rate for Colon57-UK₀ is 100% for almost all classifiers with more than 19 genes. With only two genes, we misclassify a cancer sample (T20; the sample identifiers used are as in Alon *et al.* (1999) for this and the subsequent samples) as a normal sample. Since the best split for Soukup and Lee (2003) is unlikely to be the best split for our method, we believe that our results are comparable with those obtained by Soukup and Lee (2003) for classification. For Colon57-UK₁-N, our best results misclassified a cancer sample (T8), whereas for Colon57-UK₁-T, our prediction accuracy is 0.89, with five cancer samples (T2, T3, T7, T8, T37) misclassified as normal. For Colon57-UK₂, our best results also misclassify five cancer samples (T2, T6, T8, T11, T16).

We then reanalyze the entire data set containing all 62 samples, with the five additional samples all placed in the test sets. For Colon62-UK₀, the best prediction have four misclassifications (N34, N36, T33, T36), all involving the hard-to-classify ones. For Colon62-UK₁-N, there are five misclassified samples, including the four misclassified in Colon62-UK₀ and an additional sample (T8). For

Colon62-UK₁-T, there are now 9 misclassified samples, including the 4 in Colon62-UK₀ and the 5 tumor samples misclassified in Colon57-UK₁-T. These results indicate that having training samples for the cancer type is important for correct classification of the tumor samples. For Colon62-UK₂, when there is no training data at all, our best result has 5 misclassified samples (T2, T19, T30, T31, T37) achieved with a classifier of size 206. Interestingly, for this variant, T30 is the only sample among the five hard-to-classify ones, while T30 is the only one among them to be classified correctly in the other three variants. In summary, for the more difficult colon dataset, our method performs well compared to other methods for the pure classification or class discovery problems. Furthermore, encouraging results are also obtained for the situations with partial training sets, which are not being tackled by the other methods.

Legends of Supplementary Figures

Figure S1 (page 4): Prediction accuracy rates, for Leukemia Variants UK0 - UK3, using preprocessing method (A). The 8 Figures on the left are based on BIC model selection, while those on the right half are based on AIC model selection. For these figures as well as those in S2 and S3, red, blue and green lines indicate that the number of classes selected by BIC or AIC is $M < 3$, $M = 3$, and $M > 3$, respectively.

Figure S2 (page 5): Prediction accuracy rates for Leukemia variants UK0 - UK1, using preprocessing method (B). The 4 figures on the left half are based on BIC model selection, while the 4 figures on the right half are based on AIC model selection.

Figure S3 (page 6): Prediction accuracy rates for Leukemia variants UK0 - UK1, using preprocessing method (C). The 4 figures on the left half are based on BIC model selection, while the 4 figures on the right half are based on AIC model selection.

Figure S4 (page 7): Results for simulation studies based on Leukemia variants UK0 - UK3, using preprocessing method (A). The 8 figures on columns 1 and 3 present percents of datasets for which BIC and AIC select models with M components. Red, blue, green lines indicate datasets where BIC selects models with $M < 3$, $M = 3$, and $M > 3$ components, respectively, while black lines indicate datasets where AIC selects models with $M = 3$ components. The 8 figures on columns 2, 4 present several summary statistics of the number of errors among the datasets for which $M=3$ is correctly inferred using BIC. Green, red, and blue curves indicate the first quartile, median, and the third quartile, respectively, of the number of the errors. Note that the gaps in the summary statistics plots for UK2-AML-ALLB and UK2-ALLB-ALLT are due to the fact that there is no dataset whose number of classes is correctly inferred.

Figure S5 (page 8): Prediction accuracy rates for Colon57 variants UK0-UK2 (left column) and the corresponding variants for Colon62 (right column). Red, blue and green lines indicate that the number of classes selected by BIC is $M < 2$, $M = 2$, and $M > 2$, respectively.







