Chapter 3

Supervised machine learning (II)

In the previous chapter, we have covered major supervised machine learning methods. In this chapter, we will discuss several fundamental machine learning concepts in real applications. Chapter 3.1 focuses on feature selection problems. Chapter 3.2 discusses issues related to performance evaluations. The cross validation technique when no external validation data are available and various performance evaluation criteria are addressed. In Chapter 3.3, the concept of overfitting and underfitting is introduced. Chapter 3.4 discusses the issue of choosing the most suitable machine learning method(s) in a given application. Finally, Chapter 3.5 discusses common mistakes and further issues in machine learning of genomic data.

3.1 Feature selection

In genomic data analysis, thousands of genes (features) are assessed for each sample. Since large portion of the genes provide noisy information irrelevant to the class label, adequate filtering procedures to eliminate such irrelevant genes are often desirable to improve the prediction performance.

3.1.1 filtering methods

Example of filtering:

Figure 3.1 shows a heatmap of correlatioin matrix between any pair of samples in a leukemia microarray data set (from "ALL" package in Bioconductor). When all 12,625 genes are used, the heatmap does not show any clear pattern. We then filter out genes that have standard deviation (across samples) smaller than 1. In the remaining 379 genes, the heatmap of correlation matrix shows clearer with-group patterns (black: ALL1/AF4, red: BCR/ABL, green: E2A/PBX1, blue: NEG). Particularly, in the NEG group there seems to be two clear subclusters. Further investigation finds that the two subclusters belong to NEG B-cell and NEG T-cell (Figure 3.2). See exercise 1 for steps to repeat this analysis. This example demonstrates the need to filter out irrelevant genes (i.e. genes do not highly fluctuate in this case) and its power to improve the analysis.

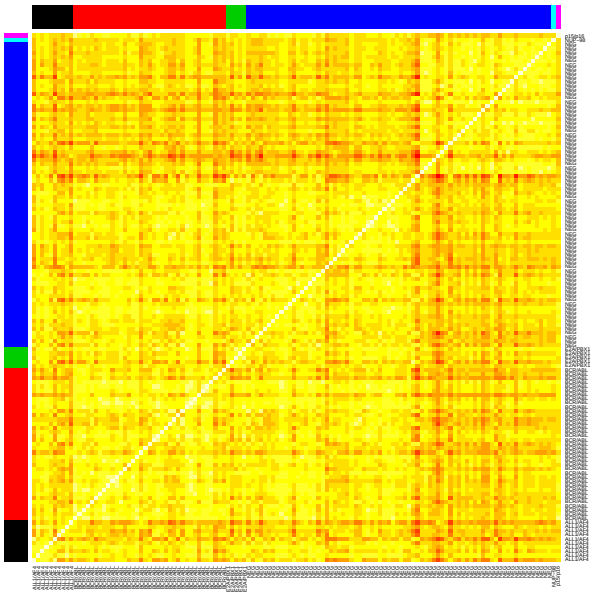

Figure 3.1: Heatmap of the correlatioin matrix of samples in the ALL package using all genes (12,625 genes) without any filtering.

In supervised machine learning of genomic data, such a simple gene-by-gene filtering is very useful. One common practice is to pick the top

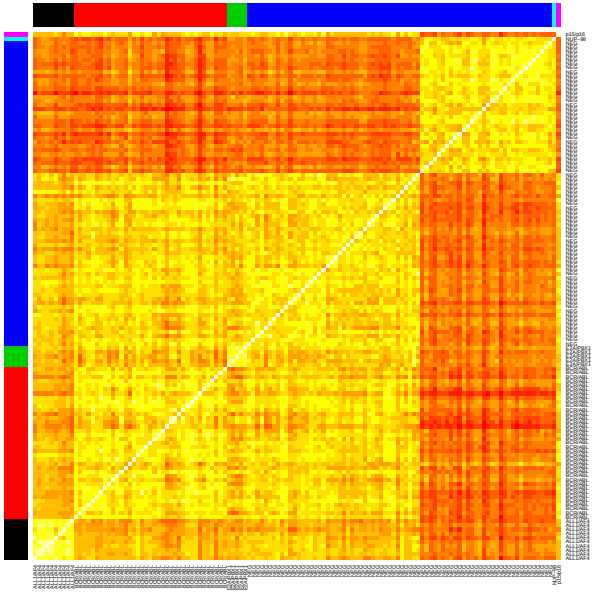

Figure 3.2: Heatmap of the correlatioin matrix of samples in the ALL package using 379 genes after filtering out weak-fluctuating genes.

N genes (features) that generate the highest absolute t-statistics (or F-statistics or certain signal-to-noise ratio) and/or the largest fold change. High absolute statistic guarantees statistical significance (i.e small within-group variation and large between-group variation) while fold change requires biological significance (e.g. more than 20% fold change). The latter requirement is sometimes important since it happens very often that a discriminant gene generate a very small p-value but has only small, say 2%, fold change. This is often not biologically interesting. The number of top genes N is a parameter. It is usually determined by cross-validation.

In addition to t-statistics, moderate t-statistics have been proposed: ?? (add more details later). It can avoid a high t-statistic value due to very small variance, which often is a result of artifact and does not contribue much generalizable classification power.

3.1.2 wrapper methods

The filtering method is computationally fast and easy to interpret. It, however, has several drawbacks. When a group of highly correlated genes with very high discriminant power exists, this set of genes may occupy the top gene list and the redundancy forces the algorithm to ignore other preditive genes. Secondly, some genes may jointly interact to provide high prediction power while individual consideration of each gene produces low prediction power. Thirdly, the filtering method is independent of the classification method used for model construction. Thus the feature selection may not choose the optimal gene set based on the special characteristics of the classification method.

– add a two-gene example figure where individual gene has small predictive power but jointly have high predictive power. Create a simulation (50 highly correlated genes of pattern 1 and 50 highly correlated of pattern 2) for exercise.

Wrapper methods consider any subset of gene selection in the algorithm. The problem is very similar to variable selection in linear regression. Popular algorithms includes forward selection, backward selection or their combinations. The Recurrsive Feature Elimination (RFE) (add citation) is a famous backward selection method applied in microarray analysis for SVM and LDA. The method starts from the full gene set. In each iteration, the gene with the smallest (or, say the bottom five smallest, to speed up the computation) estimated absolute weight in SVM or LDA is eliminated. The elimination is recurrsively performed until N genes is left. Again, the estimation of N is from cross validation.

3.1.3 embedded methods

Embeded methods jointly or simultaneously train the classifier and select the feature subset. Nearest Shrunken centroids, CART, random forest, are other tree-based classifiers belong to this category. Intuitively, embedded methods are more desirable.

3.2 Assessing and comparing classification algorithms

3.2.1 Generalizability and cross-validation

One important consideration for supervised machine learning is the "generalizability" of the classifier. In the situation that a training data set and a test (validation) data set are available, the classifier should be constructed in training data without including any information in the test data. The classifier can then be applied to the test data set to assess the classification accuracy.

In many situations, we are given only one genomic data set for supervised machine learning. A "cross validation" scheme is often used to assess an unbiased prediction accuracy. The whole samples are split into V equal portions. In each iteration, one portion of the data is left out as the test data set. The remaining $V-1$ portions are used as the training data to constructe a classifier. The classifier is then applied to the left-out test portion to assess the prediction accuracy. The procedure is repeated for V times untill all V portions are evaluated by the cross validation. Each sample is evaluated exactly once and the total prediction accuracy can be calculated. In the literature, $V = 5$ or 10 are often used. Another popular selection is when $V = S$. This is called leave-one-out cross validation (LOOCV) where only one sample is left out as test data in each iteration.

3.2.2 Performance assessment measures

2X2 contingency table (confusion matrix) TP, TN, FP, FN, sensitivity (recall rate)=TP/(TP+FN), specificity=TN/(TN+FP), positive predictive value (precision)=TP/(TP+FP), negative predictive value=TN/(TN+FN), false discovery rate (FDR)=FP/(TP+FP) (Note: The term FDR is usually used for multiple testing, not for evaluating prediction accuracy. In that case the definition of FDR is by taking expected value E[FP/(TP+FP)]).

Table 3.1: 2X2 contigency table.

	actual value (p)	actual value (n)
predicted (p)	true positive (TP)	false positive (FP)
predicted (n)	false negative (FN)	true negative (TN)

ROC curve and AUC: (1) definition of ROC curve (2) AUC is used

to summarize the sensitivity and specificity trade-off.

Genomic example: Simulate two classes of 2-D data. Use LDA to separate. Show trade-off of sensitivity and specificity.

Problem of ROC curve: (1) instability from small sample size and machine learning methods that causes difficulty to obtain an accurate estimate of AUC (2) Partial AUC may be preferred in specific situations (e.g. in population screening test, focus on region of low false positive rate)

3.3 Over-fitting and under-fitting

over-fitting or under-fitting from allowed classifier searching space.

Simulation: True model linear separation, use LDA (best); True model linear separation, use QDA (over-fitting); True model quadratic separation, use LDA (under-fitting); True model quadratic separation, use QDA (best)

overfitting from using test data in model construction
Show an exercise of gene filtering using both training and testing data.

3.4 How to choose a classifier?

There are so many machine learning methods available. Each method has its own pros and cons. Some methods have stronger data assumtptions or higher limitation on the data structure (e.g. logistic regression requires that the number of features should be smaller than sample size) that limit their use in wider applications. Some methods have found wider successful applications than the others (e.g. support vector machines and random forest). In general, the choice of the best classifier highly depends on the underlying data structure and the biological goal. There have been several comparative studies for comparing performance of different machine learning methods in microarray data. Below are three common considerations in genomic applications:

(1) accuracy focus on total accuracy, sensitivity, specificity or Youden index=sensitivity+specificity-1

(2) simplicity and interpretability of a classifier
(compare CART, LDA, ANN)

Nearest shrunken centroids (PAM)

(3) hard classification versue classification with assignment probability

3.5 Common mistakes and further notes

3.5.1 Misintepretation of accuracy

Interpretation: experimental population and empirical population

3.5.2 Use testing data information in classifier construction

Example: Below is an example that illustrates the importance of generalizability and cross-validation. It is a common mistake to perform feature selection that contains information from test data.

3.5.3 Determining better performance of a new method

Develop a new method, test on several data sets, find that its cross-validation accuracy is better than existing methods and claim that it is better.
small sample size problem $\rightarrow$ inaccurate accuracy estimation

3.5.4 Choose the minimal-error classifier among many classifiers

Note: problem of testing many classifiers and choose the best

3.5.5 Difference between DE gene detection and classification

– DE gene detection identifies "all" genes that are differentially expressed across conditions. False discovery rate is usually the concern. – Classification analysis focuses on constructing a good prediction model that is

generalizable to future samples. The prediction accuracy (sensitivity and specificity) is the concern. The analysis usually also generate key gene features that participate in the model construction. However, not all genes with good prediction power are obtained. For example, if several genes are highly correlated and all have good prediction power, it is possible that only one of them are used in the prediction model construction.

Exercise:

1. Repeat the filtering analysis of ALL data

References

Classification based upon gene expression data: bias and precision of error rates. Bioinformatics (2007) 23(11): 1363-1370
Avoiding model selection bias in small-sample genomic datasets Bioinformatics (2006) 22(10): 1245-1250
An empirical assessment of validation practices for molecular classifiers Brief Bioinform (2011) 12(3): 189-202
Pitfalls of supervised feature selection Bioinformatics (2010) 26(3): 440-443
Mistakes in validating the accuracy of a prediction classifier in high-dimensional but small-sample microarray data Stat Methods Med Res (2008) 17(6): 635-642

Comparison of Discrimination Methods for the Classification of tumors using gene expression data. Terry Speed. JASA
Genome Res. 2005 May;15(5):724-36. Classification of a large microarray data set: algorithm comparison and analysis of drug signatures. Natsoulis G, El Ghaoui L, Lanckriet GR, Tolley AM, Leroy F, Dunlea S, Eynon BP, Pearson CI, Tugendreich S, Jarnagin K.
BMC Genomics. 2008;9 Suppl 1:S13. A comparative study of different machine learning methods on microarray gene expression data. Pirooznia M, Yang JY, Yang MQ, Deng Y.
A Comparative Study of Classification Methods for Microarray. Data Analysis. Hong Hu1. Jiuyong Li1. Ashley Plank1. Hua Wang1. Grant
An extensive comparison of recent classification tools applied to microarray data. Jae Won Lee... CSDA 2005

Predictor correlation impacts machine learning algorithms: implications for genomic studies Bioinformatics (2009) 25(15): 1884-1890
Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data Bioinformatics (2005) 21(10): 2394-2402

A review of feature selection techniques in bioinformatics Bioinformatics (2007) 23(19): 2507-2517
Logistic regression for disease classification using microarray data: model selection in a large p and small n case Bioinformatics (2007) 23(15): 1945-1951
A protocol for building and evaluating predictors of disease state based on microarray data Bioinformatics (2005) 21(19): 3755-3762

Estimating classification probabilities in high-dimensional diagnostic studies Bioinformatics (2011) 27(18): 2563-2570
Probabilistic classifiers with high-dimensional data Biostatistics (2011) 12(3): 399-412
The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data J Am Med Inform Assoc (2011) 18(4): 370-375
Bayesian Model Averaging: Theoretical Developments and Practical Applications POLIT ANAL (2010) 18(2): 245-270