Classification of Gene Expression Profiles

BMI/CS 576
www.biostat.wisc.edu/bmi576/
Mark Craven
craven@biostat.wisc.edu
November 2003

Classifying Gene Expression Profiles: The Learning Task

• **given:**
  – a fixed set of classes of interest
  – expression profiles for a set of genes or experiments/individuals/time points (whatever columns represent) each labeled with its corresponding class

• **do:** induce a model that is able to predict a class label for any given expression profile (hopefully with high accuracy)
Classifying Gene Expression Profiles: The Classification Task

- **given:**
  - a model that is able to predict a class label for any given expression profile
  - expression profiles whose classes are not known

- **do:** predict a class label for each of the expression profiles (hopefully with high accuracy)

Molecular Classification of Cancer

- Golub et al., *Science* 1999
- the first published application of supervised learning methods to microarray data
- measured activity of 6817 genes in 38 leukemia patients using Affymetrix chips
- patients had one of two types of leukemia, *acute lymphoblastic leukemia* (ALL) or *acute myeloid leukemia* (AML)
Molecular Classification of Cancer

- learning task
  - given: expression profiles of leukemia patients
  - do: learn a model for distinguishing AML vs. ALL patients from expression data

- classification task
  - given: learned model, expression profile of a new patient
  - do: predict whether the patient has AML or ALL

Golub et al.’s Prediction Approach

- rank genes by their correlation with class variable (AML/ALL)
- select subset of “informative” genes
- have these genes do a weighted vote to classify a previously unseen patient
Ranking Genes

- split the expression values for the $i$th gene into two pools – one for each class
- determine the mean $\mu_i$ and standard deviation $\sigma_i$ of each pool
- rank genes by:

$$\text{weight}(i, c) = \frac{\mu_i^{\text{ALL}} - \mu_i^{\text{AML}}}{\sigma_i^{\text{ALL}} + \sigma_i^{\text{AML}}}$$

- where $c$ indicates which class the weight is associated with (ALL for +ve values, AML for −ve)

Selecting Genes

- select the $k_{\text{ALL}}$ top ranked genes (highly expressed in ALL) and the $k_{\text{AML}}$ bottom ranked genes (highly expressed in AML)

$$\text{weight}(i, c) = \frac{\mu_i^{\text{ALL}} - \mu_i^{\text{AML}}}{\sigma_i^{\text{ALL}} + \sigma_i^{\text{AML}}}$$
Informative Genes

Weighted Voting

• suppose that $x_i$ is the normalized expression level measured for the $i$th gene in a given patient

$$V = \text{weight}(i, c) \times \left( x_i - \frac{\mu_i^{\text{ALL}} + \mu_i^{\text{AML}}}{2} \right)$$

distance from the measurement to the class boundary
Weighted Voting

- the vote $V$ is assigned to a particular class depending on its sign
  
  $$V_{ALL}^+ = V \quad \text{if } V > 0$$
  $$V_{AML}^+ = |V| \quad \text{if } V < 0$$

Prediction Strengths

- can assess the “strength” of a prediction as follows:
  
  $$PS = \frac{V_{\text{winner}} - V_{\text{loser}}}{V_{\text{winner}} + V_{\text{loser}}}$$

  where $V_{\text{winner}}$ is the summed vote for the winning class,
  and $V_{\text{loser}}$ is the summed vote for the losing class
Prediction Strengths

- when classifying new cases, can pass (no-call) on cases where the strength of the prediction does not exceed a threshold

\[
\text{prediction} = \begin{cases} 
\text{ALL} & \text{if } V_{\text{ALL}} > V_{\text{AML}}, PS > \theta \\
\text{AML} & \text{if } V_{\text{AML}} > V_{\text{ALL}}, PS > \theta \\
\text{no-call otherwise} 
\end{cases}
\]

Two Experiments

- *cross validation* with the original set of patients for i=1 to 38
  - hold the ith patient aside
  - use other 37 patients to determine weights
  - with this set of weights, make prediction on the ith gene

- train on all 38 patients, test on a *separate* set of 34 patients
Golub et al. Experimental Results

- cross-validation experiments
  - all trials that used at least 3 genes had 0 prediction errors, with 1-4 no-calls

- using the 50-gene model on a test set of 34 additional patients
  - 29 correct predictions
  - 5 no-calls

Breast Cancer Outcomes Prediction

- microarray and clinical data from 86 lymph-node positive breast cancer patients
  - 12,625 genes measured using Affymetrix arrays
- goal is to distinguish between high risk (recurrence w/in 5 years) and low risk (recurrence-free for 5 years)
Calculating “Metagenes”

- the features used in their model are not mRNA measurements from individual genes
- instead they compute “metagenes”, which consist of linear combinations of gene measurements
- procedure
  - ran $k$-means clustering (with $k=500$) on original microarray data set
  - computed first principal component of each cluster
  - each of these principal components becomes a metagene

A Decision Tree Classifier

- low risk/high risk cases in training set that reach this node
- smoothed probability estimate of high risk
- outcome of test at internal node above
Decision Tree Classifiers

- tree-based classifiers partition the data using axis-parallel splits

Inducing Tree-Based Classifiers

- there are many decision-tree learning methods
- two most common are
  - C4.5 (Quinlan)
  - CART (Breiman, Friedman, Olshen, Stone)
- Nevins et al. use their own method
- all DT learning methods have the same basic algorithm structure, recursively grow a tree top-down
Generic DT Induction Pseudocode

MakeSubtree(set of instances $I$)
   if stopping criteria met
      make a leaf node $N$
      determine class label/probabilities for $N$
   else
      make an internal node $N$
      select best splitting criterion for $N$
      for each outcome $k$ of the split
         $I_k$ = subset of instances that have outcome $k$
         $k$th child of $N$ = MakeSubtree($I_k$)
   return subtree rooted at $N$

Final Comments Gene
Expression Analysis

- we discussed two computational tasks
  classification: do this when you do know the categories of interest
  clustering: do this when you don’t know the categories of interest
- class discovery is an interesting task that falls between classification and clustering
  - identify classes of profiles that don’t seem to fit into any of the modeled categories
  - e.g. new subtypes of cancer, new types of toxic substances
Final Comments Gene Expression Analysis

- there are other interesting statistical/computational tasks we didn’t cover
  - designing microarray experiments
  - reducing noise in microarray data sets
  - deciding when genes are differentially expressed across two or more conditions
  - inferring networks of interacting genes and the “programs” that govern them