to a response. The computational complexity is that of Random Forest, $O(\sqrt{M}\ N \log N)$, where M is the number of variables and N is the number of observations. This is, in fact, lighter than that of any of the benchmark methods.

References

1. L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, 1996, pp. 123–140.
2. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, 2001, pp. 5–32.
3. M. Hall, "Correlation-Based Feature Selection for Machine Learning," PhD thesis, Dept. of Computer Science, Waikato Univ., 1998.
4. I. Guyon et al., "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, nos. 1–3, 2002, pp. 389–422.
5. J.H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, tech. report, Dept. of Statistics, Stanford Univ., 1999.

Minimum Redundancy–Maximum Relevance Feature Selection

Hanchuan Peng, Chris Ding, and Fuhui Long, *Lawrence Berkeley National Laboratory*

A critical issue in pattern analysis is feature selection. Instead of using all available variables (features or attributes) in the data, one selects a subset of features to use in the discriminant system. Feature selection has numerous advantages: dimension reduction to reduce the computational cost, noise reduction to improve classification accuracy, and more interpretable features or characteristics that can help, for example, to identify and monitor target diseases or function types. These advantages are important in applications such as gene marker selection for microarray gene expression profiles[1,2] and medical image morphometry.[3] For example, selecting a small set of marker genes could be useful in discriminating between cancerous and normal tissues.

Two general approaches to feature selection exist: *filters* and *wrappers*.[4] Filter methods select features on the basis of their relevance or discriminant powers with regard to the targeted classes. Simple methods based on mutual information and statistical tests (t-test, F-test) have proven effective. In this approach, feature selection isn't correlated to any specific prediction methods. So, the selected features have better generalization properties—that is, the selected features from training data generalize well to new data.

Wrapper methods wrap feature selection around a specific prediction method; the prediction method's estimated accuracy directly judges a feature's usefulness. One can often obtain a set with a very small number of features, which gives high accuracy because the features' characteristics match well with the learning method's. Wrapper methods typically require extensive computation to search the best features.

One common practice of filter methods is to simply select the top-ranked features—say, the top 50. A deficiency of this simple approach is that these features could be correlated among themselves. For the gene-marker selection problem, if gene g is ranked high for the classification task, the filter method will likely select other genes highly correlated with g. Simply combining one very effective gene with another doesn't necessarily form a better feature set, because the feature set contains a certain redundancy. Several recent studies have addressed such redundancy.[3,5,6]

This leads to minimum redundancy–maximum relevance (mRMR) feature selection;[1,2] that is, selected features should be both minimally redundant among themselves and maximally relevant to the target classes. The emphasis is direct, explicit minimization of redundancy.

mRMR feature selection

For categorical features (variables), we use mutual information to measure the level of similarity between features. Let S denote the features subset that we're seeking and Ω the pool of all candidate features. The minimum redundancy condition is

$$\min_{S\subset\Omega} \frac{1}{|S|^2} \sum_{i,j\in S} I(f_i, f_j) \tag{1}$$

where $I(f_i,f_j)$ is mutual information between f_i and f_j, and $|S|$ is the number of features in S.

To measure features' level of discriminant power when they're differentially expressed for different targeted classes, we again use mutual information $I(c, f_i)$ between the targeted classes $c = \{c_1, \ldots, c_K\}$ (we call c the classification variable) and the feature f_i. So, $I(c, f_i)$ quantifies the relevance of f_i for the classification task. The maximum relevance condition is to maximize the total relevance of all genes in S:

$$\max_{S\subset\Omega} \frac{1}{|S|} \sum_{i\in S} I(c, f_i) \tag{2}$$

We obtain the mRMR feature set by optimizing these two conditions simultaneously, either in quotient form

$$\max_{S\in\Omega} \left\{ \sum_i I(c, f_i) \Big/ \left[\frac{1}{|S|} \sum_{i,j\in S} I(f_i, f_j) \right] \right\} \tag{3}$$

or in difference form

$$\max_{S\in\Omega} \left\{ \sum_{i\subset S} I(c, f_i) - \left[\frac{1}{|S|} \sum_{i,j\in S} I(f_i, f_j) \right] \right\} \tag{4}$$

The exact solution to mRMR requires $O(N^{|S|})$ search to obtain (N is the number of features in Ω). In practice, a near-optimal solution is sufficient, which the incremental-search algorithm obtains. The first feature is selected according to equation 3 or 4—that is, the feature with the highest $I(c, f_i)$. The rest of the features are selected incrementally. The solution can be computed efficiently in $O(|S|\cdot N)$.

For features taking continuous values, we compute quantities such as the F-statistic between features and the classification variable c as the score of maximum relevance

$$\max_{S\subset\Omega} \frac{1}{|S|} \sum_{i\in S} F(f_i, c) \tag{5}$$

and the average Pearson correlation coefficient of features as the score for minimum redundancy,

$$\min_{S\subset\Omega} \frac{1}{|S|^2} \sum_{i,j} \left| c(f_i, f_j) \right| \tag{6}$$

where we assume that both high positive and high negative correlation mean redundancy. We can also consider the distance function $d(f_i, f_j)$ (for example, L_1 distance) for the minimum redundancy condition:

$$\max_{S\subset\Omega} \frac{1}{|S|^2} \sum_{i,j\in S} d(f_i, f_j) \tag{7}$$

Mutual information formalism

As a theoretical basis of mRMR feature selection, we consider a more general feature-selection criterion, *maximum dependency* (MaxDep).[2] In this case, we select the feature set $S_m = \{f_1, f_2, \ldots, f_m\}$, of which the joint statistical distribution is maximally dependent on the distribution of the classification variable c. A convenient way to measure this statistical dependency is mutual information,

$$I(S_m; c) = \iint p(S_m, c) \log \frac{p(S_m, c)}{p(S_m)\,p(c)} dS_m dc \quad (8)$$

where $p(.)$ is the probabilistic density function. The MaxDep criterion aims to select features S_m to maximize equation 8. Unfortunately, the multivariate density $p(f_1, \ldots, f_m)$ and $p(f_1, \ldots, f_m, c)$ are difficult to estimate accurately when the number of samples is limited, the usual circumstance for many feature selection problems. However, using the standard multivariate mutual information

$$J(y_1, \cdots y_n) = \iint p(y_1, \cdots y_n) \log \frac{p(y_1, \cdots, y_n)}{p(y_1) \cdots p(y_m)} \boldsymbol{d}y_1 \cdots dy_n \quad (9)$$

we can factorize equation 8 as

$$I(S_m; c) = J(S_m, c) - J(S_m). \quad (10)$$

Equation 10 is similar to the mRMR feature selection criterion of equation 4: The second term requires that features S_m are maximally independent of each other (that is, least redundant), while the first term requires every feature to be maximally dependent on c. In other words, the two key parts of mRMR feature selection are contained in MaxDep feature selection.

Experiments on gene expression data

We've found that explicitly minimizing the redundancy term leads to dramatically better classification accuracy. For example, for the lymphoma data in figure 3a, the commonly used MaxRel features lead to 13 *leave-one-out cross-validation* errors (about 86 percent accuracy) in the best case. Selecting more than 30 mRMR features results in only one LOOCV error (or 99.0 percent accuracy). For the lung cancer data in figure 3b, mRMR features lead to approximately five LOOCV errors, while maxRel features lead to approximately 10 errors when more than 30 features are selected. We present more extension results elsewhere.[1,2] The performance of mRMR features is good, especially considering that the features are selected independently of any prediction methods.

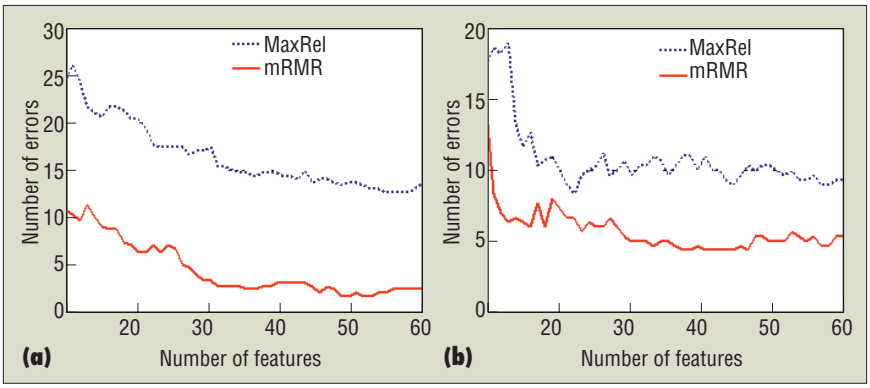

Figure 3. Average leave-one-out cross-validation errors of three different classifiers—Naïve Bayes, Support Vector Machine, and Linear Discriminant Analysis—on two multiclass data sets, lymphoma (a) and lung cancer (b), which contain microarray gene expression profiles. Lymphoma: 4,026 genes and 96 samples for nine subtypes of lymphoma; Lung cancer: 918 genes and 73 samples for seven lung cancer subtypes. More information on these data sets is available elsewhere.[1,2]

Extension

The mRMR feature-selection method is independent of class-prediction methods. One can combine it with a particular prediction method.[2] Because mRMR features offer broad coverage of the characteristic feature space, one can first use mRMR to narrow down the search space and then apply the more expensive wrapper feature-selection method at a significantly lower cost.

Acknowledgments

Chris Ding's work is partially supported by the Office of Science, US Department of Energy, under contract DE-AC03-76SF00098.

References

1. C. Ding and H.C. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data," *Proc. IEEE Computer Soc. Bioinformatics Conf.* (CSB 03), IEEE CS Press, 2003, pp. 523–528.

2. H.C. Peng, F.H. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, 2005, pp. 1226–1238.

3. E. Herskovits, H.C. Peng, and C. Davatzikos, "A Bayesian Morphometry Algorithm," *IEEE Trans. Medical Imaging*, vol. 23, no. 6, 2004, pp. 723–737.

4. R. Kohavi and G. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, vol. 97, nos. 1–2, 1997, pp. 273–324.

5. J. Jaeger, R. Sengupta, and W.L. Ruzzo, "Improved Gene Selection for Classification of Microarrays," *Proc. 8th Pacific Symp. Biocomputing* (PSB 03), World Scientific, 2003, pp. 53–64.

6. L. Yu and H. Liu, "Efficiently Handling Feature Redundancy in High-Dimensional Data," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (KDD 03), ACM Press, 2003, pp. 685–690.

Fostering Biological Relevance in Feature Selection for Microarray Data

Michael Berens, *Translational Genomics Research Institute*
Huan Liu, Lance Parsons, and Zheng Zhao, *Arizona State University*
Lei Yu, *State University of New York, Binghamton*

Microarray-based analysis techniques that query thousands of genes in a single experiment present unprecedented opportunities and challenges for data mining.[1] Gene filtering is a necessary step that removes noisy measurements and focuses further analysis on gene sets that show a strong relationship to phenotypes of interest. The problem becomes particularly challenging because of