Mutual information formalism

As a theoretical basis of mRMR feature selection, we consider a more general feature-selection criterion, *maximum dependency* (MaxDep).[1] 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,

$$(S_m; c) = \iint p(S_m, c) \log \frac{p(S_m, c)}{p(S_m)\,p(c)} dS_m a$$

(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, developed when the number of samples is limited, the usual circumstance for many feature selection problems. However, using the standard multivariate mutual information

$$y_n) = \iint p(y_1, \cdots y_n) \log \frac{p(y_1, \cdots, y_n)}{p(y_1) \cdots p(y_m)} d$$

(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 2a, 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 2b, 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.

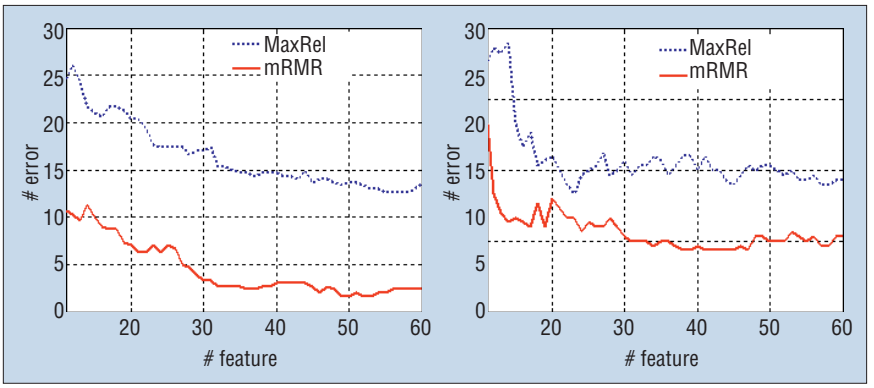

Figure 3. Average leave-one-out cross-validation errors of three different classifiers—Naïve Bayes, Support Vector Machine, and Linear Discriminant Analysis—on two multiclass data sets, lymphoma (a) and lung cancer (b), which contain microarray gene expression profiles. Lymphoma: 4,026 genes and 96 samples for 9 subtypes of lymphoma; Lung cancer: 918 genes and 73 samples for 7 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

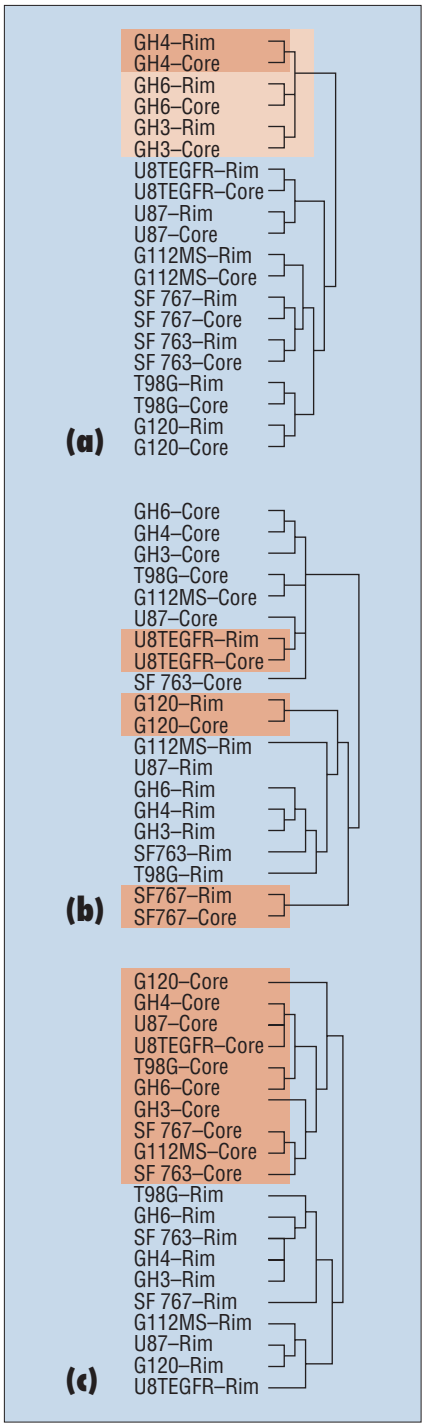

Figure 4. Hierarchical clustering of core and rim samples from 10 glioma cell lines. (a) clustering using all genes (features), (b) clustering using 22 genes selected by 2-sample t-test, and (c) after supervised feature selection, core and rim samples are clustered together, respectively.

the large number of features (approximately 30,000–40,000 genes) and the small number of samples (about 100 experiments). So, dimensionality reduction is necessary to enable effective data mining such as classification, clustering, or discriminant analysis. Feature selection, a technique that selects a subset of features from the original ones, is a frequently used preprocessing technique in data mining.[2,3]

A recent experiment on glioma cell line data reveals the importance of feature selection in microarray analysis.[4] By applying hierarchical clustering, we can visualize the discriminative power of various gene sets emerging from the two phenotypes' gene expression profiles. Figure 1a shows the dendrogram generated by hierarchical clustering based on all of the genes. Core and rim samples from the same specimen are uniformly grouped together, indicating that the core-to-rim variations are less significant than specimen-to-specimen variations. The two-sample t-test is commonly used to identify genes showing differential expression and selects 22 genes with p-values < 0.01. Figure 1b shows the dendrogram produced using these 22 genes; the clusters in red boxes still contain both core and rim samples. After the application of supervised feature selection,[5] the core-to-rim variations are far more pronounced and the samples cluster neatly into a core cluster and a rim cluster (see figure 1c). The clustering results indicate that feature selection selects discriminatory features better than statistical criteria such as a t-test do.

Beyond statistical significance in feature selection

Machine learning and statistical approaches can effectively identify both statistically relevant genes and those with redundant information. However, many statistically significant patterns found in datasets with a huge feature space and few samples might not be biologically relevant. Microarray studies' goal is often to determine which genes and pathways determine a target phenotype or clinical condition. In other words, statistically significant patterns are interesting, but it would be even better if these patterns could help identify genes with biological relevance.

A high-level goal of microarray analysis is to elucidate the developmental model of the phenotypes under study. Researchers use microarray experiments to identify genes and pathways for further study (for example, to find potential drug targets). Researchers might wish to develop diagnostic or prognostic tools, which are practical only when the number of genes is small and the classification is robust across many samples and noise levels. Suitable genes and pathways are those with not only statistical significance in the data but also certain biological or molecular traits. The additional downstream requirements necessitate the evaluation of not only microarray data but also factors such as the availability of antibodies for a given protein or the ability to interrupt a pathway with minimal harmful side effects.

The complexities of biological information can often mean that the class labels might be unreliable or too coarse, suggesting the use of unsupervised or semisupervised techniques. For example, a class label might be the histological categorization of a cancer. While those categories are quite useful, they often don't tell the entire story. Histologically similar cancers can, in fact, be molecularly distinct, with different underlying causes and clinical outcomes.

Fostering biological relevance in feature selection

We define three types of biological relevance:

- genes with known functions, which contribute to learning efficiency,
- genes with unknown functions, which present opportunities to contribute to high-impact results, and
- genes that are known to be good targets a priori (for example, genes with readily available antibodies or those suspected owing to independent evidence).

We developed a tool, Reporter-Surrogate Variable Program, which reduces the number of selected genes while increasing the overall discriminative power and helps biologists select more biologically relevant genes for subsequent biological and clinical validation.[4] Specifically, RSVP identifies a small subset of reporter genes that are mutually nonredundant and jointly provide a profile for discriminating the two phenotypes under study. In addition, for each reporter gene, RSVP identifies and presents a set of surrogate genes that are highly correlated to the reporter gene. So, biologists can replace reporter genes with genes from the surrogate lists that provide greater biological relevance without jeopardizing the overall discrimina-

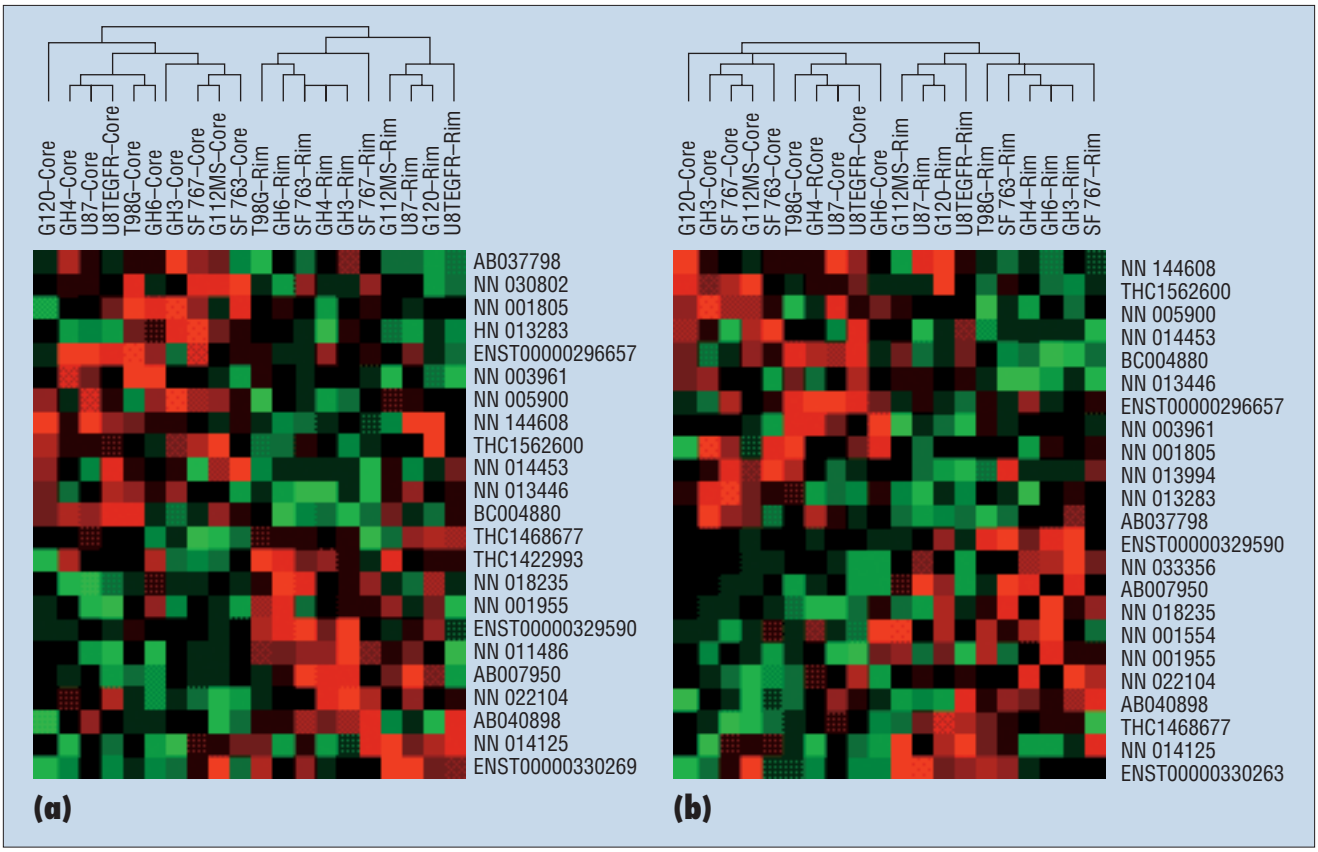

Figure 5. Hierarchical clustering results based on genes selected by the Reporter-Surrogate Variable Program tool: (a) a dendogram with an expression heatmap from the 23 reporter genes, and (b) a similar result from 20 reporter genes and three surrogate genes of biological relevance, replacing three reporter genes.

tive power. RSVP aims to produce results that are both statistically significant and biologically relevant.

RSVP identified 23 reporter genes and their corresponding surrogate genes from the 306 genes selected by the 2-sample t-test ($p < 0.1$). Figure 2a shows the clustering dendrogram and a heatmap based on the 23 reporter genes' expression values. In the heatmap, log ratios of 0 are black, and increasingly positive or negative log ratios are increasingly red or green, respectively. The 20-sample dendrogram forms two distinct clusters corresponding to the two phenotypes. Simply removing the reporter genes with unknown functions or replacing them with randomly selected genes resulted in reduced discriminative power. However, simultaneously replacing the three unknown reporter genes (NM_014486, THC1422993, NM_030802, marked by arrows) with their surrogate genes with known functions produced very similar cluster results, as figure 2b shows. Coexpression of genes in the reporter gene set and surrogate lists might also help reveal the functions of many genes for which such information is currently unavailable.

Feature selection with clinical impact

Enriching statistically significant gene lists with biologically relevant genes can help expedite biological discovery and downstream analysis. Despite public knowledge bases' increasing accessibility, the process remains largely manual, with little consistency among researchers or labs. By incorporating additional biological knowledge directly into feature selection, we can automate much of the process and improve researchers' ability to leverage the increasing amounts of publicly available research data. Interdisciplinary researchers could limit results to those targets for which antibodies are readily available to enable further study. Researchers can also more easily target drug research to particular locations in the cell. As microarray techniques advance into DNA and protein research, the number of features is increasing to millions. Sophisticated feature selection techniques that can leverage existing domain knowledge will become even more important.

References

1. G. Piatetsky-Shapiro and P. Tamayo, "Microarray Data Mining: Facing the Challenges," *SIGKDD Explorations Newsletter*, vol. 5, no. 2, 2003, pp. 1–5.

2. H. Liu and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no. 3, 2005, pp. 1–12.

3. L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *J. Machine Learning Research*, vol. 5, Oct. 2004, pp. 1205–1224.

4. L. Yu et al., *Exploiting Statistical Redundancy in Expression Microarray Data to Foster Biological Relevancy*, tech. report TR-05-005, Computer Science and Eng. Dept., Arizona State Univ., 2005.

5. L. Yu and H. Liu, "Redundancy Based Feature Selection for Microarray Data," *Proc.*

10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD 04), 2004, pp. 737–742.

Feature Selection: We've Barely Scratched the Surface

George Forman, *Hewlett-Packard Labs*

Selecting which inputs to feed into a learning algorithm is important but often underappreciated. People usually talk about "the" clusters in a data set as if there were one set of them. But if you were to cluster, for example, the vehicles in a parking lot into groups, your answer would depend completely on the features you considered: color? model? license plate? Without prior knowledge of which sorts of clusters are desired, no right or wrong choice exists. However, if someone paid you to generate a predictive model for gas mileage, you would consider vehicle weight and ignore color. These examples are meant to be obvious, but real-world data sets tend to involve large and often complex feature selection choices, whether or not they're made deliberately.

If feature selection is done poorly, no clever learning algorithm can compensate—for example, predicting gas mileage from color and trim. If done well, the computational and memory demands of both the inducer and the predictor can be reduced, and usually more important, the prediction accuracy improved. The performance of naïve Bayes—ever popular for its ease of programming—is highly sensitive to feature selection; even relatively insensitive algorithms, such as support vector machines, can benefit substantially. In some circumstances, such as biochemistry wet labs, eliminating all but the essential features can reduce the cost of obtaining measurements. Finally, feature selection by itself has useful applications, such as the statistically improbable phrases now appearing at www.amazon.com to help end-users characterize books.

While several good feature-selection techniques exist, I contend that feature selection is still in its infancy and major opportunities await. (For a survey on feature selection, refer to the 2003 special issue on variable and feature selection in the online *Journal of Machine Learning Research* (http://jmlr.csail.mit.edu/papers/special/feature03.html) or to the recent survey by Huan Liu and Lei Yu.[1])

Low-hanging fruit

A first avenue is simply to bring known successful techniques into mainstream usage. Too often an available data set is used as-is with all its features, no matter how they came to be. People generally give much more thought to the induction algorithms than to the features. Part of the solution lies in just streamlining user interfaces to make automated feature selection part of the natural process.

Of course, people don't want to be bothered with more knobs to tune. Just as you can use cross-validation to select which of several learning models performs best for a given training set, so too can it automate decisions about feature selection. (Cross-validation involves breaking a data set into, say, 10 pieces, and on each piece testing the performance of a predictor trained from the remaining 90 percent of the data. In this way, you can estimate how well each of several learning algorithms performs on the available data and then choose the best method to apply to all of the training data.) But this has its limits. Cross-validation on large data sets can exceed the user's patience budget, and cross-validation on small training sets is more likely to produce overfit models than true improvements in generalization accuracy. You can combat this with knowledge about which combinations of feature selection and learning algorithms perform well for different kinds of data. This is an open opportunity for metalearning research.

Several trends will increase the demand for feature selection. One is obviously the growing size of data sets, requiring either random subsampling of rows or purposeful feature selection of columns.

Accuracy vs. robustness

While a great deal of machine learning research seeks to improve accuracy, it sometimes comes at a cost in brittleness. To enable more widespread use of feature selection, there's a valuable vein of research in developing robust techniques. We at Hewlett-Packard have faced industrial data sets where most feature selection techniques fail spectacularly. For example, in a multiclass task for document classification where one class is very easy to predict—for example, German documents—most feature-selection methods will focus on the many strongly predictive foreign-word features for the easy class, leaving the other classes hard to distinguish.[2] Although we devised a solution for this specific type of problem, certainly more research into robust methods is necessary. I urge practitioners to share the failures they encounter on real data sets; most public benchmark data sets don't expose these issues.

Trends

I predict several trends will increase the demand for feature selection. One is obviously the growing size of data sets, requiring either random subsampling of rows or purposeful feature selection of columns. The former is easier, but the latter may be more beneficial. Feature selection might be the only reasonable choice for reducing wide data sets with many more columns than rows (for example, often more than 100,000 features in genomics or document classification).

And data sets are generally widening, with the increasing ability to link to additional databases and join with other tables. In my car example, you could link each vehicle to external databases with pollution ratings, sales figures, and review articles, potentially adding thousands of features. Today such linking requires human thought and effort, but tomorrow it could be automated.[3] This increases the pressure on automated feature selection to efficiently determine which widening is useful. The demand for this research will come primarily from practitioners who seek optimal prediction for economically valuable tasks, not from pure machine-learning researchers who care about optimizing performance on fixed, self-contained benchmark data sets for comparable, publishable results.

Rich data types

The trend toward richer data types is pushing feature selection in both scale and complexity. Natural language text features and image features are becoming common-