

# Application of Relief-F Feature Filtering Algorithm to Selecting Informative Genes for Cancer Classification using Microarray Data

Yuhang Wang  
Department of Computer Science  
Dartmouth College  
Hanover, NH 03755  
wyh@cs.dartmouth.edu

Fillia Makedon  
Department of Computer Science  
Dartmouth College  
Hanover, NH 03755  
makedon@cs.dartmouth.edu

## Abstract

Numerous recent studies have shown that microarray gene expression data is useful for cancer classification. Classification based on microarray data is very different from previous classification problems in that the number of features (genes) greatly exceeds the number of instances (tissue samples). It has been shown that selecting a small set of informative genes can lead to improved classification accuracy. It is thus important to first apply feature selection methods prior to classification. In the machine learning field, one of the most successful feature filtering algorithms is the Relief-F algorithm. In this work, we empirically evaluate its performance on three published cancer classification data sets. We use the linear SVM and the  $k$ -NN as classifiers in the experiments, and compare the performance of Relief-F with other feature filtering methods, including Information Gain, Gain Ratio, and  $\chi^2$ -statistic. Using the leave-one-out cross validation, experimental results show that the performance of Relief-F is comparable with other methods.

## 1. Introduction

Recent studies have shown that microarray gene expression data is useful for differentiating between cancerous and normal tissues [1], and among different subtypes of the same cancer [3, 2]. Cancer classification using microarray data poses a major challenge because of the following characteristics:

- The number of features (genes) greatly exceeds the number of instances (tissue samples).
- Most features (genes) are not related to the given cancer classification problem.

It has been shown that selecting a small set of informative genes can lead to improved classification accuracy [5].

The most commonly used gene selection approaches are based on gene ranking. In these gene ranking approaches, each gene is evaluated individually and assigned a score reflecting its correlation with the class according to certain criteria. Genes are then ranked by their scores and the top-ranked ones are selected.

In the machine learning field, one of the most successful individual feature filtering algorithms is the Relief-F algorithm [4]. This algorithm has been successfully used in many large subset feature selection tasks. However, to our knowledge, its performance on gene selection has not been evaluated. In this work, we empirically evaluate its performance on three published cancer classification data sets.

## 2. Relief-F

The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. Specifically, it tries to find a good estimate of the following probability to assign as the weight for each feature  $f$ .

$$w_f = \frac{P(\text{different value of } f | \text{different class}) - P(\text{different value of } f | \text{same class})}{2}$$

This approach has shown good performance in various domains [6].

## 3. Experimental Results

In this study, we used the following three published data sets: 1) ALL/AML leukemia [3] (7129 genes, 72 samples in two classes), 2) MLL leukemia [2] (12582 genes, 72 samples in three classes), and 3) Colon tumor [1] (2000 genes, 62 samples in two classes).

We use the linear Support Vector Machine (SVM) and the  $k$ -Nearest Neighbor ( $k$ -NN) as classifiers in the experiments, and compare the leave-one-out cross validation

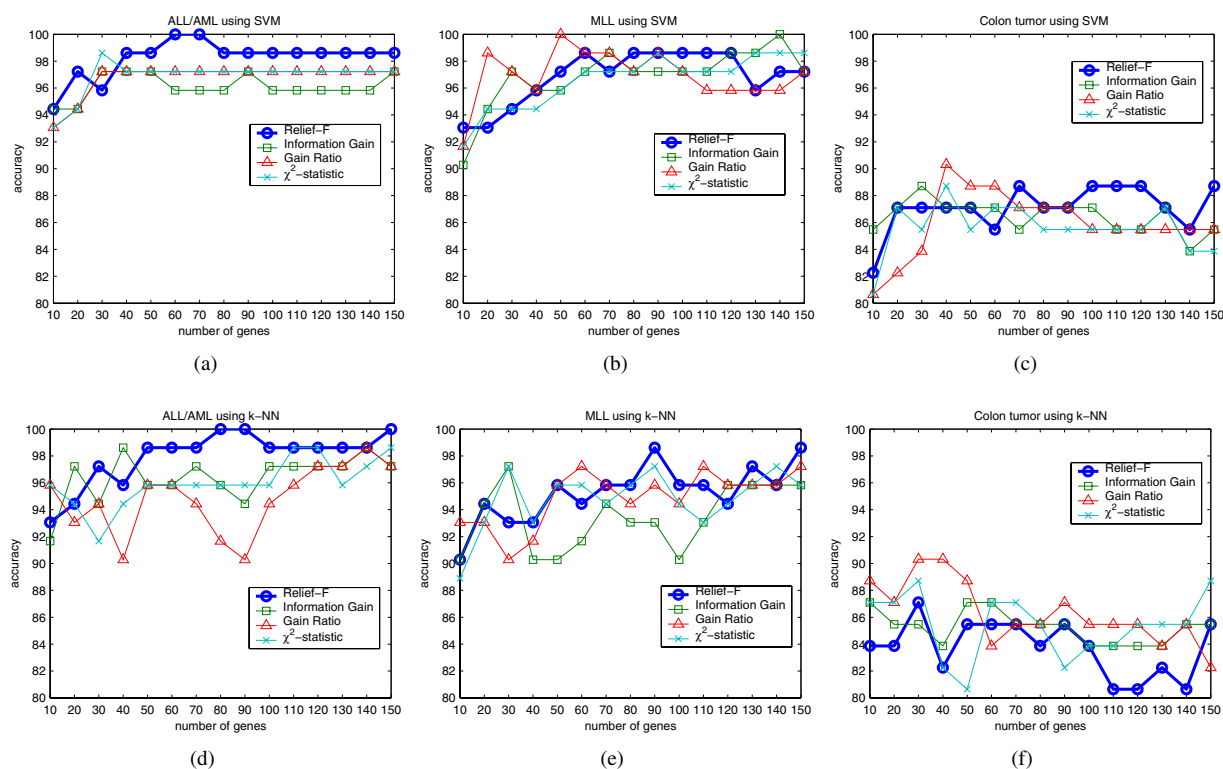


Figure 1. Comparison of the LOOCV accuracy. (a)–(c) Using linear SVM. (d)–(f) Using  $k$ -NN.

(LOOCV) accuracy of Relief-F with other feature filtering methods, including Information Gain, Gain Ratio, and  $\chi^2$ -statistic, when the top 10, 20,  $\dots$ , 150 genes are selected. When a SVM is applied to a multi-class data set, the one-versus-the-rest method is used. For the  $k$ -NN classifier, we use the Euclidean distance as the distance metric, and the best  $k$  between 1 and 10 is found by performing LOOCV on the training data.

Figure 1 shows the results. We can observe from the results that the performance of Relief-F is slightly better than other methods on the ALL/AML data set. On the other two data sets, however, the performance of different feature filtering methods is comparable. Results also show that using only the top 10 or 20 genes selected by any of the four methods doesn't lead to the best LOOCV accuracy.

### 3.1. Conclusions

This paper empirically compares the performance of the Relief-F feature filtering method with the other three methods for selecting informative genes for cancer classification using microarray gene expression data. Experimental results suggest that the performance of Relief-F is comparable with other methods.

## 4. Acknowledgments

This work was supported in part by the National Science Foundation under grants ITR-0312629 and IDM-0083423.

## References

- [1] U. Alon, N. Barkai, D. A. Notterman, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–6750, 1999.
- [2] S. A. Armstrong, J. E. Staunton, L. B. Silverman, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–47, 2002.
- [3] T. R. Golub, D. K. Slonim, P. Tamayo, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.
- [4] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Proceedings of ECML'94*, pages 171–182. Springer-Verlag New York, Inc., 1994.
- [5] Y. Lu and J. Han. Cancer classification using gene expression data. *Information Systems*, 28(4):243–268, 2003.
- [6] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.*, 53(1-2):23–69, 2003.