Using Data Mining Techniques for Improving Non-Small Cell Lung Cancer Classification

Nam-Phuong Tran[1], The University of Texas at Dallas
Faculty Advisors: Dr. Kami Makki and Dr. Quoc-Nam Tran, Lamar University

Abstract

Somatic alterations in cellular DNA underlie almost all human cancers. In recent years, microarrays have opened the possibility of creating datasets of molecular information. However, finding the responsible genes for a cancer is not an easy task because a typical microarray dataset has only a small number of records while having up to thirty thousand attributes. This kind of dataset creates a high likelihood of finding false predictions because irrelevant and redundant attributes have negative impacts on the accuracy of classification algorithms. Finding the most relevant genes is often the key phase in building an accurate classification model. In this undergraduate research paper, we investigate the use of many data mining techniques for sequential pattern analysis and similarity search techniques in DNA analysis in order to improve the accuracy of Non-Small Cell Lung Carcinoma (NSCLC) cancer classification.

1. Introduction

Approximately 1.5 million new cancer cases are expected to be diagnosed in 2010 [3]. The past decade has seen an explosive growth in biomedical research including the identification and study of the human genome by discovering large-scale sequencing patterns and gene functions [2]. As a result, new techniques are needed to analyze, manage, and discover sequence, structure, and functional patterns or models from these large sequence and structural databases [4]. Somatic alterations in cellular DNA underlie almost all human cancers [1]. In recent years, microarrays have opened the possibility of creating the data sets of molecular information to represent many systems of biological or clinical interest. However, finding the responsible gene for a cancer is not an easy task because a typical microarray dataset has only a small number of records while having up to thirty thousand of attributes. This kind of dataset creates a high likelihood of finding false predictions that are due to chance because irrelevant and redundant attributes have negative impacts on the accuracy of classification algorithms. Finding the most relevant genes is often the key phase in building an accurate classification model. Since many interesting sequential pattern analysis and similarity search techniques have been developed in data mining, data mining has become a powerful tool and contributes substantially to DNA analysis [2].

Data mining refers to extracting or "mining" knowledge from large amounts of data [2]. One form of data analysis is data classification. Data classification is the process of finding a set of models that describe and distinguish data classes for the purpose of being able to use the model to predict the class of objects whose class label is unknown [2]. Classification is a two-step process. In the first step, a model is built. In the second step, the model is used for classification. The predictive accuracy of the model is estimated. The accuracy of a model is the percentage of samples that are correctly classified by the model. If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples for which the class label is not known.

In this Research Experience for Undergraduate (REU), I experiment many different data mining techniques for diagnosing the lung cancer disease using a dataset of lung cancers from the Broad Institute [5]. I used Weka, a data mining software, to implement data mining techniques [6]. Weka is a collection of machine learning algorithms and data preprocessing tools. It provides implementations of learning algorithms that you can easily apply to your dataset. First, I used attribute evaluator algorithms to rank the attributes. I took different number of best attributes given by each algorithm to create new datasets.

[1] Supported in part by an NSF Research Experience for Undergraduate grant at Lamar University.

Classification may need to be preceded by relevance analysis, which attempts to identify irrelevant attributes that do not contribute to the classification process. These attributes can then be excluded. Many of the attributes in the data may be irrelevant or redundant to the classification task. As a result, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process. Including such attributes may otherwise slow down, and possible mislead, the learning step. I then applied classification methods to the new datasets. Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size. As a result, data preprocessing is a very important issue for data mining. Data preprocessing is an important step in data mining, since quality decisions must be based on quality data. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the classification process. One technique for data preprocessing is data reduction. Complex data analysis and mining of huge amounts of data may take a very long time, making such analysis impractical or infeasible. Data reduction can reduce the data size. Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same analytical results. One method of data reduction is discretization, where raw data for attributes are replaced by ranges. Discretization techniques can be used to reduce the number of values for a given continuous attribute by replacing actual data values with intervals.

We compare the accuracy of 17 different classifiers (figures in the Appendix give details of the diagnosing accuracy for lung cancer for each of these classifiers) on datasets that have been filtered. Three different methods have been used to select small sets of relevant genes. We also compare the accuracy of 17 different classifiers on discretized attributes.

2. Preliminaries

Decision tree induction and Bayesian classification are two well-known techniques used for data classification. A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes [2]. In order to classify an unknown sample, a path is traced from the root to a leaf node that holds the class prediction for that sample. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. Examples of decision tree induction algorithms include the ID3 and C4.5 algorithms. Both algorithms use the information gain measure to select the test attribute at each node in the tree. The attribute with the highest information gain is chosen as the test attribute for the current node. This attribute minimizes the information needed to classify the samples in the resulting partitions and reflects the least randomness or "impurity" in these partitions [2]. In the ID3 algorithm, the expected information needed to classify a given sample is given by

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

where p_i is the probability that a tuple in D belongs to class C_i, estimated by $\frac{|C_{i,D}|}{|D|}$. The entropy, or expected information based on the partitioning into subsets, is given by

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j).$$

The information that would be gained by branching on A is $Gain(A) = Info(D) - Info_A(D)$.

This algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set S. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly. The C4.5

algorithm is a successor of the ID3 algorithm. In the C4.5 algorithm, the encoding information that would be gained by branching on A is

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)},$$

where $SplitInfo(A)$ is

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|}).$$

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, the probability that a given sample belongs to a particular class. Bayesian classifiers have exhibited high accuracy and speed when applied to large databases [2]. Bayesian classification is based on Bayes theorem. Bayes theorem is

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}.$$

Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given an unknown data sample, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$P(C_i \mid X) > P(C_j \mid X)$$

for $1 \le j \le m, j \ne 1$. Thus, we maximize $P(C_i \mid X)$.

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}.$$

3. Experiment

Given a dataset of lung cancers from the Broad Institute, I used Weka, a data mining software, to implement data mining techniques [5, 6]. Weka is a collection of machine learning algorithms and data preprocessing tools [7]. It provides implementations of learning algorithms that you can easily apply to your dataset [7]. First, I applied classification methods to the dataset.

Next, I used attribute evaluator algorithms to rank the attributes. I took the twenty best attributes given by each algorithm to create new datasets. Classification may need to be preceded by relevance analysis, which attempts to identify irrelevant attributes that do not contribute to the classification process [2]. These attributes can then be excluded. Many of the attributes in the data may be irrelevant or redundant to the classification task. As a result, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process. Including such attributes may otherwise slow down, and possible mislead, the learning step [2]. I then applied classification methods to the new datasets.

Today’s real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size [2]. As a result, data preprocessing is a very important issue for data mining. Data preprocessing is an important step in data mining, since quality decisions must be based on quality data. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the classification process [2]. One technique for data preprocessing is data reduction. Complex data analysis and mining of huge amounts of data may take a very long time, making such analysis impractical or infeasible [2]. Data reduction can reduce the data size. Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same analytical results [2]. One method of data reduction is discretization, where raw data for attributes are replaced by ranges. Discretization techniques can be used to reduce the number of values for a given continuous attribute by replacing actual data values with intervals.

Next, I use discretization on the new datasets. The discretization filter in Weka is entropy-based. It is based on the ID3 and C4.5 algorithms. An information-based measure called entropy can be used to recursively partition the values of a numeric attribute [2]. For each evaluation of a candidate cut point T, the data are partitioned into two sets and the class entropy of the resulting partition is computed. Let T partition the set S of examples in the subset S_1 and S_2. Let there be k classes $C_1, \dots, C_k$ and let $P(C_i, S)$ be the proportion of examples in S that have class C_i. The class entropy of a subset S is defined as:

$$Ent(S) = -\sum_{i=1}^{k} P(C_i, S)\log_2(P(C_i, S)) \; .$$

To evaluate the resulting class entropy after a set S is partitioned into two sets S_1 and S_2, we take the weighted average of their resulting class entropies. The class information entropy of the partitioned induced by T, $E(A,T;S)$ is defined as

$$E(A,T;S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2) \, .$$

A discretization for A is determined by selecting the cut point T_A for which $E(A,T;S)$ is minimal amongst all the candidate cut points [8]. Entropy-based discretization can reduce data size. Unlike the other methods, entropy-based discretization uses class information [2]. This makes it more likely that the interval boundaries are defined to occur in places that may help improve classification accuracy [2]. I then applied classification methods to the discretized datasets.

Next, I replaced data with clusters using the EM algorithm instead of replacing data with ranges to see how it affected the accuracy of the classification models. Clustering is the process of grouping the data into clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [2]. Similarity is commonly defined in terms of how “close” the objects are in space, based on a distance function [2]. In data reduction, the cluster representations of the data are used to replace the actual data. A clustering algorithm can be applied to partition data into clusters or groups. The EM (Expectation Maximization) algorithm assigns each object to a cluster according to a weight representing the probability of membership [2]. Then, I took a reduced dataset, specifically the dataset that has been reduced using the Chi Squared attribute evaluator and applied classification methods to it. I then took another reduced dataset, specifically the dataset that has been reduced using the Filtered attribute evaluator and applied classification methods to it to see if the results are consistent.

Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier will label future data [2]. Accuracy estimates also help in the comparison of different classifiers. One technique for estimating classifier accuracy is the k-fold cross-validation method. Cross-validation is a common technique for assessing classifier accuracy, based on randomly sampled partitions of the given data. In k-fold cross-validation, the initial data are randomly partitioned into k mutually exclusive subsets or “folds,” $S_1, S_2, \dots, S_k$, each of approximately equal size. Training and testing is performed k times. In iteration i, the subset S_i is reserved as the test set, and the remaining subsets are collectively used to train the classifier. The accuracy estimate is the overall number of correct classification from the k iterations, divided by the total number of samples in the initial data.

4. Results

Figure 1 compares the accuracy of different classifiers on the original dataset with 12600 genes.

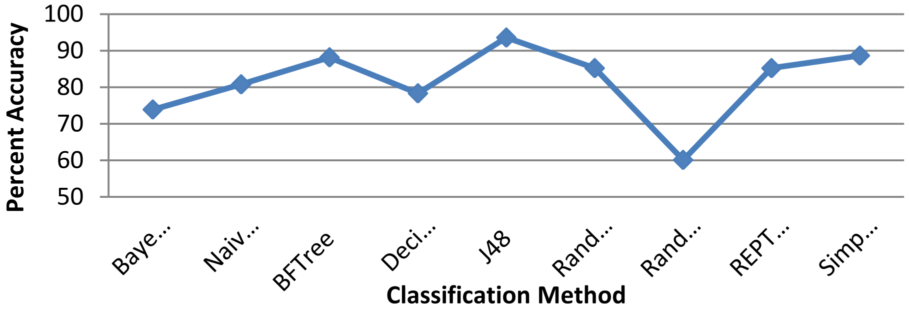

Figure 1 Accuracy of 12600 Genes

Figure 2 compares the accuracy of different classifiers on the reduced datasets. One can see that the accuracies have improved from the original dataset.

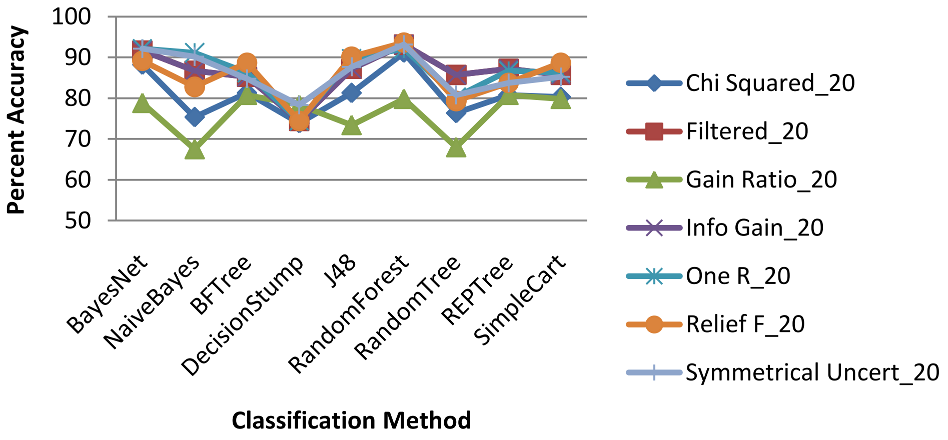

Figure 2 Accuracy of Attribute Selection Methods

Figure 3 compares the accuracy of different classifiers on the reduced datasets that have been discretized. One can see that discretization greatly improves the accuracy of classification methods.

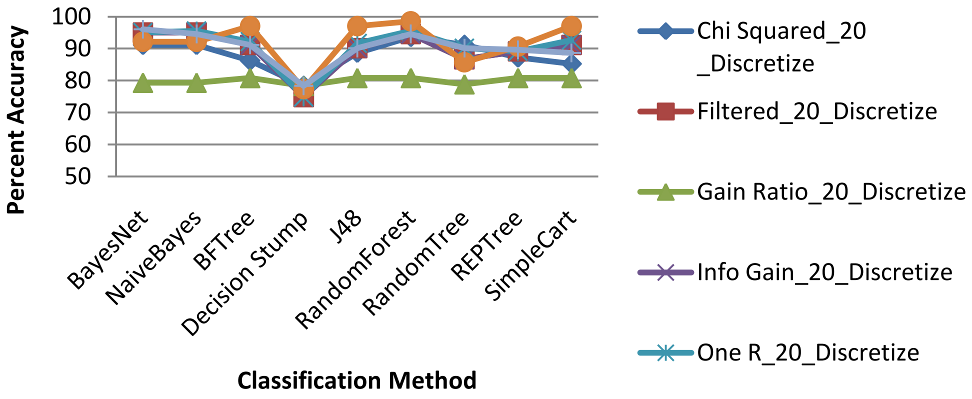

Figure 3 Accuracy of Discretization

Figure 4 compares the accuracy of different classifiers on a dataset that has been unfiltered, a dataset that has been discretized, and a dataset that has been discretized using the EM algorithm. I took a reduced dataset, specifically one that has been reduced using the Chi Squared attribute evaluator. I then applied classification methods to it. Next, I discretized the dataset and applied classification methods to the newly discretized dataset. Finally, I discretized the dataset using the EM algorithm and applied classification methods to the newly discretized dataset. Figure 4 compares the accuracy of the different datasets. One can see that though the dataset that has been discretized with the EM algorithm has greater accuracy than the original, reduced dataset, the dataset that has been discretized has the greatest accuracy.

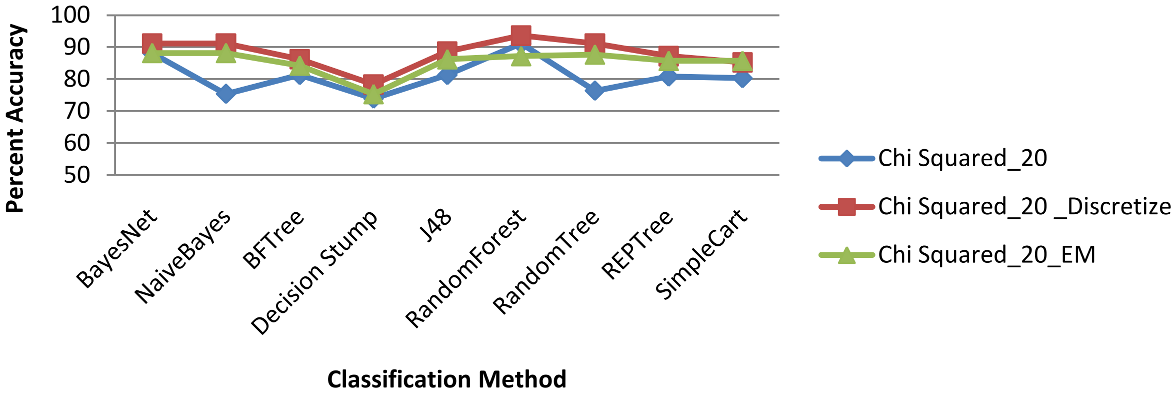

Figure 4: Accuracy of Unfiltered Dataset Vs. Discretize Filter Dataset Vs. EM Filter Dataset on the Chi Squared Dataset

Figure 5 compares the accuracy of different classifiers on a dataset that has been unfiltered, a dataset that has been discretized, and a dataset that has been discretized using the EM algorithm. I took a different reduced dataset to see if the results will be consistent, specifically one that has been reduced using the Filtered attribute evaluator. I then applied classification methods to it. Next, I discretized the dataset and applied classification methods to the newly discretized dataset. Finally, I discretized the dataset using the EM algorithm and applied classification methods to the newly discretized dataset. Figure 5 compares the accuracy of the different datasets. One can see that though the dataset that has been discretized with the EM algorithm has greater accuracy than the original, reduced dataset, the dataset that has been discretized has the greatest accuracy.

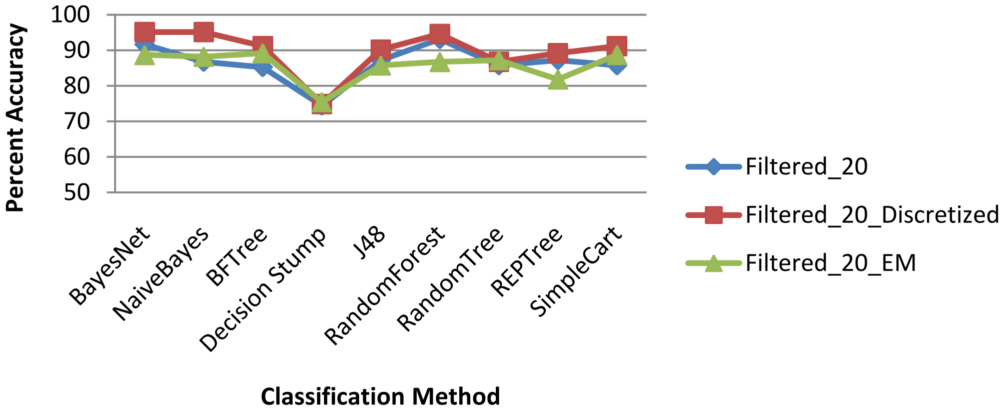

Figure 5 Accuracy of Unfiltered Dataset Vs. Discretize Filter Dataset Vs. EM Filter Dataset on the Filtered Dataset

5. Conclusion

Many of the attributes in the data may be irrelevant or redundant to the classification task. Relevance analysis, with the aim of removing any irrelevant or redundant attributes, helps improves classification accuracy and efficiency.

Data preprocessing techniques can improve the accuracy and efficiency of the classification process by improving the quality of the data. Data reduction can reduce the data size. Discretization techniques can be used to reduce the number of values for a given continuous attribute by replacing actual data values with intervals.

References

[1] B. Weir, X. Zhao, and M. Meyerson, "Somatic alterations in the human cancer genome," *Cancer Cell*, pg. 433-438, 2004.

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers, 2001.

[3] http://www.cancer.org/acs/groups/content/@nho/documents/document/acspc-024113.pdf

[4] N. Ye, Ed., *The Handbook of Data Mining*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers, 2003, pp. 573-596.

[5] http://www.broadinstitute.org/mpr/publications/projects/LUNG/DatasetA_12600gene.txt.gz

[6] http://www.cs.waikato.ac.nz/ml/weka/

[7] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers, 2005.

[8] U. M. Fayyad and K. B. Irani, "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning," *Machine Learning*, pg. 1022-1027, 1993.