Ensembling of Feature Selection Methods for HIGH DIMENSIONAL DATASET

Ruchi Gupta
M.Tech
ITMU
Gurgaon, India
ruchiguptasnp06@gmail.com

Rita Chhikara
Asst. Prof. of dept. CS/IT
ITMU
Gurgaon, India

Abstract

The feature selection is an important preprocessing step in data mining that helps in increasing the performance of the model. The aim of feature selection is to choose a subset of features with high information and to eliminate the irrelevant features with less or no predictive information. Many researches had been done to improve the performance of a single feature ranking methods, but not so much in the area of combinations of feature ranking methods. In this paper, we propose an ensemble method for feature selection, in which multiple feature selection methods are combined to yield more robust and stable results. We have done our experiment on the four feature ranking methods: ttest, entropy, Wilcoxon and Bhattacharyya methods. Through the experiment, we have shown that the combination of multiple feature ranking methods can outperform the single feature ranking method only if each individual feature selection method has high performance. The dataset we have used for our work is High Dimensional resolution Ovarian Cancer Dataset. We have used MATLAB tool in our work.

Keywords: ttest, Wilcoxon, Bhattacharyya, entropy, neural network, MATLAB.

1. Introduction

The second leading cause of death among women is ovarian cancer, as it comes directly after lung cancer. Data mining involves the use of data analysis tools to discover the previously unknown valid patterns and the relationships from large amounts of data stored in databases or other information repositories. Data mining approaches in medical domains is increasing rapidly due to the improvement effectiveness of these approaches to classification and prediction systems, especially in helping medical practitioners in their decision making. In addition to its importance in finding ways to improve patient outcomes, reduce the cost of medicine, and help in enhancing clinical studies. Data mining and machine learning depend on classification which is the most essential and important task. Many experiments are performed on medical datasets using multiple classifiers and feature selection techniques.

Feature selection is an important pre-processing step in data mining that helps in increasing the predictive performance of a model [1]. Feature selection can be categorized into feature ranking and feature subset selection. Feature ranking techniques ranks the features according to their predictive scores. The Feature subset selection groups features that can collectively have good predictive scores. Feature ranking techniques can be classified into three categories: filters, wrappers and hybrids. In this paper, we had used four filter based feature ranking techniques.

Using a single feature ranking technique may generate local optima. Ensemble approach improves the classification performance by using a combination of feature ranking techniques. The ensemble of multiple feature ranking techniques is basically performed in two steps. The first step is to create a set of different feature selection techniques, that provides its sorted order of features, and the second step aggregates the results of all feature ranking techniques.

In this paper, we studied the use of the combination of four feature ranking methods: ttest, entropy, and Wilcoxon and Bhattacharyya method. Our experimental results are obtained on high dimensional ovarian cancer dataset and shows improvement of performance on combining multiple feature selection methods.
2. Feature Selection Methods

In this section, we review four feature selection methods: ttest method, entropy method, Wilcoxon method and Bhattacharyya methods. Each of them assigns a rank to the features and generates a feature ranking list. The ranking lists contain the number of features as required by the user.

2.1 Ttest Method

The independent t-test is an inferential test designed to tell us whether we should accept or reject our null hypothesis [3]. We use an independent t-test when we want to compare the mean of one sample with the mean of another sample to see if there is a significant difference between the two samples. An independent t-test is used for the independent samples. [4]. If we carry out an experiment or collect data from two samples to see a difference between them, then we have a problem because there will almost always be some difference due to sampling.

The result of using a t-test is that we know how likely it is that the difference between our sample means is due to the sampling error. It is presented as a probability and is called p-value. The p-value tells us the probability of seeing the difference we found in two random samples if there is really no difference in the population.

Generally, if this p-value is below 0.05 (5%), we can reject the null hypothesis and thus conclude that there is a significant difference between the two population means. If we want to be particularly strict, we can decide that the p-value should be below 0.01 (1%). The level of p that we choose is called the significance level. Calculation of p-value is done by first using the t-test formula to produce a t-value. Then this t-value is converted to a probability either by software or by looking it up in a t-table.

Formula Used:
\[ z = \frac{abs(m1-m0) \sqrt{v1/n1 + v0/n0})}{1} \] (1)
Where
- \( m1 \): mean of 1 group features
- \( m0 \): mean of 2nd group features
- \( v1 \): variance of 1 group
- \( v2 \): variance of 2nd group

2.2 Entropy

The basic idea of this method is to filter out those features whose expression distributions are random. [5]. Smaller the feature’s entropy is, the more discriminatory it is. We sort the values of entropy in an ascending order and consider those features with lowest entropy values.

There exist broadly two approaches to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory. Under the first approach, the most well known measure is linear correlation coefficient. Under the first approach, the most well known measure is linear correlation coefficient.

The value of linear correlation coefficient lies between -1 and 1. If X and Y are completely correlated, r takes the value of 1 or -1; if X and Y are totally independent, linear correlation coefficient value is zero. There are several benefits of choosing linear correlation as a feature goodness measure for classification. There are several benefits of choosing linear correlation as a feature goodness measure for classification. First, it helps remove features with near zero linear correlation to the class. Second, it helps to reduce redundancy among selected features. It is known that if data is linearly separable in the original representation, then it is still linearly separable if all but one of a group of linearly dependent features were removed.

However, it is not safe to always assume linear correlation between features in the real world. It is because linear correlation measures may not be able to capture correlations that are not linear in nature. Another limitation is that the calculation requires all features contain numerical values. To overcome these shortcomings, in the solution given below we choose a correlation measure based on the information-theoretical concept of entropy which is a measure of the uncertainty of a random variable.

Formula used:
\[ z = \frac{(v1/v0+v0/v1-2)/2+(m1-m0)^2(1/v1+1/v0)/2}{1} \] (2)
Where
- \( m1 \): mean of 1 group features
- \( m0 \): mean of 2nd group features
- \( v1 \): variance of 1 group
- \( v2 \): variance of 2nd group

2.3 Wilcoxon

The Wilcoxon signed-ranks test applies to two-sample designs involving repeated measures [3], matched pairs, or "before" and "after" measures. For example, beginning with a set of paired values of \( X_a \) and \( X_b \),

- take the absolute difference \( |X_a - X_b| \) for each pair;
omit from consideration those cases where \( |X_i - X_j| = 0 \);
- rank the remaining of the differences, from smallest to largest, employing tied ranks where appropriate;
- assign to each such rank a "+" sign when \( X_i - X_j > 0 \) and a "-" sign when \( X_i - X_j < 0 \);
- And then calculate the value of \( Z \) for the Wilcoxon test, which is equal to the sum of the signed ranks. Number of signed ranks, here designated as \( n_{db} \), is equal to the number of \( X_i X_j \) pairs with which you begin minus the number of pairs for which \( |X_i - X_j| = 0 \).

General Formula:

\[
ranks = \text{tiedrank}(X) \tag{3}
\]

\[
z = \text{abs}(\text{sum}(\text{ranks(group,:)}))/n1/n0-1 \tag{4}
\]

Where

\( N1 \) = number of 1st group features
\( N0 \) = number of 2nd group features

2.4 Bhattacharyya Method

In statistics, the Bhattacharyya distance measures the similarity of two discrete or continuous probability distributions. And it is closely related to the Bhattacharyya coefficient which is a measure of the amount of overlap between two statistical samples or populations [3]. The coefficient can be used to determine the relative closeness of the two samples that are being considered. This distance measurement is used to measure the separability of the classes in classification and it is considered to be more reliable than the Mahalanobis distance. Mahalanobis distance is a case of the Bhattacharyya distance when the standard deviations of the two classes are the similar. So, when two classes have same means but different standard deviations, then the Mahalanobis distance would tend to zero. The Bhattacharyya distance would grow depending on the difference between the standard deviations.

General Formula:

\[
sd1 = \sqrt{v1} \tag{5}
\]

\[
sd0 = \sqrt{v0}
\]

\[
z = ((m1-m0)^2)/2)/sd1+sd0)/4 + log((sd1+sd0)/sqrt(s1*s0))/2
\]

3. Neural Network Classifier

A feed forward neural network is a biologically inspired classification algorithm. It consists of a (possibly large) number of simple neuron-like processing units that are organized in layers. Each and every unit in a layer is connected with all the units in the previous layer [6]. All these connections are not equal as each connection may have different weights. And the weights on these connections encode the knowledge of network. Sometimes the units in a neural network are also called nodes.

The entry of the data is at the inputs and passes through a network, layer by layer, till it arrives at the outputs. During its normal functioning, i.e. when it acts as a classifier, there is no feedback between the layers. So, this is why they are called as feed forward neural networks.

The key feature of neural networks is an iterative learning process where the data cases (rows) are presented to the network only one at a time, and then the weights associated with the input values are adjusted each time. The advantages of neural networks include its high tolerance to the noisy data, as well as their ability to classify patterns for which they have not been trained.

4. DATASET

Experiments conducted in this paper were performed on the datasets collected from the FDA-NCI Clinical Proteomics Program Databank high-resolution ovarian cancer data set [7]. This dataset consists of three files: Y, grp & MZ as shown in Appendix A. Each column in Y represents measurements taken from a patient. In file Y there are 216 columns that represent 216 patients. Out of these 216 patients, 121 are categorized as ovarian cancer patients and 95 as normal patients. Each row in Y represents the ion intensity level at a specific mass-charge value which is indicated in MZ file. 15000 mass-charge values are there in MZ and each row in Y represents the ion-intensity levels of the patients at that particular mass-charge value. The variable grp holds the index information as to which of these samples represent cancer patients and which ones represent normal patients.

5. Ensemble Feature Ranking Techniques

Ensemble of feature ranking techniques is an approach where multiple feature ranking lists obtained from corresponding feature ranking techniques are combined to generate a single ranking list [8]. Ensemble of multiple feature ranking techniques is performed to improve the classification performance. Two general steps are performed in Ensembling of feature ranking methods. The very first step is to create a set of “n” ranking lists using corresponding rankers and the second is to select the combination function i.e. the function that will transform the ranking lists obtained in
the first step into single ranking list. And the second step is the important step as it contains the method for combining the lists. Three types of combination methods are there: fusion based, selection based, and hybrid. Fusion based makes use of all the information obtained from individual rankers to produce a final outcome. Selection based methods chooses a single ranker from the list to become the final outcome. In hybrid, the final outcome is obtained after both selection and fusion methods have been used.

5.1 Proposed Work
The proposed ensemble approach is performed in two steps. It starts with creating a set of different ranking lists obtained using the rankers selected and then applies the ensemble approach to form a single feature ranking list. The first step is to select a fixed number of features from every ranking list. The ensemble approach used in this study is that for the entire feature ranking methods the accuracy classification rate is obtained first. Then we compare the accuracy rates. The ranking lists are combined with one another in the order of decreasing classification accuracy rate i.e. list 2 is attached after list 3 as the classification accuracy rate of list 3 is more than the list 2 and so on. After combining the lists into one single list we apply the classification method. And we observe that the classification accuracy have increased very much.

The algorithm steps of the proposed method can be illustrated as below:
- First step is to create i number of ranking lists.
- For each ranking list i apply the desired classification method.
- Store all the classification results in an empty list L along with the ranking method used.
- Now combine the i ranking lists in the decreasing order of the accuracy rate obtained via classification method.
- After combining the lists, apply the classification method.
- And obtain the results.

Diagrammatic representation for the proposed work is shown below:

6. Experimental Results
6.1 Experimental Result of Proposed Work
High-resolution ovarian cancer data set contains 15000 features and records of 216 patients. The ovarian cancer dataset contains three files as explained above. The experimental results on high resolution ovarian cancer dataset are taken using MATLAB. In this four feature ranking methods are applied on the dataset: T-test, Entropy, Wilcoxon and Bhattacharyya method. The classification method used for the experiment is neural network. The ranker methods were applied on the complete 15000 features to obtain top 60 and 120 features and then classification method was applied. The accuracy rate obtained using classifier neural network is given in table 5.1.

6.2 Experimental Result of Proposed Work
For our proposed work experiment we have taken top 60 i.e. 15 features each from the four feature ranking methods (15*4) and top 120 features i.e. 30 features each from the four feature ranking methods (30*4). The four feature ranking methods used are: - t-test, entropy, Wilcoxon and Bhattacharyya method. The analysis of the proposed ensemble method is given in the table 6.1.
7. CONCLUSION

In this paper, we have reviewed four filter based feature ranking techniques. They are ttest, entropy method, Wilcoxon method, Bhattacharyya method. We compared the feature ranking methods by applying the classification method neural network. We had also introduced ensemble methods for feature ranking technique that can help build stable and robust classification models.

In the table 6.1 it can be seen that when proposed ensemble approach method was used the classification accuracy rate was increased for both top 60 and 120 features as compared to the classification accuracy rate achieved from single ranking method. In the table it can be seen that for top 60 features the correct classification rate is very high for Wilcoxon method and then of ttest then entropy. Same steps for combining the lists can be followed in case of 120 features.

### TABLE 6.1

<table>
<thead>
<tr>
<th>Features</th>
<th>Ttest</th>
<th>Entropy</th>
<th>Wilcoxon</th>
<th>Bhattacharyya</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>81.39%</td>
<td>77.74%</td>
<td>87.37%</td>
<td>67.44%</td>
<td>88.37%</td>
</tr>
<tr>
<td>120</td>
<td>83.72%</td>
<td>81.39%</td>
<td>76.74%</td>
<td>81.39%</td>
<td>86.04%</td>
</tr>
</tbody>
</table>

8. REFERENCES


