Summary 7-1, 7-2, 7-3

Shaun Pritchard

Florida Atlantic University

CAP 6778

October -07-2021

M. Khoshgoftaar

Summary 7-1 - Random Forest: A Reliable Tool For Patient Response Prediction

Researchers used machine learning to predict multiple myeloma treatment responses and found that it was the most accurate prediction method. A drug called bortezomib was used in experiments using nineteen different feature selection methods using different classes and the random forest ensemble learner in Weka.

Ultimately the Random Forest learner shows that in general, we have favorable results as long as we use enough features, regardless of what strategy we use to select features. There are two approaches to selecting features, filter, and wrapper. Features are selected by filtering methods without regard to a classifier. A filter approach uses only the raw dataset to determine which features should be used to create the best classifier.

For this analysis, classes were broken down into sets. Positive classes include Complete Response, Partial Response, and Minimal Response. Other negative classes include No Change and Progressive Disease. This study performed two experiments in which these five classes were divided into a "positive" and "negative" group. Although the classes of interest did not change between the experiments, they were merged into one referred to as R. Negative classes, on the other hand, did change. The negative classes in one experiment were combined to form a single class called NR. In the other experiment, only the Progressive Disease class was used, and no changes were made to the No Change class. Random Forest learner was significantly effective in predicting bortezomib response in patients when comparing R vs NR and R vs PD.

Summary 7-2 - Feature List Aggregation Approaches for Ensemble Gene Selection on Patient Response Datasets

In the current study, feature list aggregation approaches were used to select genes based on ensemble responses from patients. Using machine learning on such complex datasets allows determining the best treatment option for the patient based on the prediction of their response to treatment.

Using high dimensionality datasets for gene expression profiles (DNA microarrays) to predict how a cancer patient will respond to treatment. Based on fifteen patient response datasets and three feature selection techniques, this study implements two feature list aggregation techniques (rank-based and score-based aggregation) with classification mean aggregation implemented through ensemble feature selection, and divided into four feature subset sizes with two classifiers.

This study was implemented using three different feature selection algorithms: Information Gain (IG), Area Under the ROC Curve (ROC), and Signal-to-Noise (S2N) with using six classification learners: 5-Nearest Neighbor (KNN), Multilayer Perceptron (P), Naive Bayes (NB), Support Vector Machines(SVM), C4.5D(J48), and C4.5N (J48). The models were tested with 5-fold cross-validation as well.

According to the study, Rank-based aggregation using nth ranks is typically used the most, and it is implemented with each new feature. Score-based aggregations begin on the feature selection step as with rank-based aggregation. The findings in this study proved that the rank-based aggregation approach outperforms the score-based aggregation approach in a majority of scenarios for both learners.

Summary 7-3 - A Novel Feature Selection Technique for Highly Imbalanced Data

A new feature selection technique is explored in this study. Additionally, a random undersampling technique (RUS) is combined with imbalance data quality assurance models to give a technique called Random Feature Selection. A study is being conducted that uses two groups of software quality data sets to indicate whether or not a particular software module contains instances that are more prone to errors.

A data set is first balanced using RUS, then subsets are selected using six Feature Ranking Techniques -chi-square (CS), information gain (IG), gain ratio (GR), and two types of ReliefF (RF and RFW). In both sets of data, three separate releases were analyzed, each with a different percentage of positive to negative classifications. A total of three learners are used, naive Bayes (NB), K-nearest neighbors (KNN),and support vector machine (SVM) learners where also used.

This study employed three scenarios: Feature ranking technique used alone (denoted NS), Data sampling followed by feature ranking (denoted nonRep), and A repetitive process of data sampling followed by feature ranking (denoted Rep). The repeatable process involved balancing the data with random undersampling (RUS), applying a feature ranking filter to the (balanced) data, and ranking all the features in order of their predictive power (scoring).

In the experimental results, results showed that data sampling improved feature selection when data sets had unequal numbers of examples in the two classes and that the repeated feature selection method outperformed other methods when the training data set was highly imbalanced.