Summary 10

Shaun Pritchard

Florida Atlantic University

CAP 6778

October -21-2021

M. Khoshgoftaar

A Review of the Stability of Feature Selection Techniques for Bioinformatics Data

This study evaluated the stability of feature selection for DNA microarray datasets with ensemble feature ranking and presented a variety of ensemble feature ranking aggregation methods.

The experimental methods used to determine the stability metrics were dataset perturbation, fixed overlap partitioning, and cross-validation procedures. These experimental procedures enable researchers to analyze and measure the stability of feature ranking methods.

According to this study, dataset perturbation refers to randomly removing instances from a dataset to create one or smaller datasets

datasets which then researchers apply feature selection to followed by creating a ranked list.

The fixed overlap partitioning generates two datasets of the same size from the original dataset with a controlled amount of overlap

When there is a slight change in the data, the overlap allows one to test the stability, and the cross-validation sampling method divides data into folds of equal data to train classifiers.

For feature selection techniques, this study explained the importance of stability but also stability measurements. The paper outlined several methods previously used by researchers, one that evaluated similarities based on the Hamming distance and others such as Kuncheva's similarity measure.

Various methods were implemented analyzing and alleviating instability as it relates to the various aspects of analyzing and devising solutions. The study found that implementing these techniques with ensemble methods yielded the best results. Due to the fact that

bioinformatics data analyzing DNA sequences require utmost precision, a technique like this is really needed.

Ensemble Feature Selection Technique for Software Quality Classification

Based on the rank order of the features of software quality engineering metrics, six filter-based feature ranking techniques and ensemble techniques were studied in this paper.

In essence, the best features were selected either through an individual ranker or through an ensemble comparison. Then the reduced data set is used to build classification models using naïve Bayes (NB), K-nearest neighbors (KNN), and support vector machine (SVM) The classification accuracy was implemented with six feature ranking techniques chi-square (CS), information gain (IG), gain ratio (GR), two forms of the ReliefF algorithm (RF and RFW), and symmetrical uncertainty (SU). Then, the AUC performance metric was applied.

Compared to a single ranker, the ensemble technique performed better in the end. Results showed that the performance of rankers fluctuated so that a certain ranker performs well for a particular data subset and classifier, but not for others when comparing independent classifiers. Despite this, the ensemble technique proved to perform better overall than any individual ranker.

The Effect of Number of Iterations on Ensemble Gene Selection

This study evaluated the effect based on the number of iterations implemented on ensemble techniques for DNA microarray datasets. Ultimately, the iterative effect creates higher computation and generates a greater number of the ranked list. The study evaluates the similarity among feature subsets generated from two different approaches. Using these two

different approaches were able to calculate the similarity between the final ranked lists generated using 10, 20, and 50 iterations using the mean aggregation function.

The study implemented two different approaches: data diversity and hybrid approaches. The data diversity method employed a single feature selection technique on multiple sampled datasets derived from the same dataset. A hybrid approach combines a set of feature selection techniques with sampled datasets derived from the same dataset.

When the additional iterations were performed, there was very little change to the Future subsets. Despite this, the study found that 20 iterations had nearly the same effect as 50 iterations. In the experiment, they used a similarity score to determine similarity.

In addition, the study showed that similarity between lists generated at different iterations increased as the size of the feature subset increased. The odds of the same gene appearing in both lists increase as the size of the feature subset increases. Additionally, the results showed that when we used data diversity, the similarity values were consistently higher than when we used hybrid diversity.