Summary 6-1 &6-2

Shaun Pritchard

Florida Atlantic University

CAP 6778

October -07-2021

M. Khoshgoftaar

**Summary_6_1 - A Comparative Evaluation of Feature Ranking Methods for High Dimensional Bioinformatics Data**

Analyses of four feature selection algorithms were implemented using bioinformatics datasets and filter-based methods. In this study, 17 high-dimensional bioinformatics datasets were analyzed to evaluate feature ranking methods to reduce feature sizes. In two of these datasets (Translation and Ovarian), microarray expression was measured. They evaluated models using three performance metrics: the area under the receiver operating characteristic (ROC) curve (AROC), the area under the precision-recall curve (APRC), and the F-measure.

In this experiment, tenfold cross-validation was implemented using the Naive Bayes classifier. Nine subsets of the data were combined to form the training dataset, and the other fold was used for training.

Out of the four feature selection filters techniques, The first three filters are the Chi-square X2, Relief-F(RF), and Information Gain(IG). As for the 4th, it is a newer feature selection filter that was recently proposed by the researchers. It is called threshold-based feature selection with AUC metric (TBFS-AUC). TBFS stands for threshold-based feature selection technique, and AUC refers to the version of this technique that is evaluated in this study.

In this new combined bivariate feature selection technique, the results were very significant and outperformed the other three filters. Additionally, X2 performed well, and both X2 and AUC performed better than the other methods. Furthermore, this research was primarily conducted to evaluate the differentiation between the two techniques and not to optimize the filters' parameters or evaluate their impact. Findings from this study showed that

developing new methods based on well-established methods lead to more successful outcomes.

**Summary_6_2 - Impact of Data Sampling on Stability of Feature Selection for Software Measurement Data**

An original research study; examined issues surrounding software fault prediction models that determine from historical, fault, and metric data whether specific software modules are prone to failure in production.

The research underlines the notion that software defect prediction accumulates issues from dimensionality and imbalance in data. Feature selection in these types of datasets is typically used for software defect prediction is a targeted concern to these issues. Never have these techniques been combined in such a way as presented in this research

In this research, they used unique experimental approaches to study how data sampling affects feature selection stability, and they evaluated the effects of various sampling techniques. The practitioners of this study examined six filter-based FS techniques and three data sampling approaches, each combined with two post-sampling proportion ratios (35:65 and 50:50).

The six Filter-Based Feature Ranking Techniques used were chi-square (CS), information gain (IG), gain ratio (GR), two types of ReliefF (RF and RFW), and symmetrical uncertainty (SU). The 3 Data Sampling Techniques used were two Random Sampling Techniques as random oversampling (ROS) and random undersampling (RUS), Synthetic Minority Oversampling Technique (SMO). Finally, the Stability Measures used were correlation coefficient, consistency index, Hamming distance, and entropy calculated between the two feature subsets.

To find feature selection stability techniques, the researchers compared ranking from each dataset to its original dataset. The results showed The results demonstrate that 1) RF and RFW performed better than other rankers in terms of their stabilities; 2) RUS35 and SMO35 produced higher stability values than other sampling approaches; and (3) the post-sampling proportion ratio between fp and nfp of 35:65 showed better stability than the other ratio of 50:50.

This study has shown that when numerous proven methods are combined issues such as stability in feature selection can be handled in with optimal results.