Summary 8-1 & 8-2

Shaun Pritchard

Florida Atlantic University

CAP 6778

October -13-2021

M. Khoshgoftaar

Summary 8-1 - The Effect of Measurement Approach and Noise Level on Gene Selection Stability

This study examines two different approaches for evaluating the stability of feature selection consisting of noise injection followed by feature ranking techniques. According to the study, stability had not been explored in real-world practice until the study was completed. It was unique in that it studied only learners trained before and after feature selection.

There are four binary datasets used in the study, including Lung cancer, ALL, Lung clean, and Ovarian cancer. Six feature rankers, Chi-Squared (CS), Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), and two types of ReliefF (RF and RFW). A 30 times cycle was then executed on each combination of four cancer-gene datasets. Followed by using the consistency index IC is used to assess the stability of noise pattern and dataset feature rankers. Lastly, compare the results of the features chosen from noisy datasets to both the clean-dataset features and the other noisy dataset features.

According to the study, noise is characteristic of errors or missing values in data from real-world environments. It consists of two types: attribute noise and class noise. When there are incorrect or missing values in the independent attributes of a dataset, attribute noise occurs, whereas class noise occurs when the label of an instance/example is incorrect.

Initially, each noisy dataset is compared with the original (clean) dataset, but then each noisy dataset is compared pairwise with each other's results.

In the study, the first approach uses noise injection, with each technique repeated 30 times. Then feature selection is performed on the original clean dataset, followed by the ranking of the attributes of the noisy datasets. According to the second approach, each ranking from the noisy datasets is compared pairwise with each ranking from the other noisy datasets as opposed to being compared individually with the original clean dataset.

Results show that the RF ranker is the most stable across all levels of noise, while the second approach is computationally more expensive and has reduced stability.

Summary 8-2 - Evaluating the Impact of Data Quality on Sampling

Data sampling is characterized by poor data quality, a large training set, and poor data sampling techniques when the data is both noisy and imbalanced. The following experiment was conducted to determine optimal sampling techniques when the data is both noisy and imbalanced.

An analysis of four training dataset characteristics such as dataset size, class distribution, noise level, and noise distribution is presented here. Using 11 learning algorithms and evaluating over 15 million models to deliver proven results.

Four datasets were used in the experiment while using four random undersampling techniques: (RUS), random oversampling (ROS), Wilson's Editing (WE), and SMOTE. Eleven different learners were also used in the experiment. Such as two nearest-neighbor techniques (2NN and 5NN), two C4.5-based techniques (C4.5D and C4.5N), Logistic Regression (LR), two types of artificial neural networks (MLP and RBFNet), Naive Bayes (NB), Random Forest (RF), RIPPER and Support Vector Machines (SVM). In the experiment, two different performance metrics were implemented: area under the ROC curve and F-measure.

In essence, the purpose of this experiment was to inject noise into the data and compare it to clean datasets for each sampling technique that was implemented with each learner and performance metric to determine significance.

Changing these data quality parameters led to the greatest performance improvements. All four data sampling techniques improved performance on noisy and imbalanced data, but RUS resulted in lower F-Measure (but usually higher AUC) when learning from imbalanced but clean data.

Wilson's editing showed it to be a safe average technique, while SMOTE and ROS improved the F-Measure when data was severe. The data also showed that SMOTE outperforms ROS no matter what the performance metric indicated.

In essence, the performance metrics used for ROC and F-measure have different effects on the results. Depending on the performance metric considered, it significantly altered the results. When the F measure is used to analyze slightly skewed or slightly noisy data, what did improve performance when severely noisy data was being analyzed? With less noise, ROC was better and with more noise, it was worse.