John Hancock

Advanced Data Mining and Machine Learning CAP 6778

September $29^{th}$, 2014

# Summary of: A Review of the Stability of Feature Selection Techniques for Bioinformatics Data

A Review of the Stability of Feature Selection Techniques for Bioinformatics Data is a survey paper reviewing attribute selection techniques what researchers have done do to find stable feature sets with bioinformatic data.

Imagine a researcher is trying to discover which genes are responsible some condition, but as he gets more data, he keeps getting different lists of genes that cause the condition. Then he is getting nowhere.

[1]mentions bioinformatic data, usually in the form of microarrays, has high dimensionality, and we typically have a low number of observations or instances.

Motivations for feature selection: eliminate irrelevant data, speed up processing time of classifiers.

We have two main classes of feature selection techniques: supervised and unsupervised. For supervised feature selection, in our data one attribute is the class attribute. We find features useful to predict the class. We have three types of feature selection techniques: filter, wrapper, and embedded. With wrapper based feature selection techniques we use a learner as part of feature selection. Filter based feature selection techniques apply statistical techniques using either one attribute, or groups of attributes at a time and gauge their

efficacy for prediction. Embedded feature selection is part of the algorithm, such as decision tree.

For unsupervised feature selection techniques we consider relationships between features.

The next section of [1]reviews methods for perturbing data. It mentions how Alelyani et. all threw the stability of some feature selection techniques into doubt because these techniques were evaluated using a high degree of overlap in subsets of data. There is also the technique of cross validation which we employ frequently in Weka. New ideas for breaking data into sets for evaluation stability include fixed overlap , and no overlap.

Next we have details on 8 stability metrics. The one we are familiar with is the consistency index, which [1]points out is due to Kuncheva.

Next in [1]we have methods for improving stability. We found two ideas to be key:

data diversity and functional diversity. We also have hybrid data diversity, which is a combination of the two.

The basic idea behind data diversity is: have original dataset, derive $n$ datasets. The method we use to create the datasets is we sample instances from original data, with replacement. For each resulting data set, we, apply one feature selection technique. We obtain $n$ lists of features, and we then apply an algorithm to aggregate the lists.

We have functional diversity method of improving stability. Here we do not fragment the dataset. We apply $n$ rankers to obtain $n$ lists of attributes. We then aggregate the lists to come up with a final list of features.

There is also a hybrid method mentioned in [1]: mix data diversity and functional diversity. We do data diversity first, then use functional diversity on each generated dataset, aggregating the resulting lists of features for a final list.

This is a survey paper, so the conclusions are a summary of previous sections; there are no results to analyze.

# References

[1] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431