John Hancock

Advanced Data Mining and Machine Learning CAP 6778

September 29^{th}, 2014

A Review of: The Effect of Number of Iterations on Ensemble Gene Selection

The goal of the paper is to determine how many iterations for feature selection we need in order for the feature selection to be stable.

Ensemble gene selection in [1]refers to using an ensemble technique, either data diversity or hybrid of data diversity and functional diversity.

Iterations in [1] refers to the number of times we apply a feature selection technique.

[1] mentions that feature selection is important when dealing with bioinformatic data because we are dealing with data that has high dimensionality. We also have a problem feature selection instability.

Related work mentioned in [1]: we know ensemble techniques are powerful for classifiers, we want to try this on feature selection, using data diversity and hybrid diversity.

The aggregation technique used in [1] is mean aggregation. [1] explains how to do mean aggregation: compute the mean rank of an attribute for all feature selection techniques used in an ensemble feature selection technique, and use this mean value as the rank of the attribute.

Figure 1 in [1] explains data diversity ensemble feature selection technique: sample data with replacement to generate n data sets, apply *same* ranker, obtain$n$ feature lists, finally aggregate the lists.

Figure 2 in [1] explains hybrid ensemble feature selection technique. It is similar to the

data diversity approach. The key difference is to use *different* rankers for each data set. After applying the rankers to the data sets, we obtain n lists of attributes, and we aggregate the results.

Ranking techniques used in [1]: The threshold based feature selection techniques used are: Mutual Information, Kolmogorov-Smirnov, Deviance, Geometric Mean, Area Under ROC, and Area Under PRC. The non-threshold based feature selection techniques are: Chi Squared, Information Gain, ReliefF, and Signal to noise.

For both kinds of ensemble feature selection - data diversity or hybrid the authors of [1] use 26 bioinformatic datasets and 80 total different feature selection techniques depending on the number of iterations. Note: 80 = 50 + 20 +10 is the different numbers of iterations used to generate bags. If we have more than 10 iterations, we use one of the 10 rankers repeatedly.

Note: paper uses 26 datasets. To get an idea of the degree of dimensionality: the numbers of attributes are all in the thousands. The lowest number of attributes is 2001, we see at least three datasets with over 50,000 attributes.

The stability measure used in [1]is the consistency index.

The stability of feature subsets is measured for different subsets ranging in size from 5 to 1000.

The results in [1] are obtained columns are taking top 5, 10, and so on features that results from feature selection using 10 vs 20 iterations, and then using the formula for stability index to compare how the sets of features are changing.

The results show that 20 iterations vs 50 produces stable results.

Tables II and III in [1] show that 10 iterations is not acceptable because there is a low consistency index score between feature subsets obtained with 10 iterations vs. 20 iterations or 10 vs.50.

Tables II and III also show that data diversity is more stable than hybrid diversity as an ensemble feature selection technique.

The conclusions in [1] explain that we find number of iterations to get stable feature set is 20 or 50. Therefore 20 or 50 iterations is the point where we can make a good trade-off between stability of features selected and more efficient use of computational resources.

Future work proposed in [1]: study how the number of iterations impacts classification results, as well as increasing the number of rankers to more than the 10 used in the current work.

References

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