College of Engineering and Computer Science

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

Wael Awada, David Dittman, Randall Wald, and Amri Napolitano, Taghi M. Khoshgoftaar

CEECS; Florida Atlantic University

Feature selection is an important step in data mining and is used in various domains including genetics, medicine, and bioinformatics. Choosing the important features (genes) is essential for the discovery of new knowledge hidden within the genetic code as well as the identification of important biomarkers. Although feature selection methods can help sort through large numbers of genes based on their relevance to the problem at hand, the results generated tend to be unstable and thus cannot be reproduced in other experiments. Relatedly, research interest in the stability of feature ranking methods has grown recently and researchers have produced experimental designs for testing the stability of feature selection, creating new metrics for measuring stability and new techniques designed to improve the stability of the feature selection process. In this paper, we will introduce the role of stability in feature selection with DNA microarray data. We list various ways of improving feature ranking stability, and discuss feature selection techniques, specifically explaining ensemble feature ranking and presenting various ensemble feature ranking aggregation methods. Finally, we discuss experimental procedures such as dataset perturbation, fixed overlap partitioning, and cross validation procedures that help researchers analyze and measure the stability of feature ranking methods. Throughout this work, we investigate current research in the field and discuss possible avenues of continuing such research efforts.

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

Wael Awada, Taghi M. Khoshgoftaar, David Dittman, Randall Wald, and Amri Napolitano

Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431

Email: waelawada@gmail.com, khoshgof@fau.edu, dittmandj@gmail.com, rwald1@fau.edu, amrifau@gmail.com

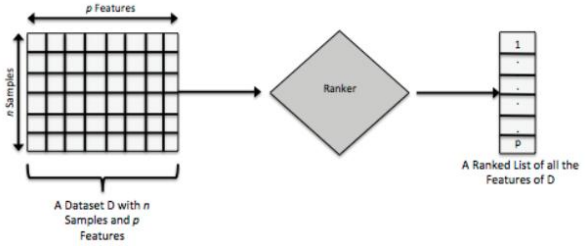
ABSTRACT

Feature selection is an important step in data mining and is used in various domains. Choosing the important features (genes) is essential for the discovery of new knowledge. Although feature selection methods can help sort through large numbers of genes based on their relevance to the problem at hand, the results generated tend to be unstable. Relatedly, research interest in the stability of feature ranking methods has grown recently and researchers have produced experimental designs for testing the stability of feature selection, creating new metrics for measuring stability and new techniques designed to improve the stability of the feature selection process. In this work, we will introduce the role of stability in feature selection with DNA microarray data.

INTRODUCTION

Two of the bigger problems associated with DNA microarrays is the high dimensionality and low sample size of the resulting datasets. High dimensionality occurs when there are a large number of features per instance of data. This combined problem causes the stability of feature rankers to decrease. However, there are a number of frameworks designed to measure or improve the stability of the feature selection process. This work is a survey of these methods.

FEATURE RANKING

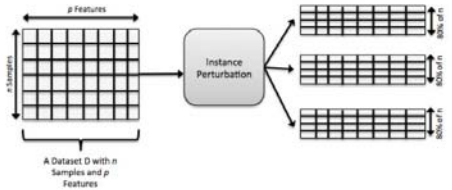


Feature ranking is the process of organizing the features on their relevance to the problem at hand based on a chosen metric. The benefits include reduced relative computational time and an intuitive output (ranked list of features).

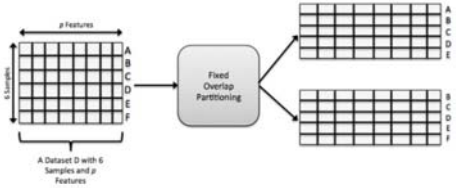
Acknowledgments

We would like to thank the IEE Research Society For the opportunity to present our work and the FAU Graduate Student Association for it's commitment to support student research.

FRAMEWORKS FOR TESTING STABILITY



Instance Perturbation randomly selects a fraction of the dataset and makes a new smaller one. This can be repeated multiple time to create any number of new datasets



Fixed Overlap Partitioning creates two new datasets of a desired level of overlap. This method allows one to account specifically for the level of overlap

METRICS FOR MEASURING STABILITY

Consistency Index

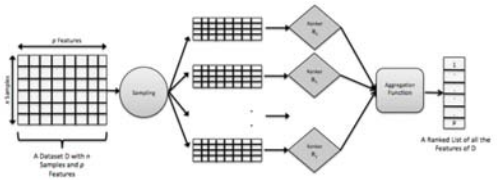
$$I_C(T_i, T_j) = \frac{dp - k^2}{k(p-k)},$$

Hamming Distance

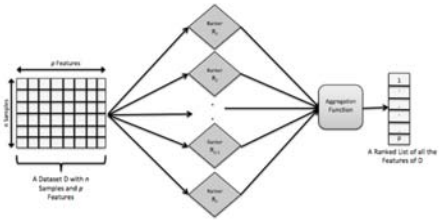
$$H(T_i, T_j) = \sum_{x=1}^{p} \left| T_i^x - T_j^x \right|,$$

In order to measure the stability of a feature selection technique, one must decide on what metric one wants to use. There are a number of different techniques each with their own strengths and weaknesses. Examples include: Consistency Index which focuses on the cardinality of the ranked feature lists and Hamming Distance which measures the distance between the feature's position in one lists vs. another list. These metrics will be applied to the pairwise combination of the feature subsets generated from the new datasets and aggregated.

ENSEMBLE FEATURE SELECTION



Data diversity uses a single feature selection technique on multiple bootstrapped or sampled datasets of the same size of the original dataset. Then the results are aggregated



Functional Diversity uses multiple feature selection technique on the original dataset and aggregates the results

Hybrid approach combines the multiple feature selection techniques from functional diversity and the bootstrapped datasets from data diversity. Each Selection technique is applied to a single dataset and the results are aggregated.

CONCLUSIONS

The stability of feature selection is extremely important in bioinformatics. Stabile feature selection will remain relevant even when changes to the data occur. Research into stability resides in two categories: measuring stability and improving the stability. Additionally there has been little work in comparing these many techniques in order to see the best performer.

References

Please refer to our publication which is referenced below:
Awada et al."A review of the stability of feature selection techniques for bioinformatics data," *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on* , vol., no., pp.356,363, 8-10 Aug. 2012
http://ieeexplore.ieee.org.ezproxy.fau.edu/stamp/stamp.jsp?tp=&arnumber=6303031&isnumber=6302564