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 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.**

***Keywords*-Stability; bioinformatics; feature selection;**

I. Introduction

Advances in genetics, chemistry, and information technology have allowed researchers to discover biomarkers ("a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." [4]) that are related to the diagnosis of a disease or its treatment. DNA contains many biomarkers, but only a small subset of these biomarkers will be related to any specific disease. Therefore one of the goals of research into these biomarkers is to identify which are relevant to the problem at hand. One of the most common ways of achieving this goal is through ordering or ranking the genes by importance.

Ordering the genes by importance is very similar to feature selection (a data preprocessing step from the domain of data mining). By providing a dataset to a feature selection technique, the feature selection algorithm returns the features (genes) that are the most important to the problem at hand. In the case of biological and genetics experiments, the problem being studied can be anything from distinguishing between healthy and diseased tissue [3], [13], [10], to identifying and accurately classifying between different types of cancer or subtypes of the same cancer [5], [12], to patient response prediction to a drug treatment [8], [26], and many more.

Two of the bigger problems associated with DNA microarrays (a recent advance in the study of genetic engineering which allows researchers to test for tens of thousands of genes simultaneously) 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. Clinical datasets usually contain few samples (instances), often less than a hundred, where each sample has been tested for the reactivity of tens of thousands of different gene probes (features). This combined problem causes the stability of feature rankers to decrease [19] and leads to generating different results after slightly changing the dataset.

Feature selection is a data preprocessing step in data mining used to reduce the dimensionality of a dataset by selecting an optimum subset of features and using only those features in subsequent analysis. Unstable feature selection leads to unstable feature subsets. The irreproducibility of a list of genes has become an issue that many researchers are trying to solve. Few research papers target the stability of feature selection, but interest in this aspect of feature selection has recently started increasing. In this paper we focus on the approaches used for analyzing, measuring, and improving the stability of feature selection techniques.

Research available so far has produced multiple types of feature selection techniques such as univariate feature ranking techniques (where each feature is studied independently of other features) and multivariate subset evaluation techniques (where every feature is studied as part of a group of multiple features).

In order to improve the stability of feature selection techniques, researchers have proposed new frameworks such as ensemble feature selection methods [27], [37], [1], [23], [40], [38], [16], which require performing feature selection multiple times and aggregating the results produced. There are three main types of ensemble feature selection techniques. The first considers data diversity [27], where the same feature selection method is used on different subsets of a dataset, and the second uses functional diversity where multiple feature selection techniques are used on the same set of data. The third is a hybrid approach combining both data diversity and functional diversity. To our knowledge, no existing research applies the latter two ensemble feature selection techniques to

bioinformatics data.

This study is a thorough survey on the current state of research into the topic of stability and its role in bioinformatics. Though this paper we will discuss topics such as: feature selection as it relates to bioinformatics, experimental procedures for the testing and measuring of stability of feature selection, causes of instability, and methods of improving the stability of feature subsets. Additionally, for each of these areas we present analysis and future areas of research to pursue.

The remainder of this paper is presented as follows. Section II contains some existing feature ranking methods. Section III outlines methods of measuring the stability of feature ranking techniques. Section IV will discuss the causes of feature selection instability. In section V, different frameworks that have been proposed to target the problem of instability are explained. Section VI presents our conclusions from our study.

II. Feature Selection Techniques

Feature selection is a commonly used tool used to reduce the effects of high dimensionality on a dataset. Specifically, feature selection seeks to reduce the number of features by targeting an optimum subset of features and removing the rest of the features from further consideration during subsequent analysis. The idea is that the features that are not being considered are either irrelevant to the problem at hand or redundant when compared to the features within the optimum subset. As the optimum subset is much smaller than the entire feature set, the computation time of subsequent analysis is greatly reduced. What is interesting is that despite the loss of data caused by the removal of a large number of features, the learners being built using the reduced subset perform on a similar level to those built using the entire feature set and in some cases perform better. A study performed by Inza et al. [18] found that classification performed on reduced feature subsets derived from the original DNA microarray datasets outperformed classification using the whole feature set in a majority of cases and they drastically reduced computation time.

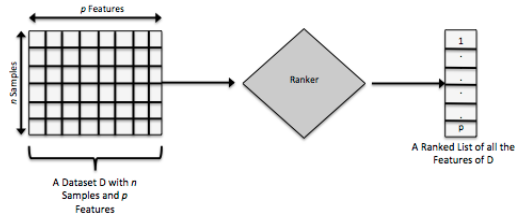

Fig. 1. Feature Ranking

At the highest level, feature selection techniques are either supervised or unsupervised. Supervised feature selection uses the class labels for deciding on the best features, while unsupervised feature selection considers only the relationships among the features without using the class values at all. Each has their advantages and disadvantages: supervised feature selection can be used to find features which are useful for predicting the class (either alone or in groups), while unsupervised feature selection can be used to better understand gene networks. In this paper we focus on supervised feature selection, because many bioinformatics datasets (such as DNA microarray datasets) have class values which are useful to predict.

Supervised feature selection techniques can be separated into three different categories: filter, wrapper, and embedded. Filter-based feature selection techniques seek to select the optimum feature subset through the use of statistical metrics. There are two main classes of filter-based feature selection: feature ranking and subset evaluation. Feature ranking (Figure 1) [28], [36], sometimes known as univariate feature selection, take each feature separately and tests it for its ability to distinguish between the classes. Subset evaluation [28] looks at the possible subsets of features and tests them for their ability to differentiate between the classes. It should be noted that in order to test each possible subset, one would have to perform $2^n - 1$ tests (there is no point is testing an empty set) where n is the number of features per instance. Although search algorithms can reduce this space significantly, they do not resolve the problem of necessitating many evaluations, and introduce the problem of finding local optima which are not the globally-preferred solutions. As one can see, when the number of features are large (tens of thousands in the case of DNA microarrays) this can become computationally expensive so as to be completely inappropriate for the problem at hand [28].

The second category is wrapper-based feature selection techniques which relies on building classification models to determine the importance of features. Typically wrapper-based feature selection is performed on subsets (much like filter-based subset evaluation), although it may be applied as a ranker. One example of this approach can be found in Abeel et al. [1], Davis et al. [7] and Slakov et al. [30]. Wrapper methods differ from filter-based feature selection in that they use a learner when evaluating the features, either separately or as subsets. Unfortunately, the building of a learner takes time and would have to be repeated for every test [15]. This aspect of wrapper-based feature selection can make the technique very computationally expensive, especially for subset evaluation. For example, if one wanted to choose the best pair of features from a dataset of 15,154 features, it would require evaluating 114,814,281 classifiers. If it takes 0.1 seconds to develop each classifier it will take 132.9 days of continuous computation to evaluate them all [31]. Again, different search techniques can improve this, but introduce their own problems.

Finally, the third category of feature selection methods is embedded techniques. Embedded feature selection techniques are found within the learners themselves. This means that when one runs a learner with embedded feature selection, the learner performs feature selection prior to analysis.

Due to the large degree of high dimensionality, the filter subset evaluators and wrapper based methods can be too computationally expensive for the use with DNA microarray datasets. This has lead to the majority of studies using filter-based feature ranking for DNA microarrays and bioinformat-

ics [28], [33]. Some commonly used methods for feature rankers are chi-squared [36], information gain [14], gain ratio [36], ReliefF [21], symmetric uncertainty [15], and SVM-RFE [36], [34].

As there is no single method which is the best for all domains and problems, future work in the area of feature selection stability should focus more on methods of improving the stability of the process of feature selection rather than devising a new technique. Methods that can improve stability in general can be used for any method in order to improve reproducibility of the feature subsets.

III. Aspects of Measuring Stability

Measuring stability requires two aspects: a testing procedure and a stability measurement . The experimental design describes how one can go about testing for the stability of a feature selection technique. The stability measurement is the specific metric for measuring stability. First, we will look at the methods for testing stability.

A. Experimental Design for Testing Stability

To measure the stability of feature selection methods, many researchers test their results by using dataset perturbation. Dataset perturbation (Figure 2) consists of randomly removing instances from a dataset in order to create one or more reduced datasets. Researchers then apply the feature selection method in question to the each of the datasets (all of the reduced datasets and sometimes the original dataset) and create a ranked list for each of the datasets. The following step is to measure the stability between the ranked lists [27], [6]. Researchers can compare the results of applying the feature selection method between the perturbed datasets or on the perturbed datasets versus the results of applying the same feature selection method on the original dataset [9].

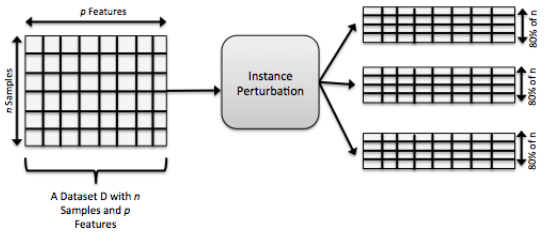

Fig. 2. Dataset Perturbation

Alelyani et al. explained in their research [2] that the stability of a feature ranker cannot be evaluated without looking into the variance between the samples of a dataset. Data variance shows the spread of different values in a given dataset. The authors relate feature selection stability to the variance in the data and show their findings by conducting an experiment with four datasets and five selection methods. They study the impact of the variance in the datasets by controlling the difference between the training samples by perturbing the original dataset and creating a number of new reduced datasets which, when compared to each other, have an overlap which is varied from 10% to 90%. Alelyani et al. then apply five selection methods on the training sets and compare the resulting rankings. The results presented show that stability is higher when the amount of overlap among the subsamples is higher. The research performed by Alelyani et al. provides a preliminary study of the role the variance in the datasets plays in terms of the stability of feature selection. This work can be extended by evaluating the stability of the feature selection techniques used in their experimental procedures when there are no samples in common between the training datasets. Studying the effect of overlap on stability is better achieved by varying the amount of overlap, including when there is no overlap at all. In addition, due to their method for producing reduced datasets, all of the subsamples have no more than 25% of the instances from the original dataset. When experimenting on domains with limited numbers of instances (such as bioinformatics datasets), discarding the majority is not an option.

Another method used to test and measure the stability of feature selection methods is through the cross validation method [33]. Cross validation (Figure 3) is when a dataset is split into multiple folds or partitions of equal size (or as close as possible). The first $n-1$ folds are used for training the learner, where n is the number of folds, and the remaining fold is used to test the learner. This process is performed n times so each of the folds will be used as the test fold. Cross validation is usually used when testing for classification accuracy. For the purpose of measuring stability, researchers apply the feature selection method in question on the $n-1$ folds (or training datasets) n times.

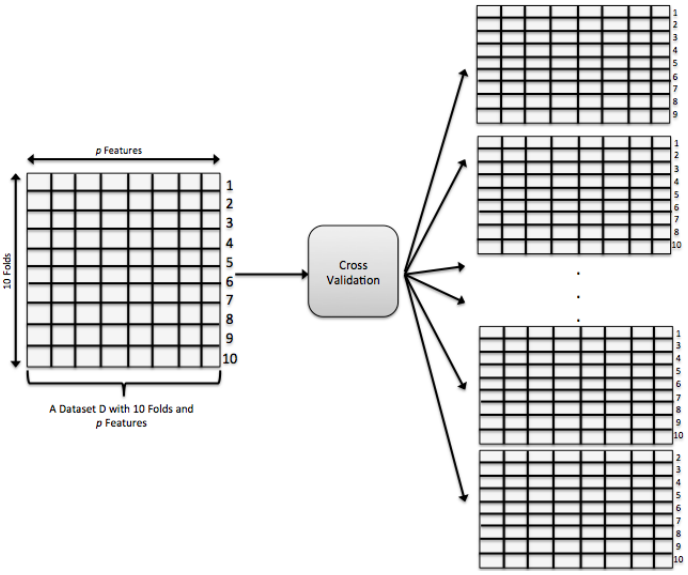

Fig. 3. Cross Validation

Wang et al. [35] propose a new method called fixed overlap partitioning which allows generating two datasets of the same size from the original dataset with a controlled amount of overlap between them. This method allows one to test the stability when there is a slight change in the data (high overlap percentage), or when the data is completely different (no overlap at all). Few researchers have controlled the overlap

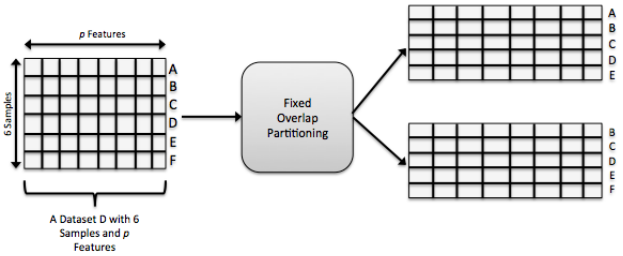

Fig. 4. Fixed Overlap Partitioning with Overlap

in their experiments before, but we have not found any other research that proposes this method in a clear way.

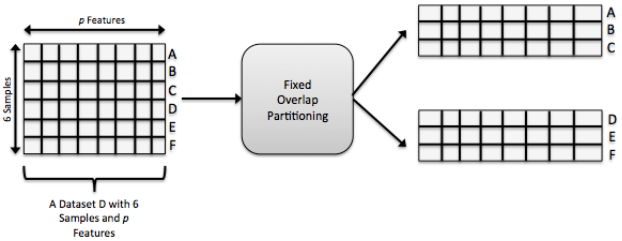

Fig. 5. Fixed Overlap Partitioning with No Overlap

In order to test the stability across datasets, Haury et. al. [16] perform three different tests on four datasets. They subsample two datasets that have 80% of the samples in the original datasets in common and then split the remaining samples between them. The authors repeat the sampling 20 times maintaining the 88.88% (80% common samples out of 90% sampled of the original dataset, i.e. (80/90)%) overlap each time and then compute the overlap between the ranked gene lists generated by the selection methods. They also calculate stability by repeating the steps above except that no overlap between the two subsampled datasets exists (e.g., each of the two subsamples contains 50% of the original data, with no instances present in both subsamples).

Haury et. al suggest that the best way to evaluate the stability of a feature ranker is through applying it to datasets with no overlap (the second of the two approaches discussed above). Their methods of evaluating stability represent a preliminary work where they tested their results on four datasets. Although their work provides a very helpful insight into stability measurement with high and no overlap, future research should consider using more than four datasets. Moreover, future work should consider studying the trend of the stability of feature selection methods as the overlap increases by varying the amount of overlap at higher degrees. (i.e.: 0%, 20%, 40%, 60%, and 80%).

While work on the specifics of these experimental design approaches and more is still ongoing, the question remains on which of these methods best measures the stability of feature selection. Additionally, there has been very little work outside of the original papers on the optimum parameters to use for these methods (level of overlap for dataset perturbation and data partitioning, and number of folds for the cross validation method).

B. Feature Selection Stability Metric

While implementing an experimental design with the goal of testing stability is important, equally as important is the stability measurement one chooses. In recent years there have been a number of different stability measurements implemented for this exact purpose.

Lustgarten et al. [25] present a number of different feature selection stability measures that apply to feature ranking as well, in addition to proposing their own measure inspired by their research. One measure they examined (originally developed by Kalousis et. al [19]) works as follows:

$$S_R\left(T_i, T_j\right) = 1 - 6 \sum_x \frac{\left(T_i^x - T_j^x\right)^2}{k\left(k^2 - 1\right)}.$$

where T_i^x and T_j^x are the rank of feature x in rankings T_i and T_j, respectively, and k is the number of features being selected. For a set U of subsets of features obtained by applying, $|U|$ times (one run per subset), a particular feature ranking technique, the overall stability is computed as follows:

$$S_t = \frac{2}{|U|\left(|U| - 1\right)} \sum_{i=1}^{|U|-1} \sum_{j=i+1}^{|U|} S_R\left(T_i, T_j\right)$$

Another approach examined is that developed by Dunne et al. [11], which evaluates similarity based on the Hamming distance:

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

where p is the total number of features in the dataset. Thus, given a set U of subsets of features generated by $|U|$ runs of a particular feature selection method, the total Hamming distance, H_t, is computed as follows:

$$H_t = \sum_{i=1}^{|U|-1} \sum_{j=i+1}^{|U|} H\left(T_i, T_j\right).$$

The overall stability is then defined by the average Normalized Hamming Distance obtained as follows:

$$\hat{H} = \frac{2 \times H_t}{p \times |U| \times \left(|U| - 1\right)}$$

The above measure is called an unadjusted stability measure since it doesn't take into consideration the role chance might play in selecting common features between two lists.

Kuncheva [22] developed an enhanced similarity measure, called consistency index, that takes chance into consideration. Let T_i and T_j be subsets of features, where $|T_i| = |T_j| = k$. The consistency index is obtained as follows:

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

where d is the cardinality of the intersection between subsets T_i and T_j, and $-1 < I_C\left(T_i, T_j\right) \leq +1$. The greater the consistency index, the more similar the subsets are. Kuncheva

et al. extended their consistency index to expand beyond the comparison of just two subsets by taking the average across all pairwise consistency indices. This extension can be used in a number of different ways including comparing the original dataset to a set of perturbed datasets derived from the original [9].

To improve Kuncheva's similarity measure, Lustgarten et al. [25] propose the similarity measure:

$$S_a\left(T_i, T_j\right) = \frac{d - \frac{k_i k_j}{p}}{min\left(k_i, k_j\right) - max\left(0, k_i + k_j - p\right)}$$

where k_i and k_j are the cardinalities of T_i and T_j respectively (in most cases, $k_i = k_j = k$). Note that S_a varies from -1 to 1, where a value of 0 represents the stability of random feature selection, positive values indicate particularly stable feature selection, and negative values represent stability lower than that of random feature selection. A final measure was devised which combines the results of multiple measures into the new stability index called adjusted stability measure (ASM), that takes role of chance into consideration and that can calculate the stability of lists of varying size. The ASM for c subsets is calculated as:

$$ASM = \frac{2}{c\left(c-1\right)} \sum_{i-1}^{c-1} \sum_{j=i+1}^{c} S_a\left(T_i, T_j\right)$$

Lustgarten et al. present their results on a single proteomic dataset using three feature selection techniques. They show that their adjusted measure presents improved evaluation of the stability of the feature selection methods.

The next step for research in this area is further comparison and evaluation of these metrics. With many to choose from, it is not obvious which are best-suited to a given problem, and what theoretical assumptions must be satisfied to ensure that each is appropriate. Because empirical evaluation doesn't necessarily make sense in the context of comparing stability metrics, this comparison must be in terms of the mathematical and statistical properties of each metric.

IV. Causes of Feature Selection Instability

Feature selection stability, especially in the biomarker discovery domain, is important for researchers looking to further validate their findings by applying the discovered biomarkers on new data. One of the biggest obstacles facing feature selection is the stability of the algorithms used. If the biomarkers chosen for the newer data are not the same as the discovered biomarkers, then the original set of biomarkers are not valid. Therefore, the stability of feature selection techniques is crucial for reliable results. In recent years stability has become a topic of interest in the fields of genetics, molecular biology, and bioinformatics [32], [29].

There are a number of reasons that could lead to reduced stability of a feature selection method. One possible cause is designing an algorithm to select the minimum number of features with the highest classification accuracy without taking stability into consideration when designing such an algorithm [39]. This minimalist view can miss important information within the redundant features. Another cause for instability is the possibility of the existence of multiple "true" markers (markers which are highly correlated with the data) in a dataset [17]. The dataset may contain various markers that are highly correlated which might lead the feature selection algorithm to select different features on different dataset samples. Alternatively, the features might be uncorrelated, but there might be multiple true markers in the dataset. High dimensionality combined with low sample size is perhaps one of the largest contributors to feature selection instability [20]. This case applies frequently in the biomedical field since DNA data can contain thousands of features but very few samples [24]. Instability could also be caused by the variance in the data [2].

V. Methods for Improving Stability

Devising methods used to improve the stability of feature selection has been a heavily researched topic of late. A large number of these methods are members of one of two categories: group feature selection and ensemble feature selection.

A. Group Feature Selection

The first category is group feature selection [39], [23], where highly correlated features are assembled into separate groups, and feature selection is performed on these resulting groups. The idea behind group feature selection is that because the groups are made of highly correlated features then they will have the same relevance in their relationship between the class labels. Group feature selection consists of two steps: group formation and feature selection. Group formation is the step where different groups of correlated features are identified by using two different methods: knowledge driven methods and data driven methods. Knowledge driven methods depend on domain knowledge, which can help in dividing the correlated features into groups. Data driven methods, on the other hand, take only the original dataset into consideration without looking at domain related information. The second step, which is feature selection, takes the resulting groups formed, and creates a compact feature subset by only selecting one feature from each group. By choosing only one feature per group the problem of redundancy has been handled. Additionally, the stability of the feature subset is increased because in addition to the compact subset, all of the features in the dataset are grouped by correlation rather than removed due to redundancy. The feature groups can give vital information about how the features relate to one another which is missed by methods which seek to remove redundancy.

B. Ensemble Feature Selection

The second approach to maximize the robustness of feature selection is to create an ensemble of feature rankers [27], [37], [1], [23], [38]. Ensemble feature selection is a subset of feature selection techniques which applies feature selection algorithms multiple times and combines the results into one decision. Because multiple results are combined, the features which are

frequently the best performers will move towards the top of the list, while those with sporadic or poor performance will move lower down; thus, the final feature list will be more stable. The idea for ensemble feature selection is derived from ensemble learning methods wherein different classifiers are applied to a dataset and their results are aggregated. The procedure for ensemble feature selection is very similar. There are three main ways to apply ensemble feature selection: through data diversity, functional diversity, and hybrid ensemble techniques. Data diversity (Figure 6) consists of applying a single feature selection method to a number of differently sampled versions of the same dataset and then an aggregation technique is used to aggregate the results [27]. Functional diversity (Figure 7) is performed by applying a set of different feature selection techniques on the same dataset. Hybrid ensembles use both of these, applying different feature selection techniques to different sampled versions. It is not always known which of these methods is superior for any given situation. This is due to the fact that existing studies only examine data diversity and do not explore the abilities of functional diversity or compare the results of the two methods. Additionally, to our knowledge, there has been no work published which combines both data and functional diversity into one hybrid method.

After multiple ranked lists are created using one of the above three methods, the second step of creating an ensemble feature ranker is to use one of many available aggregation functions to aggregate the results that are generated in the first step. Some examples of aggregation methods include exponential aggregation, mean and median aggregation, and threshold based aggregation [16]. However, each of these aggregation methods has their strengths and their weaknesses. Choosing the most appropriate aggregation technique can be a daunting task especially as there is, to our knowledge, no study found which thoroughly compares the various methods, especially in regards to stability.

One study examining ensemble feature selection is Saeys et al. [27], which uses ensemble feature selection where each feature's final rank is the sum of its rank in the different lists. In their experiment, the authors apply four different selection methods on 10 subsamples of each of the six different datasets where each subsample contains 90% of the data available in the original dataset. Out of the four methods used, we are specifically interested in the stability of Symmetric Uncertainty and ReliefF (SVMRFE and Random Forest are wrapper feature selection methods and thus are out of the scope of this study). The results show that ReliefF and Symmetric Uncertainty were weaker than the wrapper methods used but an ensemble version of each of the four selection methods has improved stability when compared with the corresponding individual selection methods. This research provides an insight into the role an ensemble feature selection method plays in improving the stability of feature selection.

VI. Conclusion

The stability of the selection of genes is extremely important to the field of bioinformatics. To this end, researchers in the field of data mining have recently been concentrating on various aspects of the problem of instability. This paper is an in-depth survey of the different aspects of analyzing and devising methods of alleviating the problem of instability.

The research into the stability of feature selection can be split into two categories: testing and measuring stability, and devising methods to improve the stability of feature selection. In terms of testing and measuring stability, there has been a lot of work in experimental design for determining the stability of a feature selection technique as well as creating measurements which allows us to quantify the level of stability. While there has been a large amount of research on the creation of these methods, only preliminary studies (and in some cases no studies) have focused on how the various methods and metrics work when compared to the other methods and metrics. For example, no research exists comparing the relative merits of cross-validation, bootstrapping, and fixed overlap partitioning for evaluating stability (and in fact, no research considers fixed overlap partitioning at all). Additionally, to our knowledge there has been no study on which measurement technique works best when applied to each method of testing.

When it comes to improving stability, one of the more popular ideas is using an ensemble method. This topic of research has a large amount of potential for future work. While most studies which use ensemble techniques choose only data diversity techniques, to our knowledge there has been no study which compares this with functional diversity or explores the concept of creating a hybrid technique which combines functional and data diversity. Additionally there has been little work on what method of aggregation is the most appropriate in any given situation. In conclusion, the area of stability in feature selection is an excellent and diverse area for research and is valid for bioinformatics experiments as well as those of other domains.

References

[1] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, February 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19942583

[2] S. Alelyani, Z. Zhao, and H. Liu, "A dilemma in assessing stability of feature selection algorithms," in *High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on*, Sept. 2011, pp. 701 –707.

[3] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999. [Online]. Available: http://www.pnas.org/content/96/12/6745.abstract

[4] A. J. Atkinson, W. A. Colburn, V. G. DeGruttola, D. L. DeMets, G. J. Downing, D. F. Hoth, J. A. Oates, C. C. Peck, R. T. Schooley, B. A. Spilker, J. Woodcock, and S. L. Zeger, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework*," *Clin Pharmacol Ther*, vol. 69, no. 3, pp. 89–95, Mar 2001. [Online]. Available: http://dx.doi.org/10.1067/mcp.2001.113989

[5] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 559–583, 2000. [Online]. Available: http://www.liebertonline.com/doi/abs/10.1089/106652700750050943

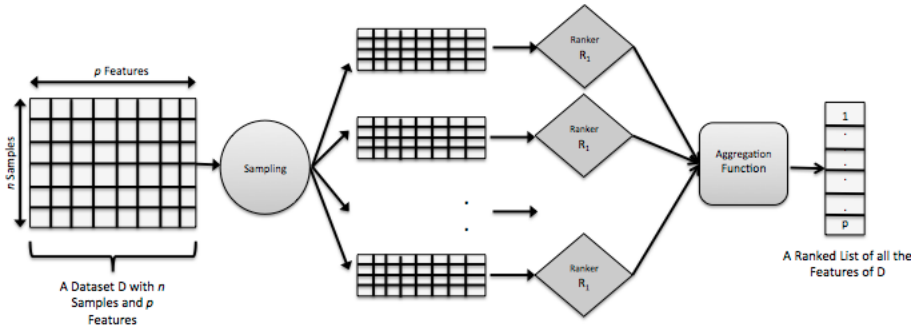

Fig. 6. Ensemble: Data Diversity

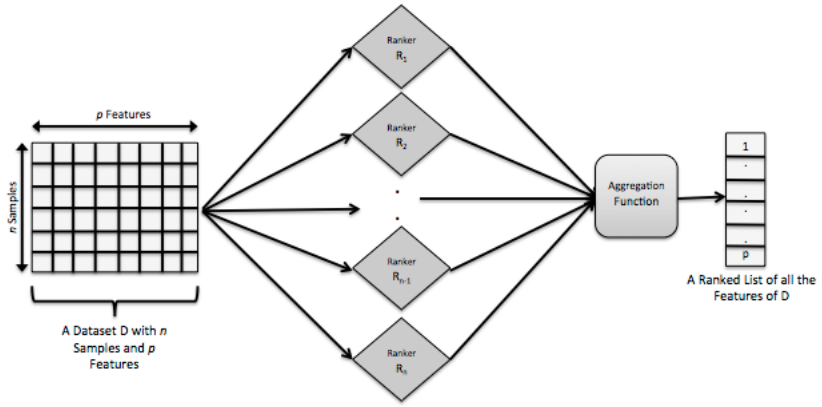

Fig. 7. Ensemble: Functional Diversity

[6] A.-L. Boulesteix and M. Slawski, "Stability and aggregation of ranked gene lists," *Briefings in Bioinformatics*, vol. 10, no. 5, pp. 556–568, 2009. [Online]. Available: http://bib.oxfordjournals.org/content/10/5/556.abstract

[7] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kffner, and R. Zimmer, "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics*, vol. 22, no. 19, pp. 2356–2363, 2006. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/22/19/2356.abstract

[8] D. Dittman, T. Khoshgoftaar, R. Wald, and A. Napolitano, "Random forest: A reliable tool for patient response prediction," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Workshops*. BIBM, 2011, pp. 289–296.

[9] D. Dittman, T. Khoshgoftaar, R. Wald, and H. Wang, "Stability analysis of feature ranking techniques on biological datasets," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. BIBM, 2011, pp. 252–256.

[10] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002. [Online]. Available: http://www.jstor.org/stable/3085760

[11] K. Dunne, P. Cunningham, and F. Azuaje, "Solutions to Instability Problems with Sequential Wrapper-Based Approaches To Feature Selection," Department of Computer Science, Trinity College, Dublin, Ireland, Tech. Rep. TCD-CS-2002-28, 2002.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, 1999. [Online]. Available: http://www.sciencemag.org/content/286/5439/531.abstract

[13] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002, 10.1023/A:1012487302797. [Online]. Available: http://dx.doi.org/10.1023/A:1012487302797

[14] M. A. Hall and G. Holmes, "Benchmarking attribute selection techniques for discrete class data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 6, pp. 392–398, November/December 2003.

[15] M. A. Hall and L. A. Smith, "Feature selection for machine learning: Comparing a correlation-based filter approach to the wrapper," in *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, May 1999, pp. 235–239.

[16] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PLoS ONE*, vol. 6, no. 12, p. e28210, 12 2011. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0028210

[17] Z. He and W. Yu, "Stable feature selection for biomarker discovery," *Computational Biology and Chemistry*, vol. 34, no. 4, pp. 215–225, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1476927110000502

[18] I. Inza, P. Larraaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in dna microarray domains," *Artificial Intelligence in Medicine*, vol. 31, no. 2, pp. 91–103, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365704000193

[19] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: a study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, no. 1, pp. 95–116, Dec. 2006. [Online]. Available: http://www.springerlink.com/index/10.1007/s10115-006-0040-8

[20] S.-Y. Kim, "Effects of sample size on robustness and prediction accuracy of a prognostic gene signature," *BMC Bioinformatics*, vol. 10, no. 1, p. 147, 2009. [Online]. Available: http://www.biomedcentral.com/1471-2105/10/147

[21] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," *Lecture Notes in Computer Science*, pp. 171–182, 1994.

[22] L. I. Kuncheva, "A stability index for feature selection," in *Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications*. Anaheim,

CA, USA: ACTA Press, 2007, pp. 390–395. [Online]. Available: http://portal.acm.org/citation.cfm?id=1295303.1295370

[23] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 81–87, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046409001117

[24] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2009, pp. 567–576.

[25] J. L. Lustgarten, V. Gopalakrishnan, and S. Visweswaran, "Measuring stability of feature selection in biomedical datasets," in *AMIA 2009 Symposium Proceedings*, 2009, pp. 406–410.

[26] G. Mulligan, C. Mitsiades, B. Bryant, F. Zhan, W. J. Chng, S. Roels, E. Koenig, A. Fergus, Y. Huang, P. Richardson, W. L. Trepicchio, A. Broyl, P. Sonneveld, J. Shaughnessy, John D., P. Leif Bergsagel, D. Schenkein, D.-L. Esseltine, A. Boral, and K. C. Anderson, "Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib," *Blood*, pp. 3177–3188, 2007.

[27] Y. Saeys, T. Abeel, and Y. Peer, "Robust feature selection using ensemble feature selection techniques," in *ECML PKDD '08: Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 313–325.

[28] Y. Saeys, I. Inza, and P. Larraaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/23/19/2507.abstract

[29] S. Schneckener, N. Arden, and A. Schuppert, "Quantifying stability in gene list ranking across microarray derived clinical biomarkers," *BMC Medical Genomics*, vol. 4, no. 1, p. 73, 2011. [Online]. Available: http://www.biomedcentral.com/1755-8794/4/73

[30] I. Slavkov, B. Senko, and S. Dzeroski, "Evaluation method for feature rankings and their aggregations for biomarker discovery," in *Journal of Machine Learning Research - Proceedings Track*, 2010, pp. 122–135. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.6967&rep=rep1&type=pdf#page=127

[31] R. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–1491, 2003. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/19/12/1484.abstract

[32] G. Stiglic and P. Kokol, "Stability of ranked gene lists in large microarray analysis studies," *Journal of Biomedicine and Biotechnology*, vol. 2010, p. 9, 2010. [Online]. Available: http://www.hindawi.com.ezproxy.fau.edu/journals/jbb/2010/616358/

[33] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature selection with high dimentional imbalanced data," in *Proceedings of the 9th IEEE International Conference on Data Mining - Workshops (ICDM'09)*. Miami, FL: IEEE Computer Society, December 2009, pp. 507–514.

[34] H. Wang, T. M. Khoshgoftaar, and A. Napolitano, "An empirical study of software metrics selection using support vector machine," in *Proceedings of International Conference on Software Engineering and Knowledge Engineering SEKE'11*, July 7-9, 2011, pp. 83–88.

[35] H. Wang, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "A novel dataset-similarity-aware approach for evaluating stability of software metric selection techniques," in *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2012)*, August 2012.

[36] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical machine learning tools and techniques*, 3rd ed. Burlington, MA: Morgan Kaufmann, January 2011.

[37] P. Yang, J. Ho, Y. Yang, and B. Zhou, "Gene-gene interaction filtering with ensemble of filters," *BMC Bioinformatics*, vol. 12, no. Suppl 1, p. S10, 2011. [Online]. Available: http://www.biomedcentral.com/1471-2105/12/S1/S10

[38] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010. [Online]. Available: http://www.ingentaconnect.com/content/ben/cbio/2010/00000005/00000004/art00006

[39] L. Yu, C. Ding, and S. Loscalzo, "Stable feature selection via dense feature groups," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '08. New York, NY, USA: ACM, 2008, pp. 803–811. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401986

[40] L. Yu, Y. Han, and M. E. Berens, "Stable gene selection from microarray data via sample weighting," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 9, no. 1, pp. 262–272, Jan. 2012. [Online]. Available: http://dx.doi.org.ezproxy.fau.edu/10.1109/TCBB.2011.47