A Survey of Stability Analysis of Feature Subset Selection Techniques

Taghi M. Khoshgoftaar*, Alireza Fazelpour*, Huanjing Wang†, and Randall Wald*
*Florida Atlantic University
†Western Kentucky University
Email: khoshgof@fau.edu, afazelpo@my.fau.edu, huanjing.wang@wku.edu, rwald1@fau.edu

***Abstract*—With the proliferation of high-dimensional datasets across many application domains in recent years, feature selection has become an important data mining task due to its capability to improve both performance and computational efficiencies. The chosen feature subset is important not only due to its ability to improve classification performance, but also because in some domains, knowing the most important features is an end unto itself. In this latter case, one important property of a feature selection method is stability, which refers to insensitivity (robustness) of the selected features to small changes in the training dataset. In this survey paper, we discuss the problem of stability, its importance, and various stability measures used to evaluate feature subsets. We place special focus on the problem of stability as it applies to subset evaluation approaches (whether they are selected through filter-based subset techniques or wrapper-based subset selection techniques) as opposed to feature ranker stability, as subset evaluation stability leads to challenges which have been the subject of less research. We also discuss one domain of particular importance where subset evaluation (and the stability thereof) shows particular importance, but which has previously had relatively little attention for subset-based feature selection: Big Data which originates from bioinformatics.**

***Keywords*-Feature selection; subset evaluation; stability; stability measure; similarity measure;**

I. Introduction

One major challenge facing researchers working with Big Data is high dimensionality, which occurs when a dataset has a large number of features (independent attributes). To resolve this problem, researchers often utilize a feature selection process to identify and remove features which are irrelevant (not useful in classifying the data) and redundant (provide the same information as other features). Selecting a subset of features that are most relevant to the class attribute is a necessary step to create a more meaningful and usable model.

Feature selection has been applied in many application domains, specifically those known for high dimensionality such as software metrics analysis [33], text mining [8], gene expression microarray analysis [9], image analysis [29], web mining [31], and intrusion detection [24]. The main goal of feature selection is to identify irrelevant or redundant features and select a subset of features which minimizes the prediction errors of classifiers. Reducing the number of features in a dataset can lead to simple and faster model training, along with improved classifier performance. Another benefit of feature selection is to gain a better understanding of the model built with a selected subset of features [10].

Feature selection techniques are divided into two broad categories: wrapper-based and filter-based. The former category utilizes a supervised learning algorithm in the process of selecting feature subsets. The downside of a wrapper-based technique is its high cost of computation and the risk of an overfitted model. On the other hand, filter-based feature selection utilizes the intrinsic characteristics of the data to evaluate the attributes for inclusion in the subset of features. The filter method has advantages over the wrapper since it is faster computationally and it selects features using the data characteristics rather than an external learner. Both of these groups can again be divided, into either ranker-based or subset evaluation-based, depending on whether they examine features individually or in groups. Rankers are much more computationally efficient, but are unable to identify redundant features, while subset evaluation is the opposite in both of these properties.

One common criterion to evaluate feature selection methods is the performance of a chosen classifier trained with the selected feature subset. More recently, some studies have considered another criterion for evaluating feature selection methods, namely, stability. The concept of stability has been extensively studied in the context of inductive learning systems [17]. The stability (e.g., robustness) of feature selection methods has been used to examine the sensitivity of these methods to changes in their input data. Stability is normally defined as the degree of agreement between its outputs to randomly selected subsets of the same input data [20], [22]. The need for consistent outputs from feature selection methods is very important in application domains where the prime concern is knowledge discovery [7]. Those methods whose outputs are insensitive to changes in the input data are said to be stable, and they are preferred over those methods that produce inconsistent outputs.

To measure the stability of a feature subset selection technique, a similarity measure is needed to assess the overlap of a pair of feature subsets. The similarity measure is used to compute the stability of a feature subset selection technique as the average similarity over all pairs of feature subsets. The similarity measures discussed in this survey paper are the Hamming distance [5], [7], the Jaccard index [27] (and the closely-related Tanimoto distance [6]), Spearman's rank correlation coefficient [17], the consistency index [20], and Shannon entropy [19]. One challenge when evaluating stability is the question of whether the metric makes sense when the feature selection technique can potentially produce feature subsets of varying sizes (as is the case for filter-based subset evaluation and wrapper-based subset selection techniques). All of the discussed measures other than consistency index, Spearman's rank correlation coefficient, and Shannon entropy are able to properly calculate stability even in this case.

The remainder of this paper is presented as follows: Section II presents an overview of feature selection techniques, including filter-based, wrapper-based, and embedded approaches. Section III outlines methods of measuring the stability of feature subset selection techniques. In Section IV, we take a closer look at one particular area of feature selection stability research which yet remains a challenge: the stability of subset evaluation for big bioinformatics data. Finally, Section V presents our ideas for future work, while Section VI covers our conclusions.

IEEE IRI 2013, August 14-16, 2013, San Francisco, California, USA
978-1-4799-1050-2/13/$31.00 ©2013 IEEE

II. Feature Selection Techniques

Due to the increased prevalence of high-dimensional datasets across many application domains, feature selection has become an essential preprocessing step in many data mining tasks, in addition to being important on its own for knowledge discovery. Saeys et al. [28] listed three main objectives for a feature selection technique: (1) to improve the performance of a classification model and avoid overfitting, (2) to improve the efficiency of the model in terms of cost and time, and (3) to provide deeper insight into the underlying characteristics of the dataset.

Feature selection techniques can be organized into three categories, depending on how they utilize various feature selection search methods when constructing the classification model: (1) filter-based methods, (2) wrapper-based methods, and (3) embedded methods. Due to space limitations, we briefly describe all three feature selection techniques in this survey paper below, as well as the different search techniques used with subset evaluation approaches, and we ask interested readers to consult Saeys et al. [28], Liu and Yu [21], and Guyon and Elisseeff [10] for more information. The first reference provides an excellent taxonomy of feature selection techniques with pros, cons, and examples of each method. The second provides a three-dimensional framework based on search strategies to identify subsets of features, data-mining tasks (classification vs. clustering), and evaluation criteria to assess the goodness of the selected feature subset. The third outlines key approaches used for attribute selection, including feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

A. Filter-based Feature Selection

Filter-based feature selection techniques are divided into two subcategories: (1) filter-based feature ranking and (2) filter-based feature subset evaluation. Filter-based feature selection utilizes the intrinsic characteristics of the data using some kind of statistical criterion to evaluate the merit of (1) each individual feature in the case of filter-based feature ranking (known as ranker or univariate techniques) or (2) the entire feature subsets in the case of filter-based subset evaluation (known as subset-evaluation or multivariate techniques). The benefits of the multivariate techniques are as follows: they model feature dependencies, they are independent of the classifier, and they are more efficient than wrapper methods in terms of computational complexity [28]. Their drawbacks are as follows: they are slower than univariate (ranker) techniques; they are less stable than univariate techniques; and they ignore interaction with the classifier, so they are outperformed by the wrapper technique in general [28].

Bolón-Canedo et al. [2] and Peteoro-Barral et al. [26] conducted some experiments to investigate the stability of three filter-based subset evaluation methods. These techniques are as follows: (1) Correlation-based Feature Selection (CFS), (2) the Consistency-based Filter [3], and (3) the INTERACT algorithm [36]. The CFS is a multivariate algorithm that evaluates feature subsets according to a correlation based heuristic function [12]. Theoretically, redundant features and irrelevant features should be ignored. The Consistency-based Filter [3] evaluates each subset of features by the level of consistency in the class values for the training dataset. The INTERACT algorithm [36] is a feature subset method based on symmetrical uncertainty (SU) that handles feature interaction and selects relevant features.

B. Wrapper-based and Embedded Feature Selection

Wrapper-based feature selection techniques utilize a supervised learning algorithm in the process of selecting feature subsets. As with filter-based techniques, wrappers can be ranker or subset evaluation techniques, although the latter is far more common, with the exception of Support Vector Machine-based feature ranking [11] which uses the individual feature weights assigned during the creation of the Support Vector Machine learner to rank the features. With wrapper-based subset evaluation, however, features are evaluated in context, i.e., dependencies and correlations between features are considered. This is useful for discovering which features are either redundant or correlated with each other. On the other hand, searches through such a large space of dimensionality to find the optimal set of features are infeasible. The benefits of the wrapper-based subset evaluation techniques are as follows: they model feature dependencies; and they interact with the classifier, so they have the potential to outperform rankers. Their drawbacks are as follows: they are computationally expensive and probably prohibitive for datasets with high-dimensionality, and they have higher risk of overfitting [28].

Due to their complexity, few papers have compared wrapper-based subset evaluation techniques with filter-based techniques. Kohavi and John [18] compared models built with wrapper feature selection to those using one form of filter-based feature ranking (as well as a model built with no feature selection), and found that wrapper-based subset selection was able to remove noisy and redundant features and help improve the performance of classification models. Hall and Holmes [13] investigated six attribute selection techniques that produce ranked lists of attributes and applied them to several data sets from the UCI machine learning repository, comparing these with a filter-based approach. They found no single best approach for all situations, but overall, wrappers were the best attribute selection schema in terms of accuracy if the speed of execution was not considered. Inza et al. [16] compared six filter-based feature rankers (four which operate on discrete features, two which operate on continuous features) to a wrapper-based approach on gene microarray data, and found that wrapper-based gene selection outperformed filter-based rankers for most learners (aside from naïve Bayes), while selecting far fewer features. This additional power came at the cost of increased computational load, however.

Embedded techniques integrate the search for an optimal feature subset into the process of building a classifier. Like wrapper techniques, embedded methods are specific to a given classification algorithm. The benefits of the embedded techniques are as follows: they model feature dependencies; they interact with the classifier; and in terms of computational complexity, they are better than wrapper methods. Their drawbacks are as follows: they are computationally more expensive than filter-based subset evaluation techniques, and they are classifier dependent [28].

C. Search Techniques

Dunne et al. [7] stated that the most common search strategies to generate subsets of features are based on stepwise addition or deletion of features. A sequential selection proceeds by adding or removing features from the current set to form a new set. Then, this new set is evaluated using some validation procedure such as cross-validation and if the new set of features is superior, then it replaces the current best set and this process continues until no more valid operations (addition or deletion) can be carried out or if no new candidate set outperforms the best current one.

Forward sequential selection (FSS) begins with an empty set of features (genes) and searches for the best feature to be added at each

step. If all the features are added or there is no improvement from adding any further features, the search stops and returns the current set as the optimal set. The FSS has a maximum search length of N iterations (N is the total number of features present in the training dataset). The objective of the search is to add only the relevant features to the current optimal set while ignoring the irrelevant and redundant features.

Backward sequential selection (BSS) starts with a full set including all features and searches for a feature to be removed at each step. The resulting set is evaluated using some validation procedure such as cross-validation and if the new set of features is optimal, then it replaces the current best set; this process continues until reaching an empty set or the subsequent removal of any feature degrades the current performance and the search ends. The objective of the BSS is to consider contribution of all feature first and tries to remove the most irrelevant or redundant feature leaving a smaller and more optimal set. The BSS is more computationally expensive than FSS because the BSS starts with all of the features at the start of the search process. However, the BSS outperforms the FSS technique in terms of classification performance due to the fact that it initially includes all the features in the set.

A third type of search strategy is called hill-climbing that either adds or removes one feature at a time. The search starts with a random set of features and then attempts the effect of toggling the current status of each feature in the set (i.e., removing an existing feature from the set or adding a feature that is not currently in the set; then evaluate the new set and choose the optimum set). The stopping criteria for the search process is when we reach the limit on the number of iterations that we set and return the last optimum set of features.

III. Stability of Feature Selection Techniques

The stability of feature subset selection techniques is defined as insensitivity of the result of a feature selection method to minor variations in the training dataset. This issue is crucial in many applications where feature selection is used as a knowledge discovery tool to identify the top features relevant to the phenomena of interest [25].

Generally, the classification performance is considered the ultimate quality measure not only for assessing classification models, but even when evaluating the feature selection algorithms [30]. The performance of a classifier based on particular sets of selected features does not necessary imply that these selected sets are robust; thus, the corresponding feature selection technique is stable. For example, in biomarker identification (knowledge discovery application), a feature selection algorithm may choose different subsets of features for different subsamples (variations in the training set), but most of these subsets can produce similar results in terms of classification performance [17], [35]. Such instability undermines the confidence of domain experts in any of the various subsets of features selected for the task of biomarker identification in particular or knowledge discovery in general.

It is well known that the stability of a feature selection technique does not reveal much about the performance of the selected features [19] because in high-dimensional datasets many features are correlated with each other and/or redundant for the task at hand. Therefore, the question of how to select the most relevant features from different runs of the feature selection technique is very crucial. The stability of feature selection is very critical and misleading conclusions may be drawn when ignoring stability issues of feature subset selection techniques.

Measuring stability requires two aspects: a framework for studying stability and a stability measurement. The framework describes how to make changes on the input datasets in order to study the stability of a given technique, while the stability measurement is the specific metric for measuring stability. We further classify the stability measurement to two categories, measurement for same size of feature subsets and for varying sizes of feature subsets.

A. Framework

One common method for making changes on the input dataset is perturbation. Consider a dataset with m instances: a smaller dataset can be generated by keeping a fraction c of instances and randomly removing $1 - c$ of instances from the original data, where c is greater than 0 and less than 1. For a given c, this process can be performed x times. This will create x new subsamples, each having $c \times m$ instances, where each of these new subsamples is unique (since each was built by randomly removing $(1 - c) \times m$ instances from the original dataset). In circumstances involving class imbalance (datasets with extreme variety in class sizes), this subsampling may be performed on a per-class basis, to ensure the subsamples have the same class ratios as the original dataset. Researchers then apply the feature selection method in question to each of the datasets (all of the reduced datasets and sometimes the original dataset as well) and create a feature subset for each of the subsamples. Some researchers consider the random subsamples (the perturbed datasets) from the original dataset and compare the feature subsets chosen on these subsamples with each other; others compare the feature subsets chosen on the subsamples with those chosen from the original dataset.

Wang et al. [32] performed an empirical study to investigate the stability (robustness) and classification performance of eighteen filter-based feature ranking techniques on four different levels of perturbation (c was set to 95%, 90%, 80%, or 66.67%) on three real-world software engineering datasets. Results demonstrated that the number of instances deleted from the dataset affects the stability of the feature ranking techniques: the fewer instances removed from a given dataset, the less the selected features will change when compared to the original dataset, and thus the feature ranking performed on this dataset will be more stable.

The perturbation method above does not control for the degree of overlap between the subsamples being compared (instead leaving this to random chance). This makes it difficult to determine whether the similarity between feature subsets is due to the similarity of the underlying datasets or is a property of the feature selection technique used. Wang et al. [34] proposed a Fixed-Overlap Partitions Algorithm, which will create two subsets which have the same number of instances and a specified level of overlap. Note in this algorithm that c, the desired degree of overlap, can vary from 0 to 1, including the endpoints. A choice of $c = 0$ will find two entirely disjoint subsets, which will each contain half of the instances from the original dataset. On the other hand, $c = 1$ will create two copies of the original dataset which share all instances. This is generally not an interesting case to study, but is permitted by the algorithm. Results show that once again, the degree of overlap and feature subset size do affect the stability of feature selection methods.

Although few works consider the impact of dataset similarity when performing perturbation experiments, one paper, by Alelyani et al. [1], did so. In this paper, the authors noted that without controlling for similarity, it is difficult to tell whether two feature subsets are different due to underlying stability issues with the ranker or due to differences in the datasets they were drawn from. To evaluate this, the researchers sampled 25% of the instances into one subset, and

then created nine more subsets with exactly c of their instances in common with the first. The pairwise stability of the features from these subsets were evaluated as c varied from 0 to 1. They found that some algorithms were not able to outperform the inherent stability of the underlying datasets, and so should not be considered "stable" regardless of their stability performance.

Haury et al. [14] considered the role of overlap when measuring the stability of gene subsets. In addition to other analysis of their datasets, the researchers considered the fraction of instances in common when comparing feature lists generated from subsamples of the original data which either have 80% or 0% overlap. They also compared feature lists among four distinct (but related) datasets. They found that the similarity measures for the 0% overlap case more closely resembled the between-datasets case than did the results from the 80% overlap case. However, unlike the 0% case, where it is noted that the original data was divided into two mutually-exclusive groups (which therefore have 0% overlap), for the 80% case the two groups were generated by adding 80% of the data from the original dataset into each group, and then splitting the remaining 20% in half and putting each half into one of the groups. Thus, the 80% refers to proportion of the original data shared by the two groups, not the overlap between the two groups. This makes it difficult to generalize the approach to create datasets with arbitrarily-chosen overlaps.

Another method to make changes on the input data is cross validation. During the experiments, x runs of n-fold cross-validation are performed. For each of the n folds, one fold is used as test data while the other $n-1$ are used as the training data. For the purpose of measuring stability, researchers apply the feature selection method on the $n-1$ folds at each run. Loscalzo et al. [22] used 10-fold cross-validation to evaluate stability of a feature selection method. They apply SVM-RFE to 9 out of the 10 folds repeatedly. The stability of SVM-RFE is calculated based on the average pair-wise subset similarity of the top 10 features selected over the 10 folds.

B. Stability Metrics

In order to measure stability, first we have to decide upon the measurement metric. In recent years there have been a number of different stability measurements implemented for this exact purpose. These stability measurements are based on the Hamming Distance [7], the Jaccard index [27] (and its generalization, the Tanimoto distance [6]), the consistency index [17] and consistency-based similarity measure [23], Spearman's rank correlation coefficient [17], and Shannon entropy.

Dunne et al. [7] evaluate the stability of a feature selection method using Hamming distance. Let S_i and S_j be subsets of features,

$$H(S_i, S_j) = \sum_{k=1}^{n} |S_{ik} - S_{jk}| \tag{1}$$

where n is the total number of features in the dataset and S_{ik} denotes the k-th feature of subset S_i. Each subset is represented by a binary vector and each value in this vector is either 1 or 0. The value 1 at the position k of this binary vector indicates that the feature k is in the set, and the value 0 at the position k indicates that the feature k is not in the set.

Thus, given a set W of subsets of features, the total Hamming distance, H_t, is computed as follows:

$$H_t = \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} H\left(S_i, S_j\right) \tag{2}$$

The overall stability across all pairwise feature subset in S is then defined by the average normalized Hamming Distance obtained as follows:

$$\hat{H} = \frac{2 \times H_t}{n \times |W| \times (|W| - 1)} \tag{3}$$

Somol and Novovičová [30] extended a number of stability metrics, including Hamming Distance. They defined the Normalized Hamming Index as:

$$NHI = 1 - \frac{H\left(S_i, S_j\right)}{n} \tag{4}$$

The overall stability across all pairwise feature subset in W is then defined by the average normalized Hamming Distance obtained as follows:

$$ANHI = \frac{2 \times (\sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} NHI\left(S_i, S_j\right))}{|W| \times (|W| - 1)} \tag{5}$$

These two measures give the variation information in a set of feature subsets. The higher the average normalized Hamming Distance, the more information. One drawback of the measures is that they do not count the intersection between two subsets.

Kalousis et al. [17], Alelyani et al. [1], and Peteiro-Barral et al. [26] used the Jaccard index [27], or Jaccard similarity coefficient, as a metric for comparing the diversity of subsets of features. Let S_i and S_j be two different subset of features, the Jaccard index is defined as the cardinality of the intersection divided by the cardinality of the union of the two sets. It is shown in Equation 6.

$$J(S_i, S_j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \tag{6}$$

Alelyani et al. [1] used the Jaccard index to measure similarity between feature sets and defined the stability as the average of similarities across all W runs of the feature selection algorithm. The stability index is shown in Equation 7.

$$S_J = \frac{2}{|W| \times (|W| - 1)} \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} J(S_i, S_j) \tag{7}$$

The stability index (S_J) defined in Equation 7 varies in the interval of [0,1] where values near zero mean the feature selection results are not stable and values near 1 mean the results are stable. The value 1 means the results are identical.

By some set operations, the Equation 6 can be written as Equation 8.

$$T(S_i, S_j) = J(S_i, S_j) = 1 - \frac{|S_i| + |S_j| - 2|S_i \cap S_j|}{|S_i| + |S_j| - |S_i \cap S_j|} \tag{8}$$

Kalousis et al. [17] used the Tanimoto coefficient that is a generalized version of Jaccard index to measure similarity between two subsets of features (genes). The stability is computed as an average across several runs of a cross-validation procedure. The Tanimoto coefficient supports multiple classes and in the case of binary classes it is reduced to the Jaccard index. The Tanimoto distance is shown in Equation 8 and it measures the amount of overlap between two sets of arbitrary cardinality.

Peteiro-Barral et al. [26] used the Jaccard index with some variations across two datasets using three filter-based subset evaluation CFS, Consistency, and INTERACT described in section II above; they concluded that the Jaccard index is more influenced by the number of instances in the training dataset.

Kuncheva [20] developed an enhanced similarity measure called consistency index that takes into account the similarity between

subsets of features due to chance (randomness). Let S_i and S_j be subsets of features with equal cardinality, i.e., $|S_i| = |S_j| = t$. The consistency index is defined in Equation 9.

$$I_C(S_i, S_j) = \frac{dn - t^2}{t(n-t)} \tag{9}$$

where d is the cardinality of the intersection between two subsets S_i and S_j, n is the total number of features in the training datasets, and -1 < $I_C(S_i,S_j)$ ≤ +1. The greater the consistency index, the more similar the subsets are. A consistency index close to zero means that both subsets are similar to being drawn by chance (randomness).

Kuncheva et al. [20] extended their consistency index to expand beyond the comparison of just two subsets by taking the average across all pairwise consistency indices. Note that all feature subsets have the same size k.

$$AI_C = \frac{2}{|W|(|W|-1)} \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} I_C(S_i, S_j) \tag{10}$$

To improve Kuncheva's similarity measure, Lustgarten et al. [23] propose the similarity measure:

$$S_a(S_i, S_j) = \frac{|S_i \cap S_j| - \frac{|S_i| \times |S_j|}{n}}{min(|S_i|, |S_j|) - max(0, |S_i| + |S_j| - n)} \tag{11}$$

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 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 W subsets is calculated as:

$$ASM = \frac{2}{|W|(|W|-1)} \sum_{i=1}^{|W|=1} \sum_{j=i+1}^{|W|} S_a(S_i, S_j) \tag{12}$$

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

The above stability metrics measure similarity on feature subsets. These subsets can be selected by either feature subset selection methods or feature ranking methods where the top t features are selected from a ranked list. Kalousis et. al [17] presented a stability measure that use Spearman's rank correlation coefficient.

$$S_P(S_i, S_j) = 1 - 6 \sum_k \frac{(S_{ik} - S_{jk})^2}{t(t^2-1)}. \tag{13}$$

where S_{ik} and S_{jk} are the rank of feature k in rankings S_i and S_j, respectively, and t is the number of features being selected. For a set W of subsets of features obtained from $|W|$ ranking lists, the overall stability is computed as follows:

$$AS_P = \frac{2}{|W|(|W|-1)} \sum_{i=1}^{|W|-1} \sum_{j=i+1}^{|W|} S_P(S_i, S_j) \tag{14}$$

Shannon entropy-based stability measurement [19] starts from the idea that feature selection is choosing one of the 2^n possible feature subsets given the original n features. Random feature selection will choose randomly from within this collection, while stable feature selection will consistently choose the same or similar subsets. Perturbing the data allows for sampling the space of feature subsets chosen by a given feature selection approach, and observing the entropy of the resulting subsets allows the user to see how much variation there is among these. Shannon entropy is defined as:

$$H(X) = -\sum_{i=1}^{m} p(x_i) \log_2(p(x_i)) \tag{15}$$

where X is a random variable representing the m possible outcomes (in this case, $m = 2^n$ for the number of possible feature subsets) and $p(x_i)$ is the probability of the ith outcome.

In total, N total trials are performed, and G_{jt} is the frequency of the jth feature subset which contains t features. Note that j ranges from 1 to $C(n,t) = \binom{n}{t}$, because that is how many possible combinations of n features can be made choosing t at a time. We let $\overline{G}_{jt} = G_{jt}/N$ be the normalized frequency of each feature subset. We then can find the overall entropy from a collection of subsets with a chosen t:

$$\gamma_k = -\sum_{j=1}^{C(n,t)} \overline{G}_{jt} \log_2(\overline{G}_{jt}) \tag{16}$$

Among these similarity measures, only the stability measure proposed in [17], [20], the Spearman's rank correlation coefficient, and the Shannon entropy require that the feature subset size be the same for all compared subsets. All other measurements can be used with varying feature subset sizes, and thus are appropriate for considering the similarity of subset evaluation techniques where the feature subset size is not guaranteed to be constant for all subsets.

IV. Spotlight Topic: Subset Evaluation for Bioinformatics

While high dimensionality is important across a number of different application domains, one of particular note is bioinformatics. This application domain (due to the underlying nature of the problem) is rife with Big Data as it frequently has extremely high numbers of features (2,000 or more) and a large number of redundant or irrelevant features, and in addition frequently leads to datasets with very small numbers of instances (200 or fewer). For example, in the case of gene microarray datasets, gene expression levels are collected for thousands of genes, but only a relatively small number of patients are available in each study, and the majority of genes will not have direct relevance to the medical question being addressed; on the other hand, those genes which are relevant may be part of a regulatory network that shows highly correlated (e.g., redundant) behavior. Although these problems are those most in need of more advanced feature selection approaches such as filter-based subset evaluation, wrapper-based subset selection, and embedded approaches, the large number of features has so far stymied most efforts. These problems apply even more so to the question of feature selection stability: the large and noisy collection of features (and the biological importance of selecting the right genes) heightens the importance of gene selection stability, but the necessity of generating multiple feature subsets to compare for stability purposes magnifies the computational challenges by at least an order of magnitude. In this section, we review a number of papers which have begun to consider the topic of stability in bioinformatics, although we also discuss how these were limited in their scope by the magnitude of the problem.

Zengyou He and Weichuan Yu [15] provided a review of the topic of stable biomarker selection. They identified the underlying biological and methodological sources of instability. They then discussed a number of approaches for improving the stability of feature selection in bioinformatics. One approach, relevant for feature ranking (but not

subset selection), is ensemble ranking, which builds multiple ranked lists (either through different perturbations of the original data or through different choices of feature ranking technique) and combines these. Alternately, one may (a priori) give higher weight to features which are known (through domain knowledge) to have importance for the problem at hand. As for more general approaches, one cause for instability is multiple biomarkers being highly redundant with one another and thus varying which is selected when using subset evaluation techniques (since the first biomarker chosen from each group will preclude the inclusion of the others, and which happens to be chosen first may vary depending on data modifications). Thus, by identifying these groups in advance (either through data-driven approaches such as clustering, knowledge-driven approaches using existing results), the stability of such feature selection approaches may be improved. Following this discussion of techniques to improve feature selection stability, the authors presented a number of metrics for measuring stability, including both metrics which require a constant feature subset size (and hence which work best with rankers) and those which do not have this requirement (and hence are appropriate for evaluating the stability of subset-based techniques). The authors concluded that stability is an important element of biomarker selection which must be considered in order to contextualize the results. Our critique of this paper is that subset selection techniques such as filter-based subset evaluators and wrapper-based subset selection are not mentioned explicitly, and in fact subset evaluation is only mentioned in passing as one application of the more general stability metrics. Also, as this is a review paper, no experiments were performed.

Somol and Novovičová [30] conducted a comprehensive study of stability of feature selection techniques and investigated the problem of evaluating the stability of feature selection techniques that produce subsets of varying size. They examined a number of existing stability metrics, and note that many of these suffer at least one of two drawbacks: (1) various stability measures are differently bounded and thus are difficult to compare, and (2) most stability measures are considered for feature selection techniques with specified subset size although many feature selection techniques allow the subset size to be optimized in the process of searching for the most relevant and discriminatory features (genes). The second issue (varying subset size) is crucial particularly for the subset-based feature selection techniques, filter-based subset evaluation and wrapper-based subset selection, since the number of selected features are optimized by the search algorithms and the same algorithm operating on slightly perturbed versions of the same data may produce subsets of varying size. In fact, stability metrics may unfairly assign greater stability to larger subsets simply due to the larger chance for random overlap. The authors referred to this problem as ”subset-size bias problem”. The authors discussed how each metric performs in terms of these drawbacks, and also proposed a number of new metrics both extending the existing metrics and (in one case) designing a metric from scratch to have neither drawback. To evaluate the new stability metrics, the researchers conduced an empirical study on real data available from the UCI Repository. They used six datasets with 10–65 features and one dataset with 10,105 features; they used three wrapper techniques (Bayesian classifier assuming normal distribution, 3-nearest neighbor with majority voting, and support vector machine) on the former datasets and applied a filter technique to the latter dataset (10,105 features). This is due to the fact that applying wrapper techniques to high-dimensional datasets is not practical in terms of computational complexities. Our critique on the shortcoming of this work is that although the proposed framework is useful and examines the problem of feature stability metrics which are appropriate for subset evaluation-based feature selection, the wrapper-based feature selection techniques were only applied to datasets which are not high-dimensional (10–65 features), while a filter-based feature selection technique (ranker) was used for the dataset with high-dimensionality (10,105 features). Thus, although the proposed stability metrics and framework are useful for understanding high-dimensional data, they were not actually demonstrated in this capacity in the context of subset evaluation-based feature selection.

Lustgarten et al. [23] proposed a stability measure called the Adjusted Stability Measure (ASM, based upon extending the consistency index to varying feature subset size), as opposed to Unadjusted Stability Measure (USM, based on the Jaccard index), that computes robustness of a feature selection technique with respect to random feature selection method. The authors stated that the ASM is superior to other measures that do not account for random feature selection (a property it shares with the consistency index). That is, ASM is capable of indicating how much better or worse a feature selection method is over one that selects features at random. The researchers also discussed many of the stability measures presented in Section III above. To demonstrate how these two measures (ASM and USM) give different results, the authors measured the stability of wrapper-based subset evaluation using greedy forward selection based on three classification algorithms (Support Vector Machine, Logistic Regression, and Naïve Bayes) with a proteomic dataset obtained from the University of Pittsburgh Cancer Institute. The dataset contains 240 instances and a total of 70 features (protein probes) and a binary class variable (cancer vs. healthy). They concluded that lower stability may indicate either the feature selection method is not robust or the data contains many correlated, redundant or noisy features; measures that evaluate quality of feature set are unrelated to the measures of stability; a stability measure provides no information on the quality of the feature set; and a quality measure such as AUC provides no information on the stability of the selected features. Our critique on the shortcoming of this work is that the chosen dataset contains only 70 features, which is not high-dimensional compared to bioinformatics datasets with features in the range of thousands or even tens of thousands.

Díaz-Uriarte and Alvarez de Andrés [4] examined the performance and stability of an embedded ensemble feature technique centered around random forests (an ensemble technique which builds a collection of decision trees which each use random subsets of the data and features). In their technique, a random forest is constructed and the genes which are used with the least frequency are discarded. This process is repeated until a stopping criterion is reached, and the genes which remain are those selected. In addition, the random forest model from this collection which best minimizes the number of features while not sacrificing performance is used as the classification model. This embedded feature selection approach was tested on both simulated and real data. The simulated data containing a variable number of classes (2–4), independent dimensions (1–3), and genes per dimension (5–100). Ten real-world microarray datasets were also used, with between 2,000 and 10,000 genes in each dataset. The random forest-based approach was compared with three more traditional classifiers, Diagonal Linear Discriminant Analysis, k-Nearest Neighbor, and Support Vector Machines, each using F-ratio gene ranking as their feature selection technique. In addition, two established embedded techniques, Shrunken Centroid and Nearest Neighbor + Variable Selection, were employed in this comparison. Classification performance was evaluated by finding the error rate weighted between resubstitution (training set) and hold-out set error, and the stability of different feature selection techniques was

estimated using the frequency with which features selected from the whole dataset were also selected when using bootstrapped versions of the data. The authors found that their proposed random forest-based approach can lead to performance comparable with the existing filter-based ranking and embedded techniques, although it does not always produce the best results. Of the ten real-world datasets, four showed their best performance with filter-based feature selection, three were best with one of the previously-described embedded techniques, one was best with plain random forest (e.g., without building multiple random forest models), and two were best with the proposed embedded ensemble random forest approach. However, the proposed approach did consistently produce small feature subsets, while the other embedded approaches each left over 1,000 features for four of the ten datasets. In terms of stability, the proposed approach gave much less stable gene subsets than the previously-described embedded techniques, although this apparent stability may have arisen from the larger feature subset sizes of those techniques. Overall, the authors felt the smaller feature subset size and similar performance justified their proposed ensemble approach. Our main critique of this paper would be that stability is analyzed in a very naïve fashion, and that only embedded techniques are considered, not filter-based feature subset selection or wrapper-based techniques. Also, although the proposed approach shows promise, the random variable subset step inherent in the construction of a random forest model poses the risk of decreasing stability and increasing the chances that an important feature from the original dataset is not considered at all.

V. Future Work

Based on the lack of research which has examined subset evaluation-based feature selection (either filter-based or wrapper-based) in the context of bioinformatics, more work is needed in this area. For example, although more works are beginning to consider the problem of choosing an appropriate stability metric, there is no consensus as to which are most appropriate. In addition, no paper considers both filter-based feature subset evaluation and wrapper-based subset selection in the same work, especially in the context of biologically-relevant datasets (e.g., $>$2,000 features). The large scale of data does pose real challenges, and thus a major research target should be approaches which can mitigate these challenges while still bringing the full power of subset evaluation to Big Data bioinformatics datasets. One potential avenue for future research is hybrid feature selection, which couples a filter-based feature ranker with a subset-based feature selection technique. The ranker can reduce the number of features down to a more manageable level, while the subset-based technique will be able to eliminate redundant features and find features that are truly able to improve performance. A related concept is performing subset-based selection on a large but not intractable dataset and then comparing these results with a ranker performed on the same dataset. This will help demonstrate how these two techniques compare, and in particular how many features must be selected from the ranker before all of the important features (as defined using the subset-based results) are included. Thus, this could lead the way for a more data-driven hybrid approach which optimizes the number of features chosen in the first stage in order to select the important features without spending too much extra time during the second (subset-based) phase. Finally, future work may consider other ensemble and resampling techniques to reduce the number of features to a manageable level before performing subset evaluation, in order to find a near-optimal subset without searching the entire feature space. All of these have the potential to open new avenues for identifying the most important genes for bioinformatics problems, even when starting with thousands or tens of thousands of genes.

More generally, future work on the topic of subset selection stability across all domains will include an empirical study using various feature selection techniques (including filter-based subset evaluation and wrapper-based subset selection) over several datasets from different application domains: social media, software quality prediction, or biomedical datasets. Initially, we will focus on social media and software quality prediction models because these datasets have lower number of features in general. Eventually, we will perform an empirical study of subset-based feature selection techniques on biomedical datasets with high-dimensionality using a cluster of computers using HPCC (High Performance Computing Cluster) that is a massive parallel-processing computing platform to solve Big Data problems.

VI. Conclusions

With the proliferation of high-dimensional datasets in recent years, feature selection has received attention of researchers and data-mining practitioners due to its capability to improve both performance and computational efficiencies. A feature subset selected by a feature selection technique is evaluated for relevance toward a task such as classification or identification of the top relevant features corresponding to a phenomenon of interest. One important characteristics of a feature selection technique is stability, which refers to insensitivity (robustness) of the selected features to minor changes in the training dataset. In this survey paper, our focus is on the problem of stability, its importance, and various stability measures used to evaluate subsets of selected features. These subsets of features may be generated using filter-based, wrapper-based, or embedded techniques, although we particularly focus on metrics which can be used for subset evaluation-based approaches. In order to measure the stability of a feature selection technique, a similarity measure is needed to assess the overlap of a pair of feature subsets. The similarity measure computes the stability of a feature subset selection technique as the average similarity over all pairs of feature subsets. We discussed various similarity measures in this survey paper such as the Jaccard index, the Hamming distance, the Consistency index, Spearman's rank correlation coefficient, and Shannon entropy. Throughout this study, we have discussed current research in stability analysis of feature subset selection techniques within the domain of bioinformatics and have identified the shortcomings of these works to explore possible opportunities for future work.

References

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

[2] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, 2013.

[3] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial intelligence*, vol. 151, no. 1, pp. 155–176, 2003.

[4] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, p. 3, 2006. [Online]. Available: http://www.biomedcentral.com/1471-2105/7/3

[5] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, no. 263, p. 286, 1995.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley, 2012. [Online]. Available: http://books.google.com/books?id=Br33IRC3PkQC

[7] K. Dunne, P. Cunningham, and F. Azuaje, "Solutions to instability problems with sequential wrapper-based approaches to feature selection," *Journal of Machine Learning Research*, 2002.

[8] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, 2003.

[9] 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/cgi/content/abstract/286/5439/531

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[11] 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. [Online]. Available: http://dx.doi.org/10.1023/A:1012487302797

[12] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, April 1997.

[13] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 359–366.

[14] 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

[15] 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

[16] I. n. Inza, P. Larrañaga, 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, June 2004. [Online]. Available: http://dx.doi.org/10.1016/j.artmed.2004.01.007

[17] 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, May 2007. [Online]. Available: http://dx.doi.org/10.1007/s10115-006-0040-8

[18] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, Dec. 1997. [Online]. Available: http://dx.doi.org/10.1016/S0004-3702(97)00043-X

[19] P. Krízek, J. Kittler, and V. Hlavác, "Improving stability of feature selection methods," in *12th International Conference on Computer Analysis of Images and Patterns (CAIP)*, ser. Lecture Notes in Computer Science. Springer, August 2007, pp. 929–936.

[20] L. I. Kuncheva, "A stability index for feature selection," in *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

[21] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, April 2005.

[22] S. Loscalzo, L. Yu, and C. Ding, "Consensus group stable feature selection," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. New York, NY, USA: ACM, 2009, pp. 567–576.

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

[24] H. Nguyen, K. Franke, and S. Petrovic, "Improving effectiveness of intrusion detection by correlation feature selection," in *International Conference on Availability, Reliability, and Security (ARES '10)*, 2010, pp. 17–24.

[25] M. S. Pepe, R. Etzioni, Z. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget, and Y. Yasui, "Phases of biomarker development for early detection of cancer," *Journal of the National Cancer Institute*, vol. 93, no. 14, pp. 1054–1061, 2001.

[26] D. Peteiro-Barral, V. Bolón-Canedo, A. Alonso-Betanzos, B. Guijarro-Berdinas, and N. Sanchez-Marono, "Scalability analysis of filter-based methods for feature selection," *Advances in Smart Systems Research*, vol. 2, no. 1, pp. 21–26, 2012.

[27] R. Real and J. M. Vargas, "The probabilistic basis of jaccard's index of similarity," *Systematic Biology*, vol. 45, no. 3, pp. 380–385, 1996. [Online]. Available: http://www.jstor.org/stable/2413572

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

[29] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009. [Online]. Available: http://dx.doi.org/10.1007/s11263-007-0109-1

[30] P. Somol and J. Novovičová, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1921–1939, 2010.

[31] R. Wald, T. M. Khoshgoftaar, A. Napolitano, and C. Sumner, "Using Twitter content to predict psychopathy," in *11th International Conference on Machine Learning and Applications (ICMLA)*, vol. 2, 2012, pp. 394–401.

[32] H. Wang, T. M. Khoshgoftaar, and R. Wald, "Measuring robustness of feature selection techniques on software engineering datasets," in *2011 IEEE International Conference on Information Reuse and Integration (IRI)*, August 2011, pp. 309–314.

[33] H. Wang, T. M. Khoshgoftaar, and A. Napolitano, "Software measurement data reduction using ensemble techniques," *Neurocomputing*, vol. 92, pp. 124–132, 2012.

[34] 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 *13th IEEE International Conference on Information Reuse and Integration*, August 2012.

[35] 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

[36] Z. Zhao and H. Liu, "Searching for interacting features," in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. Morgan Kaufmann Publishers Inc., 2007, pp. 1156–1161.