The Effect of Measurement Approach and Noise Level on Gene Selection Stability

Randall Wald, Taghi M. Khoshgoftaar, Ahmad Abu Shanab
Florida Atlantic University, Boca Raton, Florida, 33431
Email: {rwald1, khoshgof, aabusha}@fau.edu

***Abstract*—Many biological datasets exhibit high dimensionality, a large abundance of attributes (genes) per instance (sample). This problem is often solved using feature selection, which works by selecting the most relevant attributes and removing irrelevant and redundant attributes. Although feature selection techniques are often evaluated based on the performance of classification models (e.g., algorithms designed to distinguish between multiple classes of instances, such as cancerous vs. noncancerous) built using the selected features, another important criterion which is often neglected is stability, the degree of agreement among a feature selection technique's outputs when there are changes to the dataset. More stable feature selection techniques will give the same features even if aspects of the data change. In this study we consider two different approaches for evaluating the stability of feature selection techniques, with each approach consisting of noise injection followed by feature ranking. The two approaches differ in that the first approach compares the features selected from the noisy datasets with the features selected from the original (clean) dataset, while the second approach performs pairwise comparisons among the results from the noisy datasets. To evaluate these two approaches, we use four biological datasets and employ six commonly-used feature rankers. We draw two primary conclusions from our experiments: First, the rankers show different levels of stability in the face of noise. In particular, the ReliefF ranker has significantly greater stability than the other rankers. Also, we found that both approaches gave the same results in terms of stability patterns, although the first approach had greater stability overall. Additionally, because the first approach is significantly less computationally expensive, future studies may employ a faster technique to gain the same results.**

***Keywords*-Stability, feature selection, noise injection, bioinformatics**

I. Introduction

High dimensionality is a common characteristic exhibited by many real-world datasets (especially in the domain of bioinformatics), and occurs when there are an especially large number of features per sample. With a large feature space, it is likely that many of the features are redundant (containing information already represented in other features) or useless (having little or no correlation with the class). Feature selection is the main preprocessing technique used to alleviate high dimensionality, and consists of finding a minimum subset of features that have the highest correlation with the class. This goal is achieved by distinguishing the relevant features from those which are irrelevant and redundant. Much work has been done towards evaluating feature selection techniques by comparing the classification performance of learners trained before and after feature selection. However, little work has addressed the stability of feature selection techniques. In this context, stability is defined as the degree of agreement between a feature selection technique's outputs when applied to differently-perturbed versions of the same input data, e.g., versions of the data which have had some instances added, removed, or modified. Stable feature selection techniques are favored over those that produce inconsistent output, especially in domains such as bioinformatics where frequently the goal is gene discovery, not classification. In other words, if the feature selection technique produces more consistent output, researchers can feel more confident that the chosen features are those with the most relevance to the class.

Noise is another common characteristic of many real-world datasets. Noise refers to errors or missing values contained in real-world data. Examples from the domain of bioinformatics include improper labeling or handling of samples, miscalibration of machinery, incorrect or expired chemicals, and other sources. There are two types of data noise: attribute noise and class noise. Attribute noise occurs when there are incorrect or missing values in the independent attributes of a dataset (e.g., the gene expression levels), while class noise occurs when an instance/example is given an incorrect label. Since most real-world datasets are prone to errors, there is a need to understand the impact of noise on the stability of feature selection. Thus, all experiments in this study were performed on data which was first determined to be free of noise and then had artificial class noise added in a controlled fashion. In addition, the stability measured in this paper represents how consistent a feature selection technique's output is in the presence of noise: a more stable technique will give the same results even if the data to which it is applied has had some artificial noise injected, corrupting the class values.

In this study we investigate two different approaches for evaluating the stability of feature selection techniques. The two approaches involve noise injection (with each technique being performed 30 times); the difference between the two approaches is which datasets are being compared to find the stability values. The first approach consists of noise injection followed by feature ranking, and the rankings of attributes from the noisy datasets are compared to the ranking when feature selection is performed on the original (clean) dataset. The second approach is similar, except that

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

each of the rankings from the noisy datasets are compared pairwise to each of the other noisy-dataset rankings, rather than being individually compared with the original (clean) dataset. Note that we do not employ the more traditional approach of measuring the variations in the chosen feature subsets built using different random subsamples of the original training dataset, as only noise-based stability is evaluated. To evaluate these two approaches, we employ six commonly-used feature rankers, namely Chi-Squared, Information Gain, Gain Ratio, two versions of ReliefF, and Symmetrical Uncertainty. We inject 24 different noise patterns into four cancer-gene datasets, and we repeat the process 30 times for each combination of noise pattern and dataset.

Our major results show that some rankers are better able to produce stable results in the face of noise than others: in particular, the ReliefF ranker has the greatest overall stability when we consider all 24 noise injection patterns. Also, we find that the two approaches discussed earlier (comparing noisy data to clean data or noisy data to noisy data) produce similar patterns in terms of the relative performance of the rankers, but that the first approach (noisy-vs-clean) produces greater stability on average. Because the first approach is also much less computationally expensive, we recommend its use in future experiments to evaluate the stability of rankers for large bioinformatics datasets.

The remainder of this paper is organized as follows. Section II presents related work. Section III introduces the methodology for our experiments, including the datasets and noise injection mechanism, the feature selection techniques, and the stability measure. Section IV presents our experimental results. Finally, conclusions and future work are presented in Section V.

II. Related work

High dimensionality is a common problem exhibited by many real-world datasets. This problem refers to the large number of attributes/features that describe an instance/sample in the dataset. High dimensionality can make machine learning computationally expensive and reduce the prediction accuracy of classifiers, because usually, most of these attributes are redundant (containing information already represented in other attributes) or useless (not having much correlation with the class) for building an inductive model. A simpler classifier and better performance can be obtained by performing feature selection to remove the redundant/useless attributes. Feature selection has received much attention in the machine learning community. As a result, much work has been done to evaluate the performance of feature selection techniques by comparing the classification performance of models trained before and after feature selection. Forman [3] investigated multiple filter-based feature ranking techniques. Sayes et al. [7] studied the use of ensemble feature selection methods and showed that the ensemble approach provides more robust feature subsets than a single feature selection method. Guyon and Elisseeff [4] outlined key approaches used for attribute selection, including feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods.

Table I: Datasets

Data set	# attributes	# instances	% positive	% negative
Lung cancer	12534	181	17.1	82.9
ALL	12559	327	24.2	75.8
Lung clean	12601	132	17.4	82.6
Ovarian Cancer	15155	253	36.0	64.0

Although many studies have examined feature selection techniques in the context of classification performance, few assess feature selection techniques in terms of stability, the degree of agreement between a feature selection technique's outputs when applied to differently-perturbed (changed) versions of the same input data [5]. Abeel et al. [1] studied the process for selecting biomarkers from microarray data and presented a general framework for stability analysis of such feature selection techniques. They showed that stability could be improved through ensemble feature selection. Dittman et al. [2] compared the stability of 19 different feature ranking techniques on 26 different biological datasets with varying levels of class imbalance, and found that the degree of imbalance can affect which ranking technique is most stable. Stiglic and Kokol [8] used three methods to evaluate the stability of univariate and multivariate gene selection techniques as the size of the datasets used is varied. They found that univariate feature selection (gene ranking) performed better than multivariate selection (subset evaluation), producing results on smaller datasets comparable to what the multivariate approaches achieved with much larger datasets.

In addition to high dimensionality, many real-world datasets are also characterized by noise. Noise refers to incorrect or missing values contained in real-world data. A special type of noise called class noise occurs when incorrect or missing values are in the class labels of instances in the dataset [12]. Noise has a detrimental effect on the stability of a feature ranker, which is to say it can cause the ranker to produce a different ranking of attributes than that which would be obtained when using the clean dataset. Therefore, all experiments in this study were performed on data which was first determined to be free of noise and then had artificial class noise added in a controlled fashion. This way, the results can be used to determine the impact of class noise on the stability (robustness) of feature rankers by showing which rankers are robust even in the face of noise.

III. Methodology

A. Datasets

Four binary cancer-gene datasets are considered in this set of experiments. Table I lists the four datasets used in this study, including their characteristics in terms of the total number of attributes, number of instances, percentage of positive instances, and percentage of negative instances. For all datasets, the minority class is considered to be the positive class. They are all binary class datasets. That is, for all the datasets, each instance is assigned one of two class labels. Furthermore, all datasets are high-dimensional and exhibit class imbalance.

The Lung Cancer dataset is a classification of malignant pleural mesothelioma (MPM) vs. adenocarcinoma (ADCA) of the lung, and consists of 181 tissue samples (31 MPM, 150 ADCA) [10]. The Acute Lymphoblastic Leukemia (ALL) dataset consists of 327 tumor samples of which 79 are positive (24.2%). The Lung Clean dataset was derived from a noisy lung cancer dataset containing 203 instances, including 64 (31.53%) minority instances and 139 (68.47%) majority instances. To produce a dataset that both was imbalanced and could be considered 'clean' (as defined by many classifiers having relatively near perfect classification on the dataset), a supervised cleansing process was used to reduce the original lung dataset. 5-fold cross-validation was performed on the original lung dataset using a 5NN classifier, and any instances which produced a probability of membership in the opposite class that was greater than 0.1 were removed. The Ovarian Cancer dataset consists of proteomic spectra derived from analysis of serum to distinguish ovarian cancer from non-cancer [6].

We selected these datasets for this study not only because they are high-dimensional and exhibit class imbalance, but also because these datasets are relatively clean (that is, show near-perfect classification accuracy and thus have very clear distinctions between the classes). Due to the fact that the datasets are relatively clean, we avoid any problems associated with injecting noise into datasets that are already noisy.

B. Noise injection

We used the same noise injection mechanism proposed by Van Hulse et al [9] where two parameters, α and β, control noise injection. The first parameter controls the overall noise level (in this study we used $\alpha \in \{10\%, 20\%, 30\%, 40\%, 50\%\}$), and the second parameter β controls the level of noise affecting the positive class (this study used $\beta \in \{0\%, 25\%, 50\%, 75\%, 100\%\}$). Specifically, the first parameter determines the number of instances which will have their class attributes flipped ($N_C = 2 \times \alpha \times N_P$, where N_P is the number of positive instances and N_C is the number of instances to corrupt), while the second parameter determines how many of these instances will be taken from each class ($N_{CP} = N_C \times \beta$ and $N_{CN} = N_C - N_{CP}$, where N_{CP} is the number of positive instances to corrupt and N_{CN} is the number of negative instances to corrupt). Note that the case with α = 50% and β = 100% was excluded, because this would result in a dataset where no minority-class instances remain. Note also that because the number of instances to be corrupted is tied to the number of minority-class instances, the quantity of noise injected into the dataset can somewhat misleading: for example, with the Ovarian Cancer dataset, even an α value of 50% means corrupting 23 instances out of 132 total.

The noise injection process is performed 30 times on each dataset for each noise corruption pattern. Thus, 720 different noisy versions of each of the four initial datasets were generated, with varying quantity and balance of noise depending on parameters α and β. For our results, we consider the averages across all 24 noise levels.

C. Feature Ranking Techniques

In this paper, we examine filter-based feature rankers, since wrapper-based techniques and subset evaluation can be very computationally expensive with thousands of features (as in our datasets). In particular, we used six "commonly-used" rankers found throughout the literature: Chi-Squared (CS), Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), and two types of ReliefF (RF and RFW). These techniques were chosen because they are frequently found throughout the literature and have been implemented within the open-source WEKA machine learning toolkit. These metrics are presented in [11], and a brief description is presented below.

Chi-Squared is a metric based on the χ^2 distribution, which is how the feature and class values would be distributed if there were no correlation whatsoever between the two. How far the actual distribution is from the theoretical no-correlation distribution shows how well the feature is correlated with the class. Information Gain is an entropy-based performance metric, based on the amount of entropy present in the partitioning of the instances based on their class values. The amount by which this entropy is reduced when the instances are first partitioned according to their feature values is how much information is gained when using that feature. Gain Ratio is a modification of IG which takes into account the inherent entropy of the feature values to reduce the problem of features with many values having artificially-high IG scores. Symmetric Uncertainty also modifies IG by taking into account the feature's inherent entropy, but it does this by dividing the IG by the sum of the feature and class's independent entropies. ReliefF is an instance-based performance metric based on the idea of picking a random instance and comparing its feature values with those from its nearest hit (the closest instance in the same class) and its nearest miss (the closest instance in the same class). Features increase their score by being close in

value in the nearest hit, but are penalized for being close in value to the nearest miss. ReliefF-W (abbreviated RFW) is a variant of ReliefF which applies distance-based weights to the nearest neighbors prior to adjusting the feature scores.

D. Stability Measure

In this study and to avoid bias due to chance we used the consistency index [5] to evaluate the stability of feature ranking techniques. First, the original dataset is assumed to have n features. T_i and T_j are two subsets of features, where $k = |T_i| = |T_j|$. When comparing T_i and T_j the consistency index is defined as follows:

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

where d is the cardinality of the intersection between subsets T_i and T_j, and $-1 < I_C(T_j, T_i) < 1$. The greater the consistency index I_C the more similar the subsets are. For each dataset and feature ranking technique, the features are ranked according to their relevance to the class, and then a subset consisting of the most relevant ones (top k features) is selected. In this study, we used eight sizes of feature subsets for each dataset (10, 14, 25, 0.25%, 0.5%, 0.5%, 1%, 2%, and 5%). Preliminary experiments conducted on the corresponding datasets show that these numbers are appropriate.

IV. Results and Analysis

As mentioned earlier, four datasets were used in this experiment. Six commonly-used feature ranking techniques are applied. We also evaluated 24 different combinations of noise level and noise distribution. We investigated two approaches to assess the robustness of feature rankers. Both approaches involve noise injection (repeated 30 times); the difference between the two approaches is whether the rankings from the noisy datasets are compared individually to the original (clean) dataset's ranking (Approach One), or pairwise with all other noisy-dataset rankings (Approach Two). We used the consistency index I_C to evaluate the stability of feature rankers, and eight sizes of feature subsets for each dataset. As we have four datasets, 24 noise patterns, and six feature rankers, we repeat the experiment 34,560 times. Due to space limitations, only the average results of the 30 repetitions of the noise patterns, the four datasets, and the 24 different combinations of noise level and noise distribution are presented.

Tables II and III contain the aggregated stability results for every combination of approach, ranker, and feature subset size, across all 24 noise injection patterns. We also present (1) the average performance (last column of the tables) of each of the feature rankers over the four datasets, and (2) the average performance (last row of the tables) of each subset size over the six feature rankers. In all tables "Attributes" is abbreviated as "Att" for space considerations, and bold values represent the best performance for that subset size. The first notable result is that when comparing the two approaches, Approach 1 has significantly higher stability than Approach 2. This makes sense, because while Approach 1 is comparing one noisy dataset to one clean dataset in order to calculate the stability, Approach 2 is comparing two different noisy datasets. Thus, there is more noise (and more potential for difference between the ranked feature lists being compared) present in Approach 2.

The results also show that different rankers respond differently to noise injection. In particular, we see that RF is the most stable ranker for most subset sizes, although some sizes exist where it comes in second or even third place. Notably, however, it is reliably among the best, which gives us confidence that it is a safe choice.

There is no clear pattern across all rankers for how the subset size affects stability. GR, for example, consistently shows increased stability as subset size increases. However, most other rankers have some internal optimum, a value of subset size which maximizes stability (and where either larger or smaller subsets will result in lower stability). IG has an internal optimum (0.597777) when evaluated with Approach 1, but with Approach 2 shows increased stability as subset size increases. RFW has an internal optimum for both approaches (0.419624 and 0.357467 for Approaches 1 and 2, respectively). Finally, CS, RF, and SU show both patterns as well as a third: multiple peaks of performance, with a dip in the middle. Overall, these results demonstrate that it is difficult to predict stability based solely on feature subset size, and in practice this must be evaluated on a case-by-case basis.

When looking across the two approaches (noisy-vs-clean versus noisy-vs-noisy), it can be seen that the two approaches have the same stability patterns, i.e., the rankers that perform well in the first approach perform as well in the second one and poor rankers in the first approach perform poorly in both approaches. This is despite Approach 2 being computationally more expensive than Approach 1 (due to Approach 1 performing 30 comparisons while Approach 2 must perform $30 \times 29/2 = 435$ comparisons). This demonstrates that Approach 1 is more suitable for comparing the stability of rankers, because it gives the same values more quickly. In addition, Approach 2 introduces more randomness (by comparing two noisy datasets), which has the potential to obscure the true results with unintended noise.

V. Conclusion

In this paper, we compared two approaches to assess the stability of feature selection techniques. The experiment was carried out on four relatively clean datasets from the field of bioinformatics. We injected noise into these datasets, and used six commonly-used feature ranking techniques. We applied the two approaches mentioned earlier (comparing the

Table II: Average Stability of the Filters for Approach 1 (Noisy-Vs-Clean), All Noise Levels

Filter	10 Att	14 Att	25 Att	0.25% Att	0.5% Att	1% Att	2% Att	5% Att	Avg
CS	0.517373	0.553762	0.589531	**0.598355**	0.594162	0.579401	0.579593	**0.582122**	0.574287
GR	0.321460	0.321620	0.317639	0.313116	0.319201	0.345924	0.378446	0.435891	0.344162
IG	0.497426	0.533848	0.583629	0.597777	0.594199	0.584182	0.584180	0.580757	0.569500
RF	**0.601291**	**0.586959**	**0.592583**	0.596376	**0.609493**	**0.616268**	**0.605559**	0.582062	**0.598824**
RFW	0.379936	0.378346	0.395095	0.394150	0.406834	0.419624	0.417887	0.409048	0.400115
SU	0.517268	0.539659	0.569786	0.573068	0.551528	0.553275	0.555717	0.558218	0.552315
Avg	0.472459	0.485699	0.508044	0.512140	0.512569	0.516446	0.520230	0.524683	0.506534

Table III: Average Stability of the Filters for Approach 2 (Noisy-Vs-Noisy), All Noise Levels

Filter	10 Att	14 Att	25 Att	0.25% Att	0.5% Att	1% Att	2% Att	5% Att	Avg
CS	0.424189	0.457904	0.503739	0.511790	0.507028	0.498824	0.508501	0.551023	0.495375
GR	0.295402	0.294548	0.294546	0.286941	0.285075	0.318177	0.350927	0.429440	0.319382
IG	0.416764	0.451810	0.505343	0.518391	0.516114	0.510284	0.520886	**0.557891**	0.499685
RF	**0.545855**	**0.541951**	**0.541739**	**0.546841**	**0.563093**	**0.571911**	**0.563396**	0.542347	**0.552142**
RFW	0.332291	0.334639	0.342598	0.343099	0.357467	0.370056	0.367590	0.363232	0.351372
SU	0.442199	0.464453	0.499035	0.500658	0.485086	0.479677	0.485412	0.522382	0.484863
Avg	0.409450	0.424217	0.447833	0.451287	0.452311	0.458155	0.466119	0.494386	0.450470

features chosen from noisy datasets to both the clean-dataset features and to the other noisy-dataset features) to evaluate their effectiveness in assessing the robustness of feature rankers. We used the consistency index I_C to evaluate the stability of feature rankers.

We found that the RF ranker was the most stable across all levels of noise. In addition, our results show that the two approaches will give the same stability pattern, while Approach 2 is computationally more expensive and shows reduced stability across the board. Thus, we recommend the use of Approach 1 in future experiments, because it can give the same results more quickly.

Future research may involve conducting more experiments, using other feature rankers, examining more (relatively clean) bioinformatics datasets, and considering additional approaches that involves data sampling techniques.

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, Feb. 2010.

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

[3] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, March 2003.

[5] 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.

[6] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer." *Lancet*, vol. 359, no. 9306, pp. 572–577, Feb. 2002.

[7] 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.

[8] G. Stiglic and P. Kokol, "Stability of ranked gene lists in large microarray analysis studies," *Journal of biomedicine biotechnology*, vol. 2010, 2010.

[9] J. Van Hulse and T. M. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data & Knowledge Engineering*, vol. 68, no. 12, pp. 1513–1542, 2009.

[10] X. Wang and O. Gotoh, "Accurate molecular classification of cancer using simple rules," *BMC Medical Genomics*, vol. 2, no. 1, p. 64, 2009.

[11] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Morgan Kaufmann, 2005.

[12] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review*, vol. 22, no. 3, pp. 177–210, Nov 2004.