Feature List Aggregation Approaches for Ensemble Gene Selection on Patient Response Datasets

Taghi M. Khoshgoftaar, Randall Wald, David J. Dittman, and Amri Napolitano
Florida Atlantic University, Boca Raton, FL 33431
Email: {khoshgof, rwald1, ddittman}@fau.edu, amrifau@gmail.com

***Abstract*—Many cancer treatments destroy healthy cells along with cancerous ones, and can leave patients fatigued and with a compromised immune system. This makes it especially important to determine whether or not a given cancer treatment will work for the patient or will just cause further harm. Recently there has been work on using gene expression profiles (DNA microarrays) to predict how a patient will respond to a cancer treatment. However, these profiles carry the problem of high dimensionality (a very large number of features (genes) per instance), thus necessitating dimension-reducing techniques such as feature (gene) selection (data pre-processing techniques from the domain of data mining to find an ideal feature set). A particularly promising subset of feature selection techniques are ensemble feature selection techniques, which perform multiple instances of feature selection and aggregate the results into a single decision. Traditionally, this is accomplished by ranking the features in each list by a metric and aggregating the ranks of each feature into a single final decision for the feature. Many forms of aggregation have been considered, both in terms of how to generate the distinct lists and how to combine the ranks from each list. However, all of these works have assumed ranks must be created per-list and then aggregated in a separate step – rather than aggregating the scores of each list directly and performing ranking only on the final list. This work compares two feature list aggregation approaches (rank-based aggregation and score-based aggregation) using the mean aggregation technique in terms of classification. We use fifteen patient response datasets along with three feature selection techniques as the basis for the ensemble feature selection, and we employ four feature subset sizes and two classifiers. Our results show that in general, the rank-based aggregation approach outperforms the score-based aggregation approach for a majority of scenarios for both classifiers. However, this is not always the case and careful consideration is required before making a decision between the two.**

***Keywords*-Patient Response; Classification; DNA Microarray; Ensemble Feature Selection;**

I. Introduction

One of the ongoing goals in bioinformatics and biomedicine is understanding the underlying genetic causes of cancer. Cancer is especially difficult for a number of reasons, including: the diseases is the patient's own cells, there are a number of areas in the cell reproduction cycle where the problem could have initially started, and cancer can move from one area of the body to another. Much research has gone into using gene expression profiles (DNA microarrays) for identifying whether or not the patient has cancer and if so, what type of cancer. However, recently there has been work in using gene expression profiles for determining how a patient will respond to a cancer treatment.

There are a number of options for combating cancer such as: surgery, chemotherapy, and focused radiation. Often the final treatment is a combination of treatments. Unfortunately, all of these treatments can carry extreme side effects: surgery can have complications during the process and both chemotherapy and radiation frequently destroy healthy cells on top of the cancerous ones. Additionally, the treatments frequently leave the patient fatigued and their immune system can be compromised. Therefore, it would be extremely useful to know ahead of time if the patient will benefit from a treatment or if it will just be damaging to the patient.

Unfortunately, gene expression profiles can be quite difficult to work with due to their inherent high dimensionality. Each sample or instance can be tested for thousands of genes simultaneously and identifying the correct genes is a challenge. However, dimension-reducing techniques such as feature selection (a data pre-processing tool from the domain of data mining which removes redundant and irrelevant feature and uses only the remaining useful genes in subsequent analysis) have been applied in order to facilitate working with these datasets. However, for some gene expression datasets, performing a single round of feature (gene) selection can give unstable results which are sensitive to slight changes in the input data [1]. New techniques are needed to select genes reliably, producing consistent and consistently-good results.

One of the more promising methods for resolving this problem is ensemble feature selection. In general, an ensemble feature selection technique will take the results of multiple iterations of feature ranking (a feature selection technique which ranks each feature by their ability to distinguish between different classes) and aggregates the resulting ranked feature lists into a single ranked list. Benefits of ensemble feature selection include more stable feature subsets and subsets that are as good if not better than those produced by a single technique in terms of classification [2].

Before performing ensemble feature selection one must decide on both how to generate the distinct lists in the

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

ensemble and how to aggregate the resulting feature lists. There are a number of techniques to choose from ranging from simple to complex. For most techniques the decisions are based on the ranking of the different features based the chosen metric, aggregating the rankings for each feature and using this to produce a final ranking. However, there is another way to view the aggregation step: aggregating the raw scores calculated using the chosen metric(s) instead of using the rankings based on those scores. No previous work has considered the use of scores directly for aggregation, with ranking only taking place following the aggregation step.

The primary goal of this work is to compare two feature list aggregation approaches: rank-based aggregation and score-based aggregation in terms of their classification performance. To this end, we use fifteen patient response datasets along with three feature selection techniques with varying levels of stability as the basis for the ensemble feature selection, as well as four feature subset sizes and two classifiers. We apply both approaches using mean aggregation. Our results show that in general, the rank-based aggregation approach outperforms the score-based aggregation approach for a majority of scenarios for both classifiers. This holds true for the feature selection techniques with the exception of Signal-to-Noise (which performs better with rank-based aggregation for Logistic Regression and score-based aggregation for Support Vector Machines). However, for some combinations of dataset and classifier score-based aggregation is the top performer. This leads us to state that one should not blindly pick the rank-based aggregation approach over the score-based aggregation approach as both should be considered when performing new experimentation. Additionally, we found that the more stable the feature selection technique, the smaller the difference in classification performance there is between the two approaches. This indicates that for the decision between the two approaches matters more for the unstable rankers. Further investigation is required in order to confirm the trends found in this work.

The remainder of this paper is organized as follows. Section II contains some related works to our topic. Section III outlines the two aggregation approaches. Section IV contains the details of how we performed the experiment. Section V contains the results of our experiments. Lastly, Section VI presents our conclusions and possible avenues for future work.

II. Related Works

Using gene expression profiles for patient response prediction has been a popular topic in recent years. The idea is to predict whether or not the patient will respond well to a treatment in order to determine the best treatment option. Eight of the studies from which we acquired our datasets focused on drug therapy. Ma et al.[3] performed a study on whether a patient with breast cancer will react well to treatment from a drug called tamoflaxin, an antiestrogen agent, using two gene expression profiles (one derived from microdissection and one from whole tumors). Pawitan et al. [4] and Chanrion et al. [5] also used tamoflaxin, but used it in different ways. Pawitan et al. used tamoflaxin as part of a larger therapy regime and Chanrion et al. were attempting to predict if the patient, after using tamoflaxin, would relapse or not. Raponi et al. performed two studies [6], [7] on the use of tipifarnib, a farnesyltransferase inhibitor, which is believed to treat blood disorders such as acute myeloid leukemia. Thuerigen et al. [8] worked on a series of drugs for the treatment of breast cancer, including gemcitabine, epirubicin, and docetaxel. Mulligan et al. [9] performed analysis on a protease inhibitor, bortezomib, used for patients whose multiple myeloma had relapsed.

The last three of the studies focused not on drug treatments but on treatments designed to eradicate the tumor directly. Watanabe et al. [10] used preoperative radiotherapy on patients of rectal cancer. Larsen et al. [11] focused on whether surgery alone is enough for the removal of lung squamous cell cancer. Wang et al. [12] focused on breast cancer patients who had their tumors surgically removed, of which a majority of them also received post surgery radiotherapy.

Univariate selection techniques, also known as feature ranking techniques, have become very popular tools in the analysis of genomic, medical, and bioinformatics data. There are a number of reasons for this, including: smaller computational demands when compared to other methods and producing output which is intuitive to understand (in particular, a list ranking the features in order of importance) [1]. Unfortunately, feature subsets derived from these techniques can be unstable. One solution to this problem is ensemble feature selection.

The use of ensembles has been most frequently applied to the creation of learners for building inductive models. It has been shown that these ensemble learners are competitive with other learners and in some cases are superior. In 2011, Dittman et al. [13] found that the ensemble classifier Random Forest outperformed five other learners in terms of accurately predicting the patient's response to a cancer treatment. Recently, there have been studies on applying the ensemble concept to the process of feature selection [14]. In 2012, Awada et al. [2] performed a survey of current methods of improving the stability of feature selection in bioinformatics data. Current research has shown that not only do models built with feature subsets created using ensemble methods have comparable (or better) classification performance (when compared to models built using a single feature selection method), but the feature subsets themselves are more robust and can be appropriately applied to other data from the same problem. Also in 2012, our research group [15] performed analysis on three different ensemble feature selection designs and compared them to a single run

Table I
DETAILS OF THE DATASETS

Name	Total # of Instances	# of Attributes	Average AUC
Mulligan 2007 (R vs. NR) [9]	169	22284	0.5931
Mulligan 2007 (R vs. PD) [9]	126	22284	0.6527
Ma 2004 (Microdissection) [3]	60	22576	0.5158
Ma 2004 (Whole Tumors) [3]	60	22576	0.6808
larsen2006 [11]	51	21323	0.4134
Raponi 2007 (R vs. PD) [6]	54	22284	0.44200
Raponi 2007 (SD + R vs. PD) [6]	58	22284	0.4739
Raponi 2008 (R vs. PD) [7]	26	22284	0.6425
Raponi 2007 (SD + R vs. PD) [7]	34	22284	0.5995
Watanabe 2006 [10]	46	12626	0.4487
Chanrion 2008 [5]	155	22657	0.6721
Thuerigen 2006 (Cy3 / Cy5) [8]	96	21881	0.6268
Thuerigen 2006 (Cy5 / Cy3) [8]	94	21881	0.5915
Wang 2005 [12]	286	12066	0.5698
Pawitan 2005 [4]	159	12066	0.6108

of feature selection and found that the ensemble techniques were frequently the top performers. These techniques all used rank-based aggregation, however.

III. AGGREGATION APPROACHES

Aggregating the various outputs from the multiple runs of feature selection into a single decision is a key component in ensemble feature selection. Therefore, how one decides to aggregate the outputs is an important decision to make when considering applying ensemble feature selection designs. There are a number of different techniques for feature list aggregation including: median, lowest rank, highest rank, and mean. However, with each of these techniques there are two approaches towards the aggregation: rank-based aggregation and score-based aggregation. As this is a preliminary study we decided to use the most commonly used ranked list aggregation technique: mean aggregation.

Rank-based aggregation (See Figure 1) is the most commonly used approach toward aggregating. In this approach each feature is scored by a feature selection technique and then ranked based on the score with "1" being the top ranked feature all the way to "n" being the nth ranked feature. This process is repeated for each iteration of feature selection, resulting in a collection of ranked lists. These ranked lists are then combined based on their rank (in our case, using the mean of the ranks) and one final ranking using the aggregated ranks is created.

Score-based aggregation (See Figure 2) begins with the feature selection step as with rank-based aggregation. However, unlike rank-based aggregation, the features are not ranked from top to bottom but are only scored based on the feature selection technique. It is this score, not a rank, that is used in the aggregating step. Once the aggregation is completed the features are finally ranked based on the aggregated score.

IV. CASE STUDY

A. Datasets

Table I contains the list of datasets used in our experiment along with their characteristics. All of the datasets focus on the problem of patient response prediction: each instance is labeled based on how the patient in question responded to a specific treatment regime. As some of the gene selection techniques used in this paper require that there be only two classes, we can only use datasets with two classes. The datasets in Table I show a variety of different characteristics such as number of total instances (samples or patients) and number of features. High dimensionality is clear in the datasets used due to the large number of attributes, especially compared with the number of instances.

The last column (Average AUC) represents how difficult the datasets are to learn from by observing the classification performance when no feature selection technique is used. The performance is measured using AUC, the Area under the ROC Curve, where the ROC Curve itself is a plot of True Positive Rate versus False Positive Rate, and thus the AUC shows how the model balances these two values. We produce these values by applying 5-fold cross-validation (the instances in the dataset are split into five equal folds and training and testing are repeated five times so that each of the folds is used as a testing set and the remaining folds are used to train the classifier) using six classification approaches or learners: 5-Nearest Neighbor (an instance based learner), Multilayer Perceptron (an artificial neural network), Naive Bayes (a Bayesian learner), Support Vector Machines (see section IV-D), C4.5D (C4.5 decision tree using default values), and C4.5N (C4.5 decision tree using Laplace smoothing and no pruning). All of these learners are available with the Weka machine learning toolkit [16], with the following changes to default values: 5-Nearest Neighbor used "weight by 1/Distance;" Multilayer Perceptron used a hidden layer with three nodes and held 10% of the data as a validation set to determine when to stop training; Support Vector Machines used a complexity constant "c" of 5.0 and did build logistic models; and C4.5N (as noted) used Laplace smoothing and no pruning. These learners were chosen for evaluating difficulty-of-learning due to their diversity and general high performance across many datasets; by observing how these work on the datasets with no feature selection, we can get a general sense of how difficult the datasets are to learn from in general.

B. Feature Selection Techniques

We use three different feature selection algorithms in this research: Information Gain (IG) [17], Area Under the ROC Curve (ROC) [18], and Signal-to-Noise (S2N) [19]. Each of these techniques are filter-based feature ranking techniques. The reason we only use filter feature ranking techniques is that for the large degree of high dimensionality seen in

Figure 1. Rank-Based Aggregation

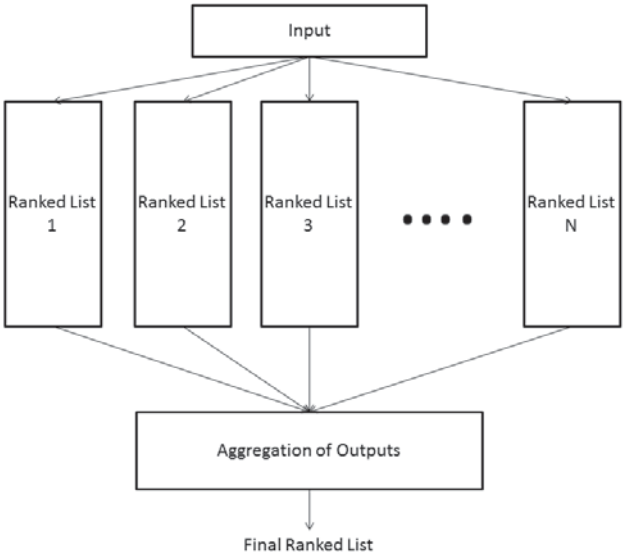
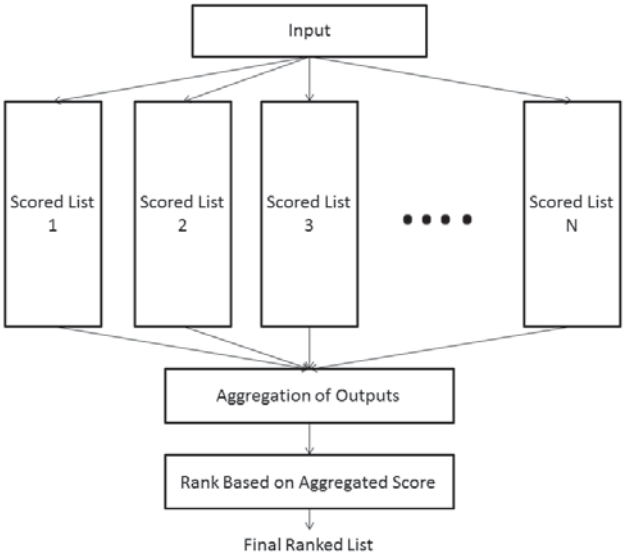


Figure 2. Score-Based Aggregation



these datasets, techniques such as filter-based subset evaluation and wrapper-based techniques are too computationally expensive to be of use.

Information Gain (IG) is one of the simplest and fastest feature ranking techniques, and is thus popular in bioinformatics where high dimensionality makes some of the more complex techniques infeasible. IG determines the significance of a feature based on the amount by which the entropy of the class decreases when considering that feature.

Receiver Operating Characteristic, or ROC, curves are a graph of the true positive rate on the y-axis versus the false positive rate on the x-axis. In the context of feature selection, this is found by considering a dataset which contains only the class feature and the feature being evaluated, and treating the normalized feature value as a posterior probability: values above a certain threshold are considered positive instances, while those below the threshold are negative instances. (The reverse, with values above the threshold being negative and those below being positive, is also tested, and the direction with the best result is used.) The ROC curve is graphed as this threshold varies, and the curve itself represents the trade-off between the rate of detection and the rate of false alarms. The area under this curve (AUC) is thus used as a single-value metric for the importance of the feature.

S2N is less often used in the context of feature selection.

The signal-to-noise ratio, or S2N, as it relates to classification or feature selection, represents how well a feature separates two classes. The equation for signal to noise is:

$$S2N = (\mu_P - \mu_N)/(\sigma_P + \sigma_N)$$

where μ_P and μ_N are the mean values of that particular attribute in all of the instances which belong to a specific class, either P or N (the positive and negative classes). σ_P and σ_N are the standard deviations of that particular attribute as it relates to the class. The larger the S2N ratio, the more relevant a feature is to the dataset.

One of the main reasons we chose these techniques is that they represent different levels of feature list stability. We feel this consideration is important as one of the goals of ensemble feature selection is to improve the stability of feature selection techniques. Thus, it is important to observe how the choice between rank-based aggregation and score-based aggregation affects rankers with different levels of inherent stability. According to previous research [20] we see that IG has average to below average stability; ROC is one of the most stable feature selection techniques; and S2N is above average in terms of stability.

C. Ensemble Design

For this experiment, we bootstrap twenty different bags (sampling with replacement) [14] from each of the fifteen patient response datasets used. We apply a single ranker to each of the twenty bags. This process is performed a total of three times, one for each of the rankers.

The twenty lists generated per ranker are aggregated using one of two aggregation approaches to generate one ranked list. This process will be performed for each of the two approaches towards aggregation. Therefore, in the course of our study (15 datasets $\times$ 20 bags $\times$ 3 feature ranking techniques $\times$ 2 feature list aggregation approaches $\times$ 4 runs $\times$ 5 folds) = 36,000 ranked feature lists were computed.

D. Classifiers

We used two different classifiers to create inductive models from the features chosen by the ensemble feature selection technique. These models are used to evaluate the predictive power of the genes chosen by applying them to a set of learners with varied attributes. In terms of inductive models we built (15 datasets $\times$ 3 feature selection techniques $\times$ 2 feature list aggregation techniques $\times$ 2 learners $\times$ 4 feature subset sizes $\times$ 4 runs $\times$ 5 folds) = 14,400 inductive models. Note that feature selection was performed within the cross-validation procedure, such that feature subsets were chosen based on the same training folds as the classification models they would be used with. The two learners used are described below.

Support Vector Machines, or SVM, is one popular choice of classification algorithm. One of the most efficient ways to classify between two classes is to assume that both classes are linearly separated from each other. This assumption allows us to use a discriminant to split the instances into the two classes before looking at the distribution between the classes. A linear discriminant uses the formula $g(x|\mathbf{w}, \omega_0) = \mathbf{w}^T x + \omega_0$. In the case of the linear discriminant the only data that needs to be learned is the weight vector, $\mathbf{w}$ and the bias ω_0. One aspect that must be addressed is that there can be multiple discriminants that correctly classify the two classes. SVM is a linear discriminant classifier which assumes that the best discriminant maximizes the distance between the two classes. This is measured in the distance from the discriminant to the samples of both classes [16]. For our study, we used a complexity constant "c" value of 5.0, and the "buildLogisticModels" parameter was set to "true."

Logistic Regression (LR) is a statistical technique that can be used to solve binary classification problems. Based on the training data, a logistic regression model is created which is used to decide the class membership of future instances [21].

V. Results

In this work we compare two approaches toward aggregating the multiple runs of feature selection that results from ensemble gene (feature) selection: rank-based aggregation or score-based aggregation. We use fifteen patient response datasets and three feature selection techniques in the process of our experiments. Additionally, we use four feature subset sizes (10, 25, 50, and 100) along with the SVM and LR classifiers in order to test the two techniques in terms of classification performance. The base aggregation technique which will utilize the two approaches is mean aggregation. Tables II and III contain the average Area Under the ROC Curve across the fifteen datasets where the feature selection technique and the feature subset size are kept static. For each combination of learner, feature selection technique, and feature subset size, we compare the rank-based aggregation results and the score-based aggregation results and the top performer in each scenario is in **boldface**.

Looking at the results using SVM (See Table II) we see that in a majority of the scenarios (eight out of twelve) the rank-based aggregation approach outperforms the score-based aggregation approach. From the feature selection technique point of view, two of the three (ROC and IG) techniques perform better using rank-based aggregation for all feature subset sizes with the exception of using ROC and a feature subset size of ten. The last technique, S2N, produces better results with score-based aggregation for all feature subset sizes except when using twenty-five features.

Looking at the results using LR (See Table III) we see that in a majority of the scenarios (ten out of twelve) the rank-based aggregation approach outperforms the score-based aggregation approach. From the feature selection technique point of view, all three techniques prefer rank-based aggregation over score-based aggregation for all feature subset sizes with two exceptions: using S2N with ten features and using

Table II
AVERAGE CLASSIFICATION RESULTS: RANK-BASED AGGREGATION VS SCORE-BASED AGGREGATION - SVM

Filter	Feature Subset Sizes							
	10		25		50		100	
	Type of Aggregation		Type of Aggregation		Type of Aggregation		Type of Aggregation	
	Rank	Score	Rank	Score	Rank	Score	Rank	Score
ROC	0.59503	**0.59603**	**0.59381**	0.59149	**0.59456**	0.59302	**0.59148**	0.58752
IG	**0.61227**	0.59623	**0.61159**	0.58943	**0.60956**	0.59562	**0.59877**	0.59009
S2N	0.60513	**0.60964**	**0.60593**	0.59966	0.59742	**0.60121**	0.59107	**0.59540**

Table III
AVERAGE CLASSIFICATION RESULTS: RANK-BASED AGGREGATION VS SCORE-BASED AGGREGATION - LR

Filter	Feature Subset Sizes							
	10		25		50		100	
	Type of Aggregation		Type of Aggregation		Type of Aggregation		Type of Aggregation	
	Rank	Score	Rank	Score	Rank	Score	Rank	Score
ROC	**0.60102**	0.59948	0.58031	**0.58359**	**0.57865**	0.57043	**0.56791**	0.56144
IG	**0.61490**	0.58838	**0.60536**	0.58241	**0.59548**	0.56730	**0.57688**	0.56778
S2N	0.61176	**0.61327**	**0.58821**	0.57282	**0.57265**	0.56671	**0.56591**	0.55443

Table IV
ABSOLUTE VALUE OF THE CLASSIFICATION DIFFERENCE BETWEEN AGGREGATION APPROACHES

Filter	SVM				Logistic Regression			
	Feature Subset Sizes				Feature Subset Sizes			
	10	25	50	100	10	25	50	100
ROC	*0.00100*	*0.00231*	*0.00155*	*0.00396*	0.00154	*0.00328*	0.00822	*0.00647*
IG	**0.01604**	**0.02216**	**0.01394**	**0.00868**	**0.02652**	**0.02295**	**0.02818**	0.00910
S2N	0.00451	0.00627	0.00379	0.00434	*0.00151*	0.01538	*0.00595*	**0.01147**

ROC with twenty-five features. This leads us to state that both learners and all three feature selection techniques will, in general, have rank-based aggregation outperform score-based aggregation with the only exception being using S2N with SVM.

Looking deeper with each scenario we see that the differences between the techniques varies depending on the feature selection technique being used. Table IV contains the absolute values of the differences between the average classification results for the rank-based and score-based aggregations for both learners. Additionally for each combination of learner and feature subset size the largest difference is in **boldface** and the smallest is in *italics*.

In SVM we see that for all feature subset sizes the feature selection technique with the largest difference between the two approaches is IG and the techniques with the smallest difference is ROC. In terms of LR we see that for all but 100 features the feature selection technique with the largest difference between the two approaches is IG. The feature selection technique with the largest difference between the approaches using 100 features and LR is S2N. In terms of the smallest differences we have two techniques which have the smallest difference between the approaches: ROC when using twenty-five and one-hundred features and S2N when using ten and fifty features. These results indicate that the decision between rank-based and score-based aggregation is especially important for IG and less important for ROC. We believe this trends stem from the relative stability of the feature rankers. We see that the largest difference is generated from the least stable ranker (IG) and the smallest difference is from the most stable ranker (ROC). Further research is required in order to determine if these trends will persist with other rankers of similar stability.

Lastly, when we look at the results from the individual datasets (tables omitted due to space considerations), we see that for seventeen out of the thirty possible pairwise dataset/learner combinations the rank-based aggregation approach outperforms the score-based aggregation approach in a majority of scenarios. Of the remaining thirteen combinations, five of them have the rank-based aggregation approach outperforming the score-based aggregation approach in exactly 50% of the scenarios. Additionally, there is not a dataset where for both learners score-based aggregation outperforms rank-based aggregation in a majority of scenarios. These results show that while rank-based aggregation performs best in the majority of scenarios, one cannot simply pick rank-based aggregation and expect the optimum results; the decision should be on a dataset by dataset basis.

VI. CONCLUSION

When working with ensemble gene selection, the choice of how to aggregate the results is an important decision. However, in addition to choosing the aggregation technique (e.g., mean aggregation) it is important to also choose the aggregation approach (e.g., rank-based or score-based aggregation). In this work we compare two aggregation ap-

proaches (rank-based and score-based aggregation) in terms of their effects on classification performance using fifteen patient response datasets. In addition to the datasets we use three feature selection techniques, four feature subset sizes, and two classifiers.

In general, we find that the rank-based aggregation approach outperforms the score-based aggregation approach in a majority of scenarios for both learners. In terms of the feature selection technique, two of the three techniques, ROC and IG, prefer rank-based aggregation over score-based aggregation for both learners. The final ranker, S2N, prefers score-based aggregation for SVM and rank-based aggregation for LR. These results allow us to state that rank-based aggregation is generally preferred, but one should not just simply pick it without considering score-based aggregation.

We also noticed that different feature selection techniques will have smaller or larger differences between the two approaches. ROC generally had the smallest difference between the two approaches with two exceptions: using LR with ten or fifty features. IG had the largest differences with one exception: using SVM and one-hundred features. Therefore, the decision between the approaches is more important for IG than for other feature selection techniques like ROC and (to a lesser extent) S2N. We believe this trend is due to the relative stabilities of the rankers (ROC being a very stable ranker, IG being of average to below average stability, and S2N having above average stability) [20].

Lastly, we looked at the individual datasets themselves. What we observed was that while a majority of the dataset/learner combinations either performed better with rank-based aggregation or were split between the two approaches, there were still combinations which excelled using score-based aggregation. This gives further evidence that although rank-based aggregation is preferred the majority of the time, some scenarios will prefer score-based aggregation.

Future work in this area will include more feature selection techniques with different levels of stability. This would allow us to confirm the trends based on the stability of the feature selection technique. Another option for future work is to use other groups of datasets (tumor identification, patient diagnosis) to observe the trends in these groups.

REFERENCES

[1] 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

[2] W. Awada, T. M. Khoshgoftaar, D. J. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *Information Reuse and Integration (IRI), 2012 IEEE International Conference on*, Aug. 2012, pp. 356–363.

[3] X.-J. Ma, Z. Wang, P. D. Ryan, S. J. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, J. Tuggle, Y. Tran, D. Tran, A. Tassin, P. Amon, W. Wang, W. Wang, E. Enright, K. Stecker, E. Estepa-Sabal, B. Smith, J. Younger, U. Balis, J. Michaelson, A. Bhan, K. Habin, T. M. Baer, J. Brugge, D. A. Haber, M. G. Erlander, and D. C. Sgroi, "A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen," *Cancer Cell*, vol. 5, no. 6, pp. 607 – 616, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1535610804001412

[4] Y. Pawitan, J. Bjohle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. Liu, L. Miller, H. Nordgren, A. Ploner, K. Sandelin, P. Shaw, J. Smeds, L. Skoog, S. Wedren, and J. Bergh, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts," *Breast Cancer Research*, vol. 7, no. 6, pp. R953–R964, 2005. [Online]. Available: http://breast-cancer-research.com/content/7/6/R953

[5] M. Chanrion, V. Negre, H. Fontaine, N. Salvetat, F. Bibeau, G. M. Grogan, L. Mauriac, D. Katsaros, F. Molina, C. Theillet, and J.-M. Darbon, "A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer," *Clinical Cancer Research*, vol. 14, no. 6, pp. 1744–1752, 2008. [Online]. Available: http://clincancerres.aacrjournals.org/content/14/6/1744.abstract

[6] M. Raponi, J.-L. Harousseau, J. E. Lancet, B. Lwenberg, R. Stone, Y. Zhang, W. Rackoff, Y. Wang, and D. Atkins, "Identification of molecular predictors of response in a study of tipifarnib treatment in relapsed and refractory acute myelogenous leukemia," *Clinical Cancer Research*, vol. 13, no. 7, pp. 2254–2260, 2007. [Online]. Available: http://clincancerres.aacrjournals.org/content/13/7/2254.abstract

[7] M. Raponi, J. E. Lancet, H. Fan, L. Dossey, G. Lee, I. Gojo, E. J. Feldman, J. Gotlib, L. E. Morris, P. L. Greenberg, J. J. Wright, J.-L. Harousseau, B. Lwenberg, R. M. Stone, P. De Porre, Y. Wang, and J. E. Karp, "A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia," *Blood*, vol. 111, no. 5, pp. 2589–2596, 2008. [Online]. Available: http://bloodjournal.hematologylibrary.org/content/111/5/2589.abstract

[8] O. Thuerigen, A. Schneeweiss, G. Toedt, P. Warnat, M. Hahn, H. Kramer, B. Brors, C. Rudlowski, A. Benner, F. Schuetz, B. Tews, R. Eils, H.-P. Sinn, C. Sohn, and P. Lichter, "Gene expression signature predicting pathologic complete response with gemcitabine, epirubicin, and docetaxel in primary breast cancer," *Journal of Clinical Oncology*, vol. 24, no. 12, pp. 1839–1845, April 20, 2006. [Online]. Available: http://jco.ascopubs.org/content/24/12/1839.abstract

[9] 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.

[10] T. Watanabe, Y. Komuro, T. Kiyomatsu, T. Kanazawa, Y. Kazama, J. Tanaka, T. Tanaka, Y. Yamamoto, M. Shirane,

T. Muto, and H. Nagawa, "Prediction of sensitivity of rectal cancer cells in response to preoperative radiotherapy by dna microarray analysis of gene expression profiles," *Cancer Research*, vol. 66, no. 7, pp. 3370–3374, 2006. [Online]. Available: http://cancerres.aacrjournals.org/content/66/7/3370.abstract

[11] J. E. Larsen, S. J. Pavey, L. H. Passmore, R. Bowman, B. E. Clarke, N. K. Hayward, and K. M. Fong, "Expression profiling defines a recurrence signature in lung squamous cell carcinoma," *Carcinogenesis*, vol. 28, no. 3, pp. 760–766, 2006. [Online]. Available: http://carcin.oxfordjournals.org/content/28/3/760.abstract

[12] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. M. van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671 – 679, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673605179471

[13] D. J. Dittman, T. M. 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.

[14] 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.

[15] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and A. Napolitano, "Comparing two new gene selection ensemble approaches with the commonly-used approach," in *Proceedings of the Eleventh International Conference on Machine Learning and Applications (ICMLA): Health Informatics Workshop*. ICMLA, 2012, p. In Press.

[16] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[17] 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.

[18] W. J. Conover, *Practical Nonparametric Studies*. John Wiley and Sons, 2nd edition, 1971.

[19] M. Wasikowski and X. wen Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1388–1400, 2010.

[20] D. J. Dittman, T. M. 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.

[21] S. Le Cessie and J. C. V. Houwelingen, "Ridge estimators in logistic regression," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, pp. 191–201, 1992.