The Effect of Number of Iterations on Ensemble Gene Selection

Wael Awada, Taghi Khoshgoftaar, David Dittman, and Randall Wald
Florida Atlantic University, Boca Raton, FL 33431
Email: waelawada@gmail.com; khoshgof@fau.edu; ddittman@fau.edu; rwald1@fau.edu

***Abstract*—Dimensionality-reducing techniques such as gene selection have become commonplace in order to reduce the high dimensionality found within bioinformatics datasets such as DNA microarray datasets. The degree of dimensionality is reduced by identifying and removing redundant and irrelevant features or genes and leaving only an optimum subset of features for subsequent analysis. However, a number of feature selection techniques show poor stability (resistance to change in the underlying data). One approach for increasing the stability of feature subsets is ensemble feature selection. This is performed first by generating multiple ranked gene lists and then aggregating the results using an aggregation function. While research has been performed on ensemble feature selection and its effect on gene list stability, there has been little research on an important choice made in the process of ensemble feature selection: the number of iterations (or repetitions) of feature selection. The computation time of ensemble feature selection is greatly affected by the number of ranked lists generated: the higher the number of iterations, the more computation time is required. To study this, we evaluate the similarity among feature subsets generated from two different approaches to ensemble feature selection (data diversity and hybrid approach). We calculate the similarity between the final ranked lists generated using 10, 20 and 50 iterations, using the mean aggregation function. Our results show that the similarity between 20 and 50 iterations is high enough for us to recommend using 20 iterations instead of 50 and thus saving the large amount of computation time required for 50 iterations.**

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

I. Introduction

Dimensionality reduction techniques play an important role in the analysis of bioinformatics data. High dimensionality is a problem that affects bioinformatics data since the number of features is very large, while the number of samples (instances) is much smaller. One set of techniques called feature selection achieves dimensionality reduction by selecting only highly relevant genes that are related to the problem at hand and ignoring irrelevant and redundant features in subsequent analysis. Univariate feature selection techniques (techniques which analyzes each feature separate from the others) are chosen in this domain due to their low computation time and easy-to-understand output (a ranked list) [1].

Unfortunately, it has been shown that feature subsets chosen by univariate feature selection techniques can be unstable. Different subsets are generated following a small change to the data: when samples are added or removed, the resulting feature subsets change. The irreproducibility of feature subsets causes confusion among researchers who are studying the data generated.

One way to solve the stability problem is ensemble feature selection [1], which is based on ensemble learning where multiple classifiers are used and their results are aggregated. Similar to this, ensemble feature selection uses multiple instances of feature selection to generate feature subsets that are later aggregated using an aggregation function. This technique minimizes the variability of relying on a single instance of feature selection.

However, there is a significant decision to be made prior to applying ensemble feature selection: how many iterations of feature selection will be performed before aggregation. This is an important decision because the amount of computation time required is dependent on how many iterations of feature selection that occurs. Essentially, if the same feature subsets are generated for two different numbers of iterations then the smallest number is more useful due to the decreased computation time. Thus, the goal of this research is to compare different numbers of iterations and determine which (if any) give similar results.

In this study we analyze the similarity among the feature subsets generated from two ensemble feature selection approaches: data diversity (using a single feature selection technique on multiple sampled datasets derived from the same dataset) and a hybrid approach (using a set of feature selection techniques on a set of sampled datasets derived from a same dataset). We test the importance of iteration count by using three different numbers of iterations (10, 20, and 50) on 26 DNA microarray datasets. We also use 10 univariate (filter-based) feature selection techniques and 12 feature subset sizes in conjunction with the ensemble approaches. Our results show that there is a large amount of similarity between the gene lists generated when using either 20 or 50 iterations. This leads us to state that using 50 iterations of feature selection is unnecessary when one can use 20 iterations and create highly similar gene lists.

The remainder of this paper is organized as follows. Section II contains some related works to our topic. Section III outlines the feature selection processes involved in this study. Section 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

In the biomedical domain, DNA microarray data suffers from high dimensionality. High dimensionality is caused by the high number of attributes compared to the low number of samples. To solve this problem, feature selection techniques are used to select the most relevant features related to the problem (such as: cancerous / non cancerous, relapse / no relapse, etc). Selecting the most relevant features leads to using less features in the classification stage which helps in lowering the computation time [1]. Due to the large number of features, univariate filter-based techniques are used since wrapper-based techniques (those which depend on building a classifier to judge feature subsets) and subset evaluation techniques (those which employ statistical tests which operate on complete subsets, rather than individual features) are computationally expensive.

Ensembles have been used primarily for classifiers or learners with success. The use of ensemble learners is still popular for researchers today. In 2010, Schietgat et al. [2] showed that an ensemble of decision trees was superior to a single decision tree classifier in order to accurately assign biological functions to ORFs (Open Reading Frames) in *S. cerevisiae, A. thaliana*, and *M. musculus* datasets. In 2011, Li et al. [3] used ensembles to better classify gene to gene interactions. Their ensemble method was more valid and had more predictive power than a number of existing methods.

However, there is no one perfect ranker that is universally optimal for any given situation [4]. Therefore recently the concept of ensembles has been applied to feature selection in order to increase the stability of the feature subsets generated through no longer relying on a single instance of feature selection. In a study performed by Yu et al. [5], they performed the Kolmogorov-Smirnov test in different bootstrap samples to assign a probability of being selected to each peak. In 2010, Van Landeghem et al. [6] applied ensemble feature selection on biomedical text mining. In 2009, Zhang et al. [7] developed an ensemble technique which creates neural networks on each of the sampled datasets and then uses aggregation to build the best performing network. In 2010, Abeel et al. [8] used ensemble feature selection on the problem of cancer diagnosis and found it led to marked improvement in the stability and classification performance.

However, the ensembles used in these papers only use the data diversity approach (see section IV-A). Though additional steps may be involved, the ensemble portion is achieved through applying the same technique to multiple sampled sets of data. In 2012, our research group [9] introduced other approaches to ensemble feature selection including the hybrid approach.

III. Feature Selection

Feature selection is the process of choosing an optimum subset of features (or genes) and performing the analysis on only this subset of features. Ensemble feature selection comprises two pieces: ensemble approach and feature ranking techniques.

A. Ensemble Approach

Prior to performing ensemble feature selection, one must choose an approach to creating the ensemble. For this paper we chose two different ensemble approaches: data diversity (a commonly used ensemble feature selection method) and hybrid diversity (a new ensemble method created and implemented by our research group). A majority of recent works discussing the use of ensemble feature selection techniques use data diversity when creating ensemble techniques [10], [11].

Data diversity (Figure 1), as its name suggests, achieves its diversity through the use of different sets of data. The process for data diversity occurs in three steps. The first step involves creating the different datasets in order to achieve the desired diversity. This can be achieved through the use of different compiled datasets which use the same set of features or, more commonly, through the creation of multiple sampled datasets derived from the original dataset. The next step is to apply the same feature selection technique on each of these new datasets. Lastly, we aggregate the results from each of the datasets and end with a single feature subset for use in subsequent analysis.

The hybrid approach to ensemble feature selection (Figure 2) begins exactly like the data diversity approach. The first step in the hybrid approach is to create the different datasets (as with data diversity). The following step, like with data diversity, is to apply feature selection. However, the techniques differ in that the hybrid method uses an ensemble of different feature selection techniques (ten total in this paper) and applies each one to $n/10$ of the created datasets, where n is the number of created datasets (which naturally must be a multiple of 10). The final step is aggregating the results from each of the dataset/feature selection combinations and producing a single subset for subsequent analysis.

In order to create the differing datasets for data diversity, we use bagging (sampling with replacement) [12] to create ten, twenty, or fifty sampled datasets (bags) from each of the eleven biomedical datasets used. We use multiples of ten because the hybrid approach requires multiples of the number of rankers used, which is ten in this study.

After the feature ranking occurs the resulting lists (that is, the results of each of the ten, twenty, or fifty bags) are then aggregated into a single ranked feature list. The aggregation is achieved by using the Mean Aggregation technique. Mean Aggregation is simply taking the average rank across all of the ranked feature lists and using that mean value as the

Figure 1. Ensemble: Data Diversity

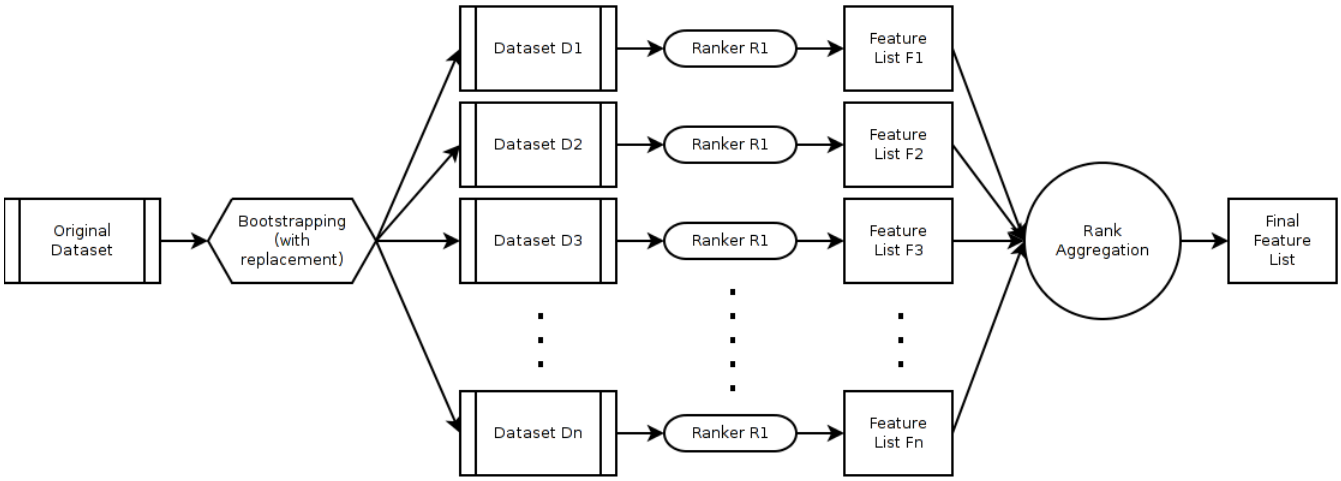
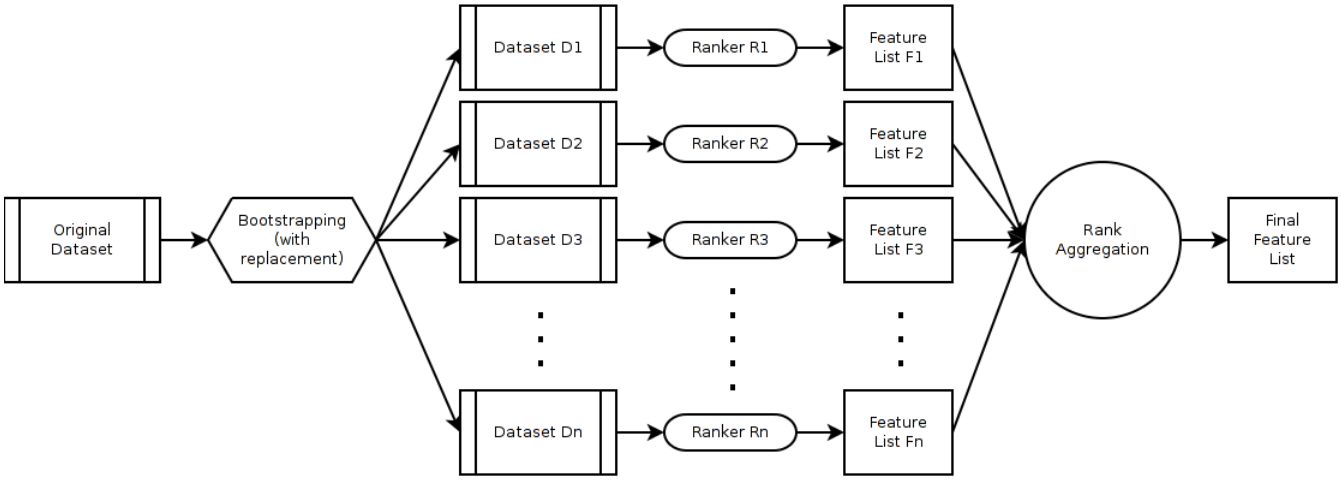


Figure 2. Ensemble: Hybrid



final rank of the feature. This method is quite simple and easy to implement which accounts for its frequent use when using ensemble techniques. Throughout the process of this experiment a total of (26 datasets $\times$ 80 bags (10 bags $+$ 20 bags $+$ 50 bags) $\times$ 10 feature rankers) rankings were performed for data diversity and (26 datasets $\times$ 80 bags (10 bags $+$ 20 bags $+$ 50 bags)) rankings were performed for the hybrid approach for a total of 3,520 rankings. Lastly, a subset of the final list is chosen for subsequent analysis.

B. Feature Ranking Techniques

Both of the ensemble techniques require either a single feature selection technique (data diversity) or an ensemble of them (hybrid). In order to accommodate this requirement we use a set of ten feature selection techniques: Information Gain, Area Under the Receiver Operator Characteristic Curve, Signal-To-Noise, Chi-Squared, ReliefF, Mutual Information, Kolmogorov-Smirnov statistic, Deviance, Geometric Mean, and Area Under the Precision Recall Curve. Each of these techniques are filter-based feature ranking techniques. The reason we only use filter-based feature ranking techniques is that for the large degree of high dimensionality of the datasets, techniques such as filter-based subset evaluation and wrapper-based techniques are too computationally expensive to be of use. Additionally the output of these techniques (a ranked list of features) is easy to understand [1]. These ten feature ranking techniques were chosen because of their consistently good performance [13]

and due to computational constraints of having more rankers.

The ten feature ranking techniques can be split into two categories: Threshold-based Feature Selection (TBFS) Techniques and non-TBFS techniques. Six of the techniques (Mutual Information, Kolmogorov-Smirnov statistic, Deviance, Geometric Mean, Area Under the ROC Curve (denoted as ROC ranker), and Area Under the Precision Recall Curve) fall under the category of TBFS techniques. These feature ranking techniques were proposed and implemented recently by our research group [14]. In TBFS, each attribute is evaluated against the class, independent of all other features in the dataset. After normalizing each attribute to have a range between 0 and 1, simple classifiers are built for each threshold value $t \in [0, 1]$ according to two different classification rules (e.g., whether instances with values above the threshold are considered positive or negative class examples). The normalized values are treated as posterior probabilities: however, no real classifiers are being built. Instead, these ersatz posterior probabilities are used to calculate various classifier performance metrics, and the results of these metrics are the quality of the feature being examined.

The remaining four techniques (Chi-Squared, Information Gain, ReliefF, and S2N) are of the non-TBFS category, and are implemented in the open-source Weka machine learning toolkit [15]. The Chi-Squared test compares the observed distribution of class-feature value pairs to the distribution predicted by a chi-squared random distribution, and those features which are distinct from this null distribution are preferred. Information Gain is based on the entropy inherent in the class-value distribution, and selects features which reduce this entropy when the instances are divided up based on their values. ReliefF decides the relevance of features by seeing how much they vary when taking a randomly-selected instance and comparing its values with those of its nearest hit (instance in same class) and nearest misses (instances in different classes). Finally, Signal-to-Noise finds the signal-to-noise ratio of the feature, which is the ratio of the difference of the mean values for each class to the sum of the standard deviations for each class. For more information on TBFS or the feature selection techniques please refer to Dittman et al [16].

IV. Methodology

In this section, we present the particulars of our case study.

A. Datasets

Table I contains the list of datasets used in our experiment along with their characteristics. The datasets are all DNA microarray datasets acquired from a number of different real world bioinformatics, genetics, and medical projects. As some of the techniques require that there be only two classes, we can only use datasets with two classes (in particular, either cancerous/noncancerous or, in the case of the mulligan-r-nr dataset, relapse/no relapse following cancer treatment). The datasets in Table I show a large variety of different characteristics such as number of total instances and number of features. High dimensionality is clear in the datasets used where the ratio of the number of attributes to the sample size is high.

Table I
Details of the Datasets

Name	Total # of Instances	# of Attributes
ECML_Pancreas	90	27680
lung-Michigan	96	7130
lung-cancer	181	12534
lung50k	400	54614
Lymphoma	96	4027
acute-lymphoblastic-leukemia	327	12559
ovarian_mat	66	6001
lymphoma_mat	77	7130
Brain_Tumor	90	27680
mll-leukemia	72	12583
prostate_mat	89	6001
lung	203	12601
colon50k	400	54614
mulligan-r-pd	126	22284
cns_mat	90	7130
all-aml-leukemia	72	7130
centralNervousSystem	60	7130
colon	62	2001
ovarian-cancer	253	15155
lungcancer-ontario	39	2881
DLBCL-NIH	240	7400
prostate	136	12601
breast-cancer	97	24482
DLBCL	47	4027
mulligan-r-nr	169	22284
bcancer50k	400	54614

B. Feature Subset Size

As the goal of feature selection is to choose an optimum subset of features to perform classification, we must decide on how many of the features to use to build the classification models. Our group decided on twelve feature subset sizes for this experiment. The sizes chosen are as follows: 5, 10, 20, 25, 50, 75, 100, 200, 350, 500, and 1000. These twelve sizes are appropriate according to previous research [14].

C. Similarity Measure

We decided to use consistency index [17] because it takes into consideration bias due to chance. First, we assume that a given dataset has n features. Let T_i and T_j be subsets of features, where $|T_i| = |T_j| = k$. The consistency index [17] is obtained as follows:

$$I_C(T_i, T_j) = \frac{dn - k^2}{k(n-k)}, \tag{1}$$

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

of 1 means the subsets are identical, while a value of 0 means they share the level of overlap which would be predicted from random chance, and a value of -1 indicates that the top half of one feature list contains the same features as the bottom half of the other list.

V. Results

In this experiment, we test the similarity between ranked gene lists generated using varying numbers of iterations in data diversity and hybrid diversity ensemble feature selection approaches. The idea is that if two numbers of iterations create very similar feature subsets then there is no need to perform further research on the number of iterations which takes longer to run. In the process of our experiment three numbers of iterations are tested: 10, 20, and 50. These numbers were chosen due to algorithm constraints (multiples of ten are required for the hybrid approach using an ensemble of ten rankers). The similarity between different numbers of iterations was tested across 26 DNA microarray datasets and 10 univariate feature selection techniques. For each combination of dataset, number of iterations, approach (data or hybrid), feature subset size, and feature ranker (in the case of data diversity), we have one feature subset. We find the similarity of the subsets which share all of these attributes other than number of iterations, and then average the results across all datasets (and all rankers, for data diversity). In data diversity, for every number of iterations, 260 different lists are generated (26 datasets x 10 feature rankers). In the case of hybrid diversity, 26 ranked feature lists are generated for every number of iterations. Tables II and III contain these averages.

When we look at the two tables, we notice that the highest similarity values, across all subset sizes, are achieved when comparing lists generated when using 20 iterations and 50 iterations. The similarity between the lists is between 0.823 and 0.924 in data diversity. In hybrid diversity, the similarity between lists generated using 20 and 50 iterations is between 0.808 and 0.915. This result is expected since as the number of iterations increase, it is expected that more relevant features will be given more weight. Additionally, the lowest similarity values are achieved when we compared the lists generated when using 10 iterations with the lists generated when using 50 iterations. This was also expected since there is a higher gap in the number of iterations used. The similarity between lists generated using 10 and 50 iterations is between 0.756 and 0.888 in data diversity, while it varies between 0.707 and 0.867 in hybrid diversity. This shows that as the number of iterations increase, different features are given more weight.

The results also show the effect of the feature subset size on the similarity of the resulting lists. It is clear that as the subset size increases, the similarity between the lists increases as well. The effect of the subset size shows that while different iterations have different top ranking features, as the subset size increases, the most relevant features will be present in both lists, thus increasing the similarity values when we compare them. We notice that the similarity values increase an additional 0.084 to 0.160. However, the greatest increase happens when we compare the lists generated using 10 iterations and 50 iterations. The similarity values increase an additional 0.132 when the feature subset size is 1000 compared to when it is 5. Additionally, the similarity increases as much as 0.160 in data diversity when the same comparison happens.

Another aspect of our experiment shows that the similarity values are consistently higher across all pairwise comparisons when we use data diversity instead of hybrid diversity. This shows that the data diversity ensemble feature selection techniques will generate more similar feature subsets than the hybrid approach as the number of iterations of feature selection changes. One reason for this is that with the data diversity approach, we only find the similarity of feature subsets generated by a single ranker, while hybrid diversity inherently involves comparing all rankers collectively. This can lead to some random variation in the output of hybrid diversity, which decreases stability.

VI. Conclusion

While ensemble feature selection is used to enhance the stability of the feature selection process, we could not find a study that focuses on how many iterations of feature selection is required. In this work, we perform an experiment comparing different numbers of iterations of feature selection. This is achieved by varying the amount of iterations of feature selection performed when using two different types of ensemble feature selection: data diversity and hybrid diversity. This work helps researchers determine the appropriate number of iterations required without sacrificing the results, thus helping to save computation time that would be required for running more iterations.

The results show that lists generated at 20 iterations can be highly similar to lists generated using 50 iterations and can reach a similarity score of 0.924. These results suggest that there will be very few changes made to the feature subset when additional iterations are performed. This shows that by using 20 iterations we can have highly similar feature subsets with a much smaller computation time (compared to 50 iterations), thus helping researchers get results faster.

We also found additional trends from our experiments. We showed that the similarity between lists generated at different number of iterations increases as the feature subset size increases. We believe this is because as the feature subset size increase it becomes more likely the same gene will appear in both lists. Additionally, we also showed that the similarity values were consistently higher when we used data diversity when compared to hybrid diversity.

Since this is a preliminary study, there is the potential for future work. Future work may study the impact of the

Table II
AVERAGE SIMILARITY: DATA DIVERSITY

Comparison	Feature Subset Sizes											
	5	10	15	20	25	50	75	100	200	350	500	1000
10 Vs. 20	0.826	0.853	0.868	0.861	0.873	0.884	0.889	0.893	0.902	0.907	0.909	0.910
10 Vs. 50	0.756	0.799	0.822	0.821	0.832	0.851	0.856	0.863	0.875	0.881	0.885	0.888
20 Vs. 50	0.823	0.873	0.886	0.886	0.885	0.904	0.909	0.910	0.917	0.922	0.924	0.924

Table III
AVERAGE SIMILARITY: HYBRID

Comparison	Feature Subset Sizes											
	5	10	15	20	25	50	75	100	200	350	500	1000
10 Vs. 20	0.784	0.807	0.805	0.817	0.821	0.851	0.878	0.868	0.878	0.884	0.888	0.890
10 Vs. 50	0.707	0.742	0.748	0.749	0.770	0.810	0.832	0.831	0.849	0.856	0.862	0.867
20 Vs. 50	0.808	0.850	0.856	0.832	0.849	0.870	0.891	0.890	0.901	0.908	0.912	0.915

number of iterations on the classification results. Another possible improvement of this work is using more feature rankers to uncover the effect of the number of iterations on ensemble feature selection techniques.

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.

[2] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Dzeroski, "Predicting gene function using hierarchical multi-label decision tree ensembles," *BMC Bioinformatics*, vol. 11, no. 1, p. 2, 2010.

[3] J. Li, B. Horstman, and Y. Chen, "Detecting epistatic effects in association studies at a genomic level based on an ensemble approach," *Bioinformatics*, vol. 27, no. 13, pp. i222–i229, 2011.

[4] Y. H. Yang, Y. Xiao, and M. R. Segal, "Identifying differentially expressed genes from microarray experiments via statistic synthesis," *Bioinformatics*, vol. 21, no. 7, pp. 1084–1093, 2005.

[5] J. Yu and X.-W. Chen, "Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data," *Bioinformatics*, vol. 21, no. suppl 1, pp. i487–i494, 2005.

[6] S. Van Landeghem, T. Abeel, Y. Saeys, and Y. Van de Peer, "Discriminative and informative features for biomolecular text mining with ensemble feature selection," *Bioinformatics*, vol. 26, no. 18, pp. i554–i560, 2010.

[7] D. Zhang and Y. Wang, "A new ensemble feature selection and its application to pattern classification," *Journal of Control Theory and Applications*, vol. 7, pp. 419–426, 2009, 10.1007/s11768-009-7234-z.

[8] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, February 2010.

[9] W. Awada, T. Khoshgoftaar, D. 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.

[10] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, "A review of ensemble methods in bioinformatics," *Current Bioinformatics*, vol. 5, no. 4, pp. 296–308, 2010.

[11] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *Journal of Biomedical Informatics*, vol. 43, no. 1, pp. 81–87, 2010.

[12] 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.

[13] A. Abu Shanab, T. Khoshgoftaar, R. Wald, and J. Van Hulse, "Comparison of approaches to alleviate problems with high-dimensional and class-imbalanced data," in *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*, aug. 2011, pp. 234–239.

[14] R. Wald, T. Khoshgoftaar, D. Dittman, W. Awada, and A. Napolitano, "An extensive comparison of feature ranking aggregation techniques in bioinformatics," in *Information Reuse and Integration (IRI), 2012 IEEE International Conference on*, Aug. 2012, pp. 377–384.

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

[16] D. J. Dittman, T. M. Khoshgoftaar, R. Wald, and J. Hulse, "Feature selection algorithms for mining high dimensional dna microarray data," in *Handbook of Data Intensive Computing*. Springer New York, 2011, pp. 685–710.

[17] 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.