Ensemble Method for Prediction of Prostate Cancer from RNA-Seq Data

Yongjun Piao, Nak Hyun Choi, Meijing Li, Minghao Piao and Keun Ho Ryu
Database/Bioinformatics Laboratory, Chungbuk National University, Cheongju, South Korea

1 STAGE OF THE RESEARCH

The main idea of our research is to develop an ensemble machine learning algorithm to accurately classify prostate cancer using RNA-Seq data. To date, many studies have focused on predicting prostate cancer using microarray data. These days, RNA-Seq is rapidly being used for cancer studies as an alternative to microarrays. Thus, new machine learning algorithms are needed to analyze RNA-Seq data, which have different characteristics compared to microarray data. The current PhD research has been running for one year and has focused on analyzing existing state-of-the-art normalization methods, gene expression data analysis, and ensemble methods. Besides that, we designed an ensemble feature selection algorithm to select relevant genes from the gene expression data. Moreover, we developed a "digital" gene expression data simulator for evaluating the performance of the proposed algorithms. The next step will be to construct an accurate ensemble prediction model to diagnose prostate cancer. Finally, the model will be fine-tuned based on feedback from medical doctors.

2 OUTLINE OF OBJECTIVES

The goal of this research is to accurately predict prostate cancer using RNA-Seq data. To ensure a systemic approach, the overall research problem is divided into a set of sub-problems: i) comparative analysis of normalization methods, ii) development of RNA-Seq simulators, and iii) ensemble algorithm design and implementation. The specific objectives in each sub-problem are as follows.

- **Provide a Clear Guideline for Choosing the Appropriate Normalization Methods.** Bullard et al. (2010) indicated that the choice of normalization procedure has a decisive effect on identifying candidate genes. The aim of data normalization is to minimize the effects caused by technical variations, such as library size or sequencing depth, gene length, and GC-content. Various normalization methods have been developed. However, it is difficult to decide which normalization methods should be used among the various approaches. Therefore, comparative analysis of these normalization methods will improve the performance of final prediction.
- **Develop a Gene Expression Data Simulator.** Various computational methods have been proposed, and new methods are continuously being developed for identifying candidate genes in gene expression data. One major problem with newly developed approaches is the issue of assessing accuracy and false positive rates in the absence of a so-called "gold standard." Thus, using both simulated and real datasets to evaluate proposed methods will increase the reliability of the prediction model.
- **Propose an Ensemble Classification Algorithm.** In general, ensembles of classifiers provide better classification accuracy than a single predictor. The main characteristic of expression data is high dimensionality. Thus, we need to consider both dimension reduction techniques that identify a small set of genes and ensemble classification methods to achieve better learning performance.

3 RESEARCH PROBLEM

Cancer is a class of complex genetic diseases characterized by out-of-control cell growth. Cancer classification has been a crucial topic of research in cancer treatment. For the last decade, microarray data have been widely used to classify the different types of human cancers (Kim et al., 2013). Recently, the emergence of next-generation sequencing (NGS) technology has brought significant changes in many biological and medical applications (Metzker, 2010; Lee et al., 2013; Shon et al., 2013). Whole transcriptome shotgun sequencing, also known as

Piao Y., Choi N., Li M., Piao M. and Ryu K..
Ensemble Method for Prediction of Prostate Cancer from RNA-Seq Data.

RNA-Seq, is often being used for cancer studies as an alternative to microarrays (Rapaport et al., 2013). Generally, an RNA-Seq analysis pipeline consists of the following procedures. An RNA sample is converted to cDNA fragments or RNA fragments with adapters and sequenced on a high-throughput sequencing platform, such as Illumina or Roche 454. As a result, millions of short reads (30-400 bp) are produced. Next, these short reads are mapped back to a reference genome or transcriptome. After that, the expression levels are estimated for each gene. Then, the count data are normalized. Finally, a machine learning technique is adopted to identify candidate genes. However, several issues still exist in RNA-Seq data analysis. Various normalization methods have been developed for removing the bias of RNA-Seq experiments. However, there is no clear guideline as to how the normalization procedure affects downstream analysis. Thus, it is difficult to decide which normalization methods should be used from among the various approaches. Moreover, there is not much work focused on machine learning approaches with RNA-Seq data for prostate cancer prediction.

In general, ensembles of classifiers provide better classification accuracy than a single predictor. To improve classification accuracy, ensemble methods, also known as classifier combinations, first generate a set of base classifiers from training data and then perform actual classification by combining the results of base classifiers. To achieve better accuracy in the combined set of multiple classifiers, each base classifier should be diverse and independent. When it comes to building each base classifier, ensemble classifier generation methods can be broadly categorized into four groups (Rahman and Verma, 2013): i) by selecting different subsets of instances from a training set to build each base classifier, ii) by choosing different subsets of features from the input features to construct each base classifier, iii) by basing the selection on different categories of class labels to build each base classifier, and iv) by manipulating the learning algorithm. Theoretical and empirical results (Tumer and Oza, 1999) indicate that the most effective method of achieving independence on high-dimensional data is by training base classifiers on different feature subsets (Bryll et al., 2003; Bashir et al., 2012). The basic idea of a feature subset–based ensemble is simply to give each classifier a different projection of the training set (Rokach, 2008). Especially for high-dimensional data, adopting independent feature subsets for ensemble generation has been shown to be more efficient (Rokach, 2010) compared with manipulating the training samples. This may be due to the following: i) a feature subset-based ensemble can perform faster due to the reduced size of input space; or ii) it can reduce the correlation among the classifiers.

4 STATE OF THE ART

4.1 Gene Expression Data Analysis

Much research has been performed on analyzing microarray gene expression data for cancer classification over the past several years. For example, Fujibuchi and Kato (2007) proposed the maximum entropy kernel, which they applied in the field of support vector machine (SVM) classification of microarray data. For classifying a leukemia dataset, Cho and Ryu (2002) proposed an ensemble classifier trained from a subset selected using SNR measurement. Cho and Won (2007) also proposed an ensemble model. Hsu et al. (2011) introduced a hybrid feature selection model based on information gain and F-score, and performed the classification task using SVM. Dettling and Buhlmann (2003) modified the boosting classifiers and applied Wilcoxon's two-sample test to select discriminative genes on the breast and lymphoma dataset. Lee et al. (2005) noted that SVM with a BSS/WSS feature-ranking criterion outperforms other classifiers on a lymphoma dataset. Liu et al. (2010) proposed an ensemble gene selection method based on normalized conditional mutual information and evaluated their method on a central nervous system (CNS), lymphoma and prostate dataset with a Naïve Bayes classifier and a k-nearest-neighbor classifier. Kanna and Ramaraj (2010) used a correlation-based memetic feature selection algorithm to select genes on a CNS and leukemia dataset. Tan and Gilbert (2003) and Yeh (2008) also worked on CNS datasets, and Yang et al. (2006) proposed two gene selection methods that were not affected by unbalanced sample class sizes. Yang et al. (2009) introduced a hybrid feature selection method based on information gain and a genetic algorithm.

4.2 Ensemble Methods

4.2.1 Bagging

Bagging (Breiman, 1996) is a method for generating multiple versions of classifiers and using these to get an aggregated classifier. Each base classifier is generated by different bootstrap samples. Algorithm

1 shows the bagging algorithm (Bauer and Kohavi, 1999). The algorithm takes training data D, inducer I, and the number of bootstrap samples N as input, and then produces an ensemble classifier that is the combination of the classifiers trained from the multiple bootstrap samples. D′ is obtained by repeatedly sampling instances from a data set according to probability distribution (line 2). Since the sampling is done with replacement, some instances may appear several times in the same training set, while others may not. Consequently, N bootstrap samples, D_1, D_2, ..., D_N, are generated, from which a classifier C_i is trained by using each bootstrap sample D_i (line 3). Finally, a combined classifier C* is built from C_1, C_2, …, C_i, and C* predicts the class label of a given instance x by counting votes (line 5).

Algorithm 1 Bagging

Input: training data D, Inducer I, number of bootstrap samples N

1. for i = 1 to N {
2. D′ = bootstrap sample from D (sample with replacement)
3. $C_i = I(D')$
4. }
5. $C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} 1$

Output: Aggregated classifier C*

4.2.2 Boosting

Boosting (Freund and Schapire, 1996) is also a widely used ensemble method developed to improve the performance of learning algorithms that generate multiple classifiers and vote on them. Unlike bagging, boosting assigns a weight to each training instance and may adaptively change the weight at the end of each boosting round. AdaBoost is an improved boosting algorithm, with the pseudo code shown in Algorithm 2. The algorithm takes as input training data D containing m instances, inducer I, and iteration parameter N, and then outputs a combined classifier. Initially, all of the instances are equally assigned the same weight (line 1). Then, the algorithm gradually constructs classifiers by modifying the weights of training instances based on the previous classifier's performance (lines 2-9). This is accomplished by computing the new classifier while putting more emphasis on those objects previously found to be difficult to accurately classify. After generating each classifier, a proportion of the incorrect classification rate is calculated (line 4). If the weighted error is larger than 0.5, the current D′ will be set to a bootstrap sample with weight 1 for every instance. Otherwise, the weight of correctly classified instances will be updated by a factor inversely proportional to the error (lines 6-8). In other words, if the current classifier finds a certain object difficult to classify, then that object will be assigned a greater weight for the next iteration. Conversely, if an object is found to be easy to classify, then it will have less weight in the next iteration. Finally, the classifiers are combined using a weighted voting scheme (line 10).

Algorithm 2 AdaBoost

Input: training data D, size *m*, Inducer I, number of iterations N

1. D′ = D with instance weights assigned at 1
2. for i = 1 to N {
3. $C_i = I(D')$
4. $\varepsilon_i = \frac{1}{m} \sum_{x_j \in D': C_i(x_j) \neq y_j} weight(x)$
5. If $\varepsilon_i > 1/2$, set D′ to a bootstrap sample from D with weight 1 for every instance and go to step 3
6. $\beta_i = \varepsilon_i / (1 - \varepsilon_i)$
7. For each $x_j \in D'$, if $C_i(x_j) = y_j$ then $weigh(x_j) = weight(x_j) \cdot \beta_i$
8. Normalize the weights of instances so the total weight of D′ is *m*
9. }
10. $C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} \log \frac{1}{\beta}$

Output: Aggregated classifier C*

4.2.3 Random Forest

Random forest is an ensemble classification method that consists of multiple unpruned decision trees. Unlike bagging, random forest forms bootstrap samples by randomly partitioning the original feature space instead of using the whole input features. As shown in Algorithm 3, to construct individual decision trees, bootstrap samples are selected from the training instances, with replacement (line 2). Then, a classification and regression tree (CART) algorithm is applied to grow the decision tree. At the node selection stage, it decides the best splitting node from a randomly selected subspace of m features (lines 3-4).

Algorithm 3 Random forest

Input: training data D, number of selected variables m, number of trees N

1. for i = 1 to N {
2. D′ = bootstrap sample from D (sample with replacement)
3. S′ size of m= S (S′ will be randomly selected from original input space)
4. C_i = I(D′, S′) (I: Classification and regression tree)
5. }
6. $$C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} 1$$

Output: Aggregated classifier C*

5 METHODOLOGY

5.1 Normalization

As mentioned above, the choice of normalization procedure has a decisive effect on identifying differentially expressed genes. The aim of data normalization is to minimize the effects caused by technical variations, such as library size or sequencing depth, gene length, and GC-content. In general, a larger sequencing depth results in higher counts, which means that the observed counts are not directly comparable between different samples (Soneson and Delorenzi, 2013). Likewise, long genes tend to be mapped to a larger number of reads. These systematic variations make it difficult to capture true differential expression. Several normalization methods have been developed, such as Total Count, Upper Quality, Median, Trimmed Mean of M values (TMM), Quartile, the Reads Per Kilobase per Million mapped reads (RPKM), and RSEM, to reduce the biases existing in RNA-Seq analysis. Thus, these methods will be carefully re-evaluated by comparing the correlations with qRT-PCR data.

5.2 Simulation

To obtain a reliable prostate prediction model, the proposed methods will be tested on both real and simulated data. Thus, we developed a gene expression data simulator. As input, the simulator simply takes several parameters, such as number of samples in condition 1, the number of samples in condition 2, the number of genes, the number of differentially expressed genes, the number of co-expressed genes, the number of highly expressed genes, and the number of zero count genes, distribution types, and distribution parameters. Then, simulated data is generated as output.

5.3 Ensemble Gene Selection

In our previous work (Piao et al., 2012), we proposed a hybrid feature selection algorithm and demonstrated that there are lots of feature subsets with good discriminative capability, and the proposed algorithm was more efficient than FCBF and other feature selection mechanisms for gene expression data analysis. While the goal of our previous work was to find the best feature subset that is most relevant to the target, there needs to be an additional goal of finding a set of feature subsets. Therefore, we extend our previous work to generate a set of feature subsets by considering the relevance and redundancy of the features.

5.4 Ensemble Construction

Over the past few years, SVM has been widely used for classification because of its good performance on high-dimensional data. SVM was developed to solve the problems occurring in applications like handwritten digit recognition, object recognition, text classification, cancer diagnosis, and bioinformatics. Hence, we use SVM as the base classifier in our ensemble method. The goal of SVM is to find a hyper-plane with a maximal margin (the distance between two groups of data points) as defined and illustrated in Figure 1. Given some data points that are assumed to be divided into two groups (circles and squares), the hyper-plane can be written as:

$$w \cdot x + b = 0 \quad (3)$$

$$w \cdot x + b = \pm 1 \ , \ margin = \frac{2}{\| w \|} \quad (4)$$

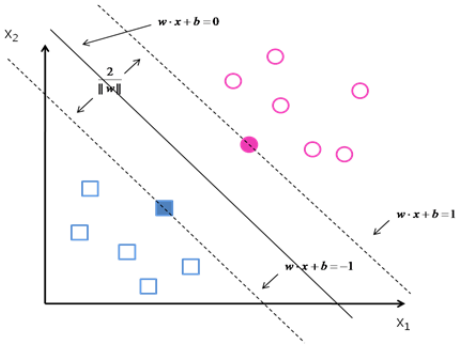

Figure 1: Example of support vector machine.

Hence, the learning task in SVM can be formalized as the following constrained optimization problem:

$$\min_w \frac{\| w \|^2}{2} \quad (5)$$

$$\text{subject to } y_i(w \cdot x_i + b) \geq 1, i = 1, 2, ..., n \quad (6)$$

This is also known as a convex optimization problem, which can be solved by using the standard Lagrange multiplier method:

$$L_p = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{N} \sigma_i (y_i (w \cdot x_i + b) - 1) \quad (7)$$

where parameters σ_i are called the Lagrange multipliers. With the Lagrange multipliers, the decision function can be written as follows:

$$f(x) = \text{sgn}(\sum_{i=1}^{n} \sigma_i y_i K(x_i, x) + b) \quad (8)$$

Additionally, the results of each classifier are combined by majority voting, and classification of unknown data is performed based on the class label to obtain the most frequent votes. The mathematical function of our ensemble method with k classifiers can be written as:

$$\text{class}(x) = \arg\max(\sum_{k} (f_k(x), c_i)) \quad (9)$$

6 EXPECTED OUTCOME

At the end of this PhD research, a new ensemble classification method will be available for predicting prostate cancer from RNA-Seq data. Moreover, a complete gene expression data simulator will be developed. The simulator may help to research non-parametric methods for cancer classification. The research will lead to new insights for understanding prostate cancer by identifying candidate genes from high-dimensional gene expression data. If this is proven successful, our approach could be applied to other types of disease.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2008-0062611) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923).

REFERENCES

Bullard, J., Purdom, E., Hansen, K., Dudoit, S., 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11:94.

Kim, Y., Yoon, H., Kim, J., Kang, H., Min, B., Kim, S., Ha, Y., Kim, I., Ryu, K., Lee, S., Kim, W., 2013. HOXA9, ISL1 and ALDH1A3 methylation patterns as prognostic markers for nonmuscle invasive bladder cancer: array-based DNA methylation and expression profiling. International Journal of Cancer 133, 1135-1143.

Metzker, M., 2010. Sequencing technologies – the next generation. Nature Reviews Genetics, 11, 31-46.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C., Socci, N., Betel, D., 2013. Comprehensive evaluation of differential gene expression analysis methods for RNA-Seq data. Genome Biology, 14:R95.

Rahman, A., Verma, B., 2013, Ensemble Classifier Generation using Non–uniform Layered Clustering and Genetic Algorithm. Knowledge-Based System 43, 30-42.

Tumer, K., Ghosh, J., 1996. Classier combining: analytical results and implications. Proc. Nat'l Conf. Artificial Intelligence, Portland, Ore, 126-132.

Tumer, K., Oza, N., 1999. Decimated input ensembles for improved generalization. International Joint Conference on Neural Network 5, 3069-3074.

Bryll, R., Gutierrez-Osuna, R., Quek, F., 2003. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets, Pattern Recognition 36, 1291-1302.

Rokach, L., 2006. Genetic algorithm-based feature set partitioning for classification problems, Pattern Recognition 41, 1676-1700.

Rokach, L., 2010. Ensemble-based classifiers. Artif. Intell. Rev. 33, 1-39.

Fujibuchi, W., Kato, T., 2007. Classification of heterogeneous microarray data by maximum entropy kernel. BMC Bioinformatics 8, 267-277.

Cho, S., Ryu, J., 2002. Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proceedings of the IEEE, 90(11), 1744-1753.

Bashir, M., Lee, D., Li, M., Bae, J., Shon, H., Cho, M., Ryu, K., Trigger Learning and ECG Parameter Customization for Remote Cardiac Clinical Care Information System. IEE Transactions on Information Technology in Biomedicine, 16, 561-571.

Cho, S., Won, H., 2007. Cancer classification using ensemble of neural networks with multiple significant gene subsets. Applied Intelligence 26, 243-250.

Hsu, H., Hsieh, C., Lu, M., 2011. Hybrid feature selection by combining filters and wrappers. Expert System with Applications 38, 8144-8150.

Dettling, M., Buhlmann, P., 2003. Boosting for tumor classification with gene expression data. Bioinformatics 19 (9), 1061-1069.

Lee, J., Lee, J., Park, M., Song, S., 2005. An extensive comparison of recent classification tools applied to microarray data, Comput. Statist. Data Anal. 48, 77-87.

Liu, H., Liu, L., Zhang, H., 2010. Ensemble gene selection for cancer classification. Pattern Recognition 43, 2763-2772.

Lee, D., Ryu, K.S., Bashir, M., Bae, J., Ryu, K., Discovering Medical Knowledge using Association Rule Mining in Young Adults with Acute Myocardial Infarction. Journal of Medical Systems, 37.

Kannan, S., Ramaraj. N., 2010. A novel hybrid feature selection via symmetrical uncertainty ranking based local memetric search algorithm. Knowledge-Based Systems 23, 580-585.

Tan, A., Gilbert, D., 2003. Ensemble machine learning on gene expression data for cancer classification. Bioinformatics 20, 3583-3593.

Yeh, J., 2008. Applying data mining techniques for cancer classification on gene expression data. Cybernetics and Systems: An International Journal 39, 583-602.

Yang, K., Cai, Z., Li, J., Lin, G., 2006. A stable gene selection in microarray data analysis. BMC Bioinformatics 7:228.

Yang, C., Chuang, L., Yang, C., 2009. IG-GA: A hybrid filter/wrapper method for feature selection of microarray data, Journal of Medical and Biological Engineering 30 (1), 23-28.

Breiman, L., 1996. Bagging predictors, Machine Learning 24, 123-140.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: bagging, boosting and variants, Appears in Machine Learning 36, 105-139.

Freund, Y., Schapire, R., 1996. Experiments with a new boosting algorithm, International Conference on Machine Learning, 148-156.

Soneson, C., Delorenzi, M., 2013. A comparison of methods for differential expression analysis of RNA-Seq data. BMC Bioinformatics, 14:91.

Piao, Y., Piao, M., Park, K., Ryu, K., 2012. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data, Bioinformatics, 28, 3306-3315.

Shon, H., Kuk, H., Whan, B., Ah, KIM., Lee, J., Ryu, K., N-terminal pro-B-type natriuretic peptide as prognostic marker for patients of non ST-segment elevation myocardial infarction. J. Cent. South Univ., 20, 2226-2232