An Iterative Algorithm of Key Feature Selection for Multi-class Classification

Daeun Jung and Hyunggon Park
Department of Electronic and Electrical Engineering
Ewha Womans University
Seoul, Korea
daeun.jung@ewhain.net, hyunggon.park@ewha.ac.kr

***Abstract*—In this paper, we propose an iterative algorithm of key feature selection for multi-class classification problems, where the data includes a large number of features but the amount of data is limited. For efficient classification, the proposed algorithm first extracts a set of key feature candidates based on Boruta algorithm and then iteratively adopts conventional machine learning based classification algorithms to determine key features. Simulation results show that the proposed algorithm can effectively determine key features, leading to improved classification accuracy compared to direct adoption of multi-class classification algorithms.**

Keywords—multi-class classification; machine learning; omics data; feature extraction; feature selection

I. INTRODUCTION

As the machine learning algorithms have been developed and show the effectiveness in practical applications, they have been deployed in a variety of industries. In particular, it is shown that machine learning algorithms can be very effective in data analytics, where inference or predictions can be made by well approaches such as regressions or classifications [1]. In this paper, we consider omics data, which is biological data set with a large number of molecular dimensions, such as genome, proteome, transcriptome, etc. [2]. An illustrative example of the omics data is shown in Fig. 1. This implies that the omics data inherently include a huge number of features. While it is inevitable to understand the interactions in genomes, proteomes, transcriptomes, etc. [2], it is significantly challenging to analyze the omics data. This become even more challenging if the omics data has multiple groups.

Our focus of this paper is on analyzing the omics data and extracting key features for multiple group classification. If the goal is to simply solve the multi-class classification problems, we may adopt algorithm adaptation techniques [3],[4] by extending several classification algorithms such as support vector machine (SVM), random forest, linear regression, K-nearest neighbors, Naïve Bayes, decision tress, etc. that are basically designed for binary class classification into multi-class classification [5]. However, this approach may not explicitly show which features are key for multi-class classification.

We propose an iterative algorithm that both feature extraction process and feature evaluation process, in order to solve multi-class classification problem as well as identify key

Features (genes)

Data points

Class	RBM47	TTC26	UBA6	KIAA1598	ARHGEF37	ILVBL	PLEKHG3	SSC5D
NM	0.649338	-5.19358	2.31718	1.3735	-5.30624	2.46925	-4.33641	-3.52004
NM	2.00035	-2.41425	2.2309	0.788849	-3.46046	1.95065	-5.33386	-3.77726
NM	1.42062	0.732737	0.125049	0.514269	-0.69913	0.15233	-4.1182	0.084689
NM	2.9477	-0.62715	2.13276	0.879399	-3.58127	0.849304	-5.09208	-3.51096
NM	2.50203	-0.67889	2.4103	2.08709	-4.11236	1.46412	-3.69574	-3.62509
NM	2.02746	-1.64112	1.6627	1.47838	-2.21354	1.67401	-3.52425	-4.11289
NM	1.36052	-0.45542	1.73973	0.624249	-3.05133	2.74077	-3.03955	-4.44097
NM	0.24694	-5.96444	2.38565	0.720784	-4.76066	0.792288	-4.72079	-2.20116
NM	1.55198	-0.64969	2.52934	1.85306	-1.28852	2.18104	-3.24259	-0.86
NM	2.37824	0.199426	1.94369	1.26013	-2.5338	1.65448	-2.62308	-1.61276
LM	0.560702	-2.68546	2.4969	-0.80661	-2.36141	0.448305	-5.21313	-0.87745
LM	1.02538	-1.92189	1.84335	1.29495	-0.39962	1.40026	-5.24296	-0.84506
LM	0.650942	-1.1295	1.14373	1.28349	-2.9202	1.65793	-2.4514	-0.76579
LM	1.54541	-1.8131	1.59697	0.74696	-1.993	1.11007	-5.66087	-1.10636

FIG. 1. AN EXAMPLE OF OMICS DATA SET (PART). THE OMICS DATA IN THIS EXAMPLE INCLUDES THREE GROUPS (NM, LM, AND M) WITH 6,304 FEATURES. THE TOTAL NUMBER OF DATA POINTS IS 30.

features. Feature selection is the process where the features that contribute most to prediction variable or output are determined and selected [1], [5]. This process may remove irrelevant or partially relevant features which can negatively impact on the performance. It is generally known that the feature selection is the first and the most important step for classification model design. For example, several machine learning algorithms such as regularized trees [6], decision tree [7] include the feature extraction process. A good feature selection should be able to avoid overfitting, improve accuracy and reduce training time, thereby leading to efficient decision making with less data and time requirements. In this paper, we adopt the Boruta algorithm [8], [9] as a feature selection algorithm that can reduce the number of features included in the data set and return a set of significantly reduced number of key feature candidates.

In the set of key feature candidates, key features can be identified and determined by adopting machine learning based classification algorithms. In particular, some feature combinations randomly selected from the set of key feature candidates are evaluated by machine learning algorithms. This process is performed iteratively until key features are identified. Note that the proposed algorithm does not need to modify existing machine learning algorithms. Rather, the propose approach can improve the performance of the algorithms by significantly reducing the number of features that need to be considered. The potential issues incurred by deploying binary classification machine learning algorithms in multi-class classification problems have been tackled by several approaches [10]. In this paper, we use Naïve Bayes, Random Forest, Linear Discriminant Analysis (LDA) as machine learning algorithms

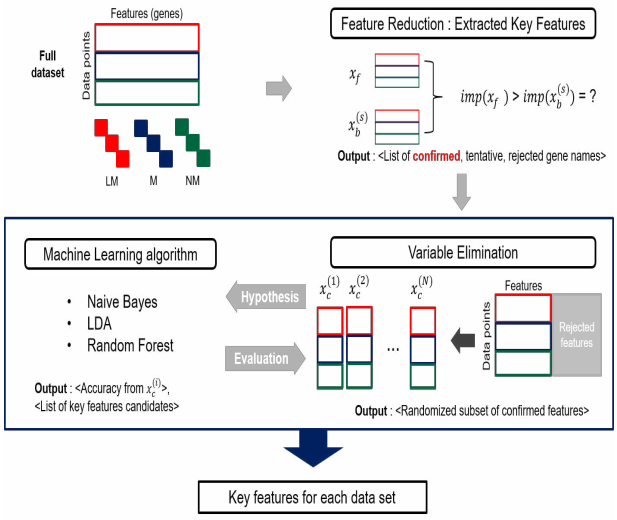

FIG. 2. OVERVIEW OF PROPOSED ALGORITHM

for multi-class classification. An overview of the proposed algorithm is shown in Fig. 2.

This paper is organized as follows. In Section II, we describe the proposed algorithm for multi-class classification problem. Then, we present experiment results that show the effectiveness of the proposed algorithm for real omics data collected from a cohort in Section III. Finally, the conclusions are drawn in Section IV.

II. PROPOSED ALGORITHM

The proposed algorithm consists of feature extraction process and feature evaluation process. The proposed algorithm does not need to modify machine learning algorithms so that we can easily adopt algorithms for classification. This can reduce the time and memory required for classification compared to the multi-class classification algorithms with the extracted features. Finally, the proposed algorithm can provide key features for classification.

A. Feature Extraction Algorithm

In the proposed algorithm, we deploy the Boruta algorithm for feature extraction. It computes the importance considering of all the features. Then, it iteratively continues comparing existing feature importance and achievable importance at random.

In our implementation, we repeatedly apply Boruta algorithm 3,000 times and then extract 50 key features based on the variable importance measure. As the number of high importance features may vary for different data sets, it is desirable to reduce the number of features, which can be achieved by using the Wrapper method [11], [12]. The Wrapper method is the procedure for deciding subset among the characteristic universal set and finding feature set based on training model comparison constructed by adding or subtracting features. Finally, it is possible to model a generalized pattern while avoiding overfitting problem.

B. Machine Learning based Feature Evalution

Given the features extracted from the feature extraction algorithm, several machine learning algorithms are deployed. Specifically, we use the following algorithms.

- Naïve Bayes: Naïve Bayes assumes that every feature of the data is equivalent and independent. It also assumes that all features contribute individually to the probability if the size of data set is small during training. Naïve Bayes classifiers are based probability models and can be trained efficiently in supervised learning [13], [14].
- LDA: LDA is a classification model by making decision of data distribution during learning. To separate each cluster properly, LDA appoints straight line to divide classes by keeping away center of each class and reducing their standard deviation. Then, a subset of the features becomes the set of selected characteristics. The dimension can be reduced by properly selecting features [15], [16].
- Random Forest: Random Forest can be used in classification or regression by training multiple decision trees. In classification, a decision tree is created randomly at the time of learning. It assembles randomly generated decision trees that make decisions based on different feature sets, allowing them to vote on the most common class [17], [18].

Note that these algorithms are adopted in order to evaluate the features selected by feature elimination process. This process is repeated until key features are determined.

III. EXPERIMENT RESULTS

In order to evaluate the performance of the proposed approach, we consider the omics data for breast cancer cells.

A. Experiment Setup

The data set with 30 data points consist of three groups, labeled as NM, LM and M, which represent no metastasis, late metastasis and metastasis, respectively. Each group has 10 data points from breast cancer cells. The data we used represent gene expression patterns in breast cancer cell lines. Each class is characterized by cell lines and therapy methods. One group (LM / NM / M) is used for three-class classification and the numbers of features in this group is 6,332. Since the number of data points for each class is 10, the data set of groups (LM / NM / M) has 30 data points in total and 6,332 features as shown in Fig. 1.

The Boruta algorithm used in the feature extraction process determines 50 key feature candidates, where they are further reduced into 8-11 features by Wrapper method. [11], [12]. The Wrapper method generates feature subsets by adding each features selected from the set of key feature candidates determined by Boruta algorithm. Then, the features included in each feature subset are used in machine learning based classification. The importance of subset features is determined based on the accuracy of 90% and features highly ranked by the importance are finally extracted. The number of key features multi-class classification algorithm is 10–24 depending on deployed machine learning algorithms described in Section II.B.

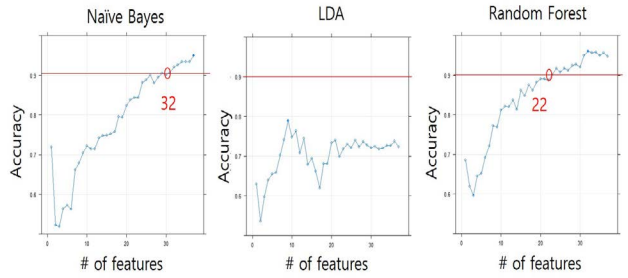

FIG. 3. RESULTS FOR WRAPPER METHOD BY DIFFERENT MACHINE LEARNING ALGORITHM

B. Experiment Results

An illustrative example of results from the wrapper method is shown in Fig. 3. The set of key feature candidates (37 features) selected from Boruta algorithm is used as an input of the wrapper method. The threshold for accuracy is set as 0.9 that is shown as a red line in Fig. 3. This is the result of evaluating the accuracy of the set with the machine learning algorithm while increasing the number of features one by one. For example, Naïve Bayes, there are 32 features above the threshold. Key features set are found through this method iteratively.

The experiment results are summarized in Table I and Table II, which present the classification accuracy obtained from 1,000 independent experiments for different settings. Table I shows the performance of the proposed algorithm while the performance of direct adoption of machine learning algorithms without feature extraction.

It is clearly shown that the overall accuracy of the proposed approach outperforms the direct adoption of multi-class classification machine learning algorithms without feature extraction. For example, machine learning based classification without key features shows 48%-65% accuracy. However, the proposed algorithm that uses extracted key features shows 80%-93% accuracy. Hence, we can conclude that the performance of machine learning algorithms can be significantly improved by using key features. It should also be noted that the proposed algorithm can explicitly identify the key features, unlike exiting multi-class classification solutions.

IV. CONCLUSIONS

In this study, we propose an iterative algorithm of key feature selection for multi-class classification based on feature extraction process and machine learning based evaluation process. The proposed algorithm is evaluated by the omics data actually collected by cohort and it is shown from the experiment results that the proposed algorithm outperforms the existing multi-class classification algorithms in terms of accuracy. Furthermore, the proposed algorithm is able to identify key features for classification.

TABLE 1. PERFORMANCE OF THE PROPOSED ALGORITHM

Accuracy	Naïve Bayes	LDA	Random Forest
Proposed	0.833	0.800	0.928
Without Feature Extraction	0.619	0.480	0.652

ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00024, Supervised Agile Machine Learning Techniques for Network Automation based on Network Data Analytics Function) and supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. NRF-2017R1A2B4005041).

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.

[2] C. Manzoni, D. A. Kia, J. Vandrovcova, J. Hardy, N. W. Wood, P. A. Lewis and R. Ferrari, "Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences", *Briefings in Bioinformatics*, vol. 19, pp. 286–302, Mar. 2018.

[3] G. Lannoy, D. Francois and M. Verleysen, "Class-Specific Feature Selection for One-Against-All Multiclass SVMs", *European Symposium on Artificial Neural Networks (ESANN)*, pp. 27-29, Apr. 2011.

[4] D. G. Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. B. Rudloff, A. Conesa and J. Tegnér, "Data integration in the era of omics: current and future challenges", *BMC Systems Biology*, vol. 8, Mar. 2014.

[5] M. Aly, "Survey on Multiclass Classification Methods", *Technical report*, 2005.

[6] H. Deng, G. Runger, "Feature Selection via Regularized Trees", *Procs. IEEE International Joint Conference on Neural Networks (IJCNN)*, Jun. 2012.

[7] F. Degenhardt, S. Seifert and S. Szymczak, "Evaluation of variable selection methods for random forests and omics data sets", *Briefings in Bioinformatics*, vol. 20, pp. 492-503, Mar. 2019.

[8] M. B. Kursa and W. R. Rudnicki, "Feature Selection with the Boruta Package," *Journal of Statistical Software*, vol. 36, Sep. 2010.

[9] V. Fortino, P. Kinaret, N. Fyhyquist, H. Alenius, D. Greco, "A Robust and Accurate Method for Feature Selection and Prioritization from Multi-class Omics Data", *PLoS ONE*, vol. 9, Sep. 2014.

[10] T. Li, C. Zhang and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", *Bioinformatics Advance Access*, vol. 20, Apr. 2004.

[11] M. A. Jayaram, A. G. Karegosda and A. S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", *International Journal of Computer Applications*, vol. 1, pp. 13-17, 2010.

[12] R. Kohavi and G. John, "Wrappers for feature subset selection", *Artificial intelligence*, vol. 97, pp. 273-324, 1997.

[13] I. Rish, "An empirical study of the naive Bayes classifier", *workshop on empirical methods in artificial (IJCAI)*, Jan. 2001.

[14] S. A. Pattekari and A. Parveen, "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES", *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, pp. 290-294, 2012.

[15] Z. Qiao, L. Zhou and J. Z. Huang, " Sparse Linear Discriminant Analysis with Applications to High Dimension Low Sample Size Data", *International Journal of Applied Mathematics (IAENG)*, vol. 39, Feb. 2009.

[16] J. H. FRIEDMA, "Regularized Discriminant Analysis", *J. American Statistical Association*, vol. 84, pp. 165-175, Mar. 1989.

[17] C. Nguyen, Y. Wang, H. N. Nguyen, "Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic", *J. Biomedical Science and Engineering,* vol. 6, pp. 551-560, May. 2013.

[18] L. Breiman, "Random Forests", *J. Machine Learning*, vol. 45, pp. 5-32, Oct. 2001.