A algorithm for multi-class classification using PLS-DA

Daeun Jung and Hyunggon Park
Multiagent Communications and Networking Laboratory
Ewha Womans University, Seoul, Republic of Korea
daeun.jung@ewhain.net, hyunggon.park@ewha.ac.kr

***Abstract*—In this paper, we propose an algorithm of finite number of key feature selection for multi-class classification problems, where the data has significantly large number of features than the amount of data. We adopt conventional machine learning algorithms for data classification. In order to improve the classification accuracy by considering the characteristics of the data set, we propose to use statistical transformation method such as Partial least squares-discriminant analysis (PLS-DA) to determine a predefined number of key features. Experiment results show that the proposed algorithm can effectively determine key features with the maximum number of features, leading to improved classification accuracy compared to direct adoption of multi-class classification algorithms.**

***Keywords*—multi-class classification, machine learning, class decomposition, omics data, feature selection.**

I. Introduction

Recently, data classification has become important in a variety of practical applications including medical treatment. For automatic data classification, several machine learning algorithms are very actively adopted. Classification problems consider data sets that have generally more data points than variables. Such multiple class classification problems can be solve by adopting machine learning [1] based for classification algorithms such as support vector machine [2], random forest [3], [4], linear regression [5], K-nearest neighbors [6], Naïve Bayes [7], decision tress [8], etc. These machine learning algorithm can be used as feature selection or data classification.

While machine learning algorithms can be directly adopted for feature selection and data classification [9], the classification performance may be degraded or guaranteed as they often depend on randomness. This becomes severe if there are significantly more features than data points. In order to improve the classification accuracy based on machine learning algorithms, only key features among all the features should be considered. To determine key features, we propose to adopt statistical methods such as Principal Component Analysis (PCA), Wrapper method, PLS for feature extraction and selection. The statistical methods can determine which features are more for data set. We propose to use Partial least squares-discriminant analysis (PLS-DA) [10] for feature extraction,

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).

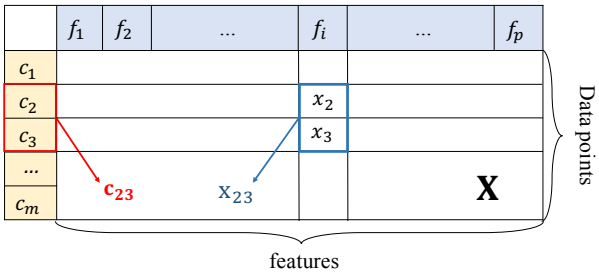


Fig. 1. A representation of data set.

which is a statistical transformation method that maximizes the variance of feature and covariance of response variable at once. For feature extraction, PLS-DA can result in importance of each feature for target classes. The PLS-DA is known to be useful in supervised classification in case where data set has significantly much more features than the data points [11], [12].

If the data set is consist of multiple classes, PLS-DA may not be efficient for feature extraction. Hence, in this paper, we propose to decompose the data set with multiple classes into multiple data sets with binary classes, referred to as *split* group of data. Then, features can be extracted from each split group of data based on PLS-DA. Since there are multiple split groups of data, features extracted from the groups of data should be merged, such that key features for the entire data set can be determined. In order to determine key features, we propose to use the importance obtained from PLS-DA for individual features. In case where only a target maximum number of features are allowed, we also propose an algorithm to further reduce the number of key features based on normalization function of the importance values.

With the features selected by the proposed algorithm, several machine learning based classification algorithms can be deployed for evaluation purpose. The proposed algorithm does not need to modify machine learning algorithms so that we can easily adopt existing algorithms for classification.

This paper is organized as follows. In Section II, we describe the problem with data set that we consider. Then, we propose an algorithm to find the key features. We present experiment results that show the effectiveness of the proposed algorithm for omics data actually collected from cohorts in Section III. Finally, the conclusions are drawn in Section IV.

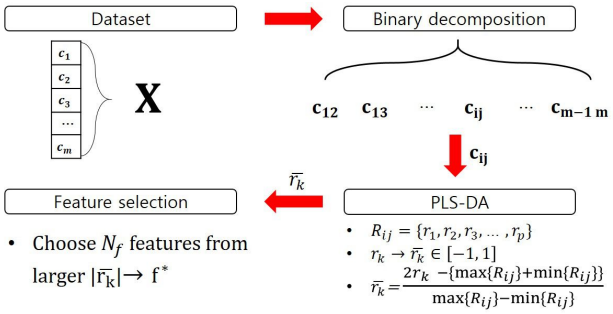


Fig. 2. Illustration for the overall algorithm

II. Problem Setup and Proposed Algorithm

In this paper, we consider a data set which is represented by $n \times p$ matrix $\mathbf{X}$, where n and p denote the number of data points and the number of features, respectively. We consider the case where $n \ll p$ in this paper, i.e., there are significantly more features than data points. Each data point x_i can be classified into a class in multiple classes, i.e.,

$$\mathbf{C} = \{c_1, c_2, \dots, c_m\}.$$

where $c_i (i \leq m)$ is the ith class in m classes. The distinct features included in $\mathbf{X}$ are denoted by

$$\mathbf{F} = \{f_1, f_2, \dots, f_p\}$$

where $f_i (i \leq p)$ represents the ith feature in p features. The subset of $\mathbf{F}$ is denoted by $\mathbf{f}$.

The goal is to find a set of at most $N_f (\leq p)$ features in $\mathbf{F}$ that maximize the accuracy η, defined as

$$\eta(\mathbf{f}) = \frac{TP + TN}{TP + TN + FP + FN} \times 100(\%) \qquad (1)$$

which is a function of features included in $\mathbf{f}$. P and N represent classes for Positive and Negative, respectively, and T and F denote the cases where the predicted and actual labels are matched or not, respectively. For example, $TP + TN$ represents the number of data points that are correctly classified. Hence, the accuracy η represents the proportion of the number of correct predictions to the number of entire predictions.

Therefore, the set of optimal features $\mathbf{f}^* \subseteq \mathbf{F}$ that maximizes the accuracy η can be determined as

$$\mathbf{f}^* = \underset{\mathbf{f} \subseteq \mathbf{F}, |\mathbf{f}| \leq N_f}{\operatorname{argmax}} \ \eta(\mathbf{f}). \qquad (2)$$

An overview of the algorithm proposed to find the solution to (2) is shown in Fig. 2. The proposed algorithm divides the entire data in $\mathbf{C}$ into $\binom{m}{2}$ groups of data included in two classes. A split group of data for classes c_i and c_j is denoted by $\mathbf{x}_{ij} = \{x_i, x_j | x_i \in c_i, x_j \in c_j\}$, $i, j \in \{1, \dots, k\}, i \neq j$. Then, the importance of each feature is evaluated based on the PLS-DA, which results in the set of importances of the features for $\mathbf{c}_{ij}$, i.e.,

$$\mathbf{R}_{ij} = \{r_1, r_2, \dots r_p\}$$

where $r_k (1 \leq k \leq p)$ represents the importance of feature k. Since the PLS-DA in binary classification generates importance of each feature to two classes as positive or negative, $f_k \in \mathbf{f}_i$ if $r_k > 0$. This enables all the features in $\mathbf{F}$ to be grouped as a key feature for class c_i or c_j, which determines the key feature sets $\mathbf{f}_i$ and $\mathbf{f}_j$ for c_i and c_j, respectively, where $\mathbf{f}_i \cup \mathbf{f}_j = \mathbf{F}$. Hence, the result of the PLS-DA for $\mathbf{x}_{ij}$ $(1 \leq i, j \leq p, i \neq j)$ is summarized as the pair $\{\mathbf{R}_{ij}, \mathbf{x}_{ij}\}$, which quantifies the impact of individual features on correct classification of the data in $\mathbf{c}_{ij}$.

We now have to determine N_f number of features from $\mathbf{f}_i$ and $\mathbf{f}_j$. Since more important features have larger absolute value of importance in the results from PLS-DA, we can choose top N_f features in $\mathbf{f}_i$ and $\mathbf{f}_j$ based on normalized importance. This eventually determines $\mathbf{f}^*$. The process to determine $\mathbf{f}^*$ features are described as follows.

- Given $r_k \in \mathbf{R}_{ij}$ for all i and j, normalize importance: mapping $r_k \rightarrow \overline{r}_k \in [-1, 1]$ by function

$$\overline{r}_k = \frac{2r_k - \{\max\{\mathbf{R}_{ij}\} + \min\{\mathbf{R}_{ij}\}\}}{\max\{\mathbf{R}_{ij}\} - \min\{\mathbf{R}_{ij}\}}$$

- Choose N_f features from larger $|\overline{r}_k|$.

TABLE I. Performance Comparison ($N_t = 20$)

ML Algorithm		Proposed	3-class
Naïve Bayes	η	**99.7**	89.7556
	$n(\mathbf{f}^*)$	15	20
LDA	η	**95.4222**	87.3333
	$n(\mathbf{f}^*)$	7	11
Random Forest	η	**98.9**	91.6222
	$n(\mathbf{f}^*)$	15	18

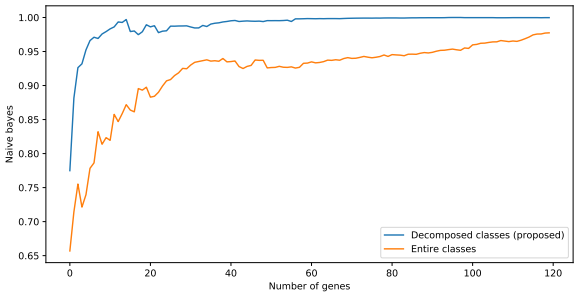
III. Experiment Results

A. Experiment Setup

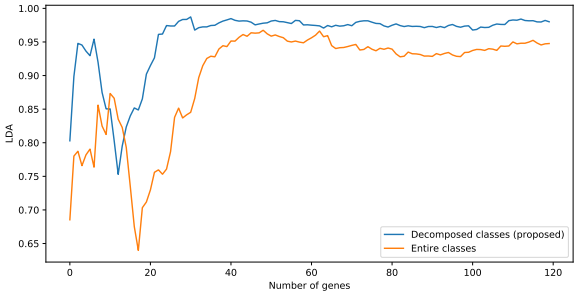
In order to evaluate the performance of the proposed approach, we use the omics data [13], [14] for breast cancer cells. The data set has 30 data points with three groups, labeled as NM, LM and M, which represent no metastasis, late metastasis and metastasis, respectively. Each group has 10 data points and the numbers of features in each group is 7,757. N_f is set as 20. Since the data consists of three classes, the split groups are $\mathbf{c}_1 = \{NM, LM\}$, $\mathbf{c}_2 = \{NM, M\}$ and $\mathbf{c}_3 = \{LM, M\}$. The N_f key feature set $\mathbf{f}^*$ is determined as described in Section II. The performance of the proposed algorithm is measured by the average accuracy computed from 1,000 independent experiments with different settings.

B. Experiment Results

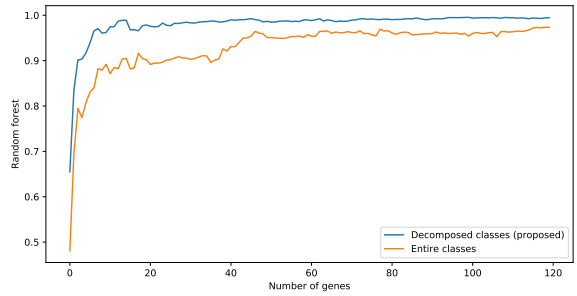
The performance comparison of the proposed algorithm with class decomposition and the classification with the entire classes is shown in Fig. 3. We observe that the proposed algorithm almost outperforms the classification for the entire data set in terms of the average classification accuracy over all the number of features (genes). The proposed algorithm shows the classification accuracy 95%-99%, while the classification for the entire classes shows 87%-97% accuracy as shown in Table I. It should be noted that that the number of features is



(a)



(b)



(c)

Fig. 3. Classification accuracy.

limited such that it does not exceed N_f.

Hence, it can be concluded that the performance of machine learning algorithms can be significantly improved by considering key features.

IV. Conclusion

In this paper, we propose an algorithm for feature selection in multiple class classification problem based on PLS-DA with split groups of binary classes. The proposed algorithm divides the entire data set of multiple classes into multiple groups of split data included in binary classes. The importance of the features in each group is evaluated by PLS-DA and then the predefined number of key features for the multiple classes are determined. The experiment results show that the features selected by the proposed algorithm outperforms the features based on absolute size in terms of classification accuracy.

References

[1] M. Aly, "Survey on multiclass classification methods," *Neural Netw*, vol. 19, pp. 1–9, 2005.

[2] G. de Lannoy, D. François, and M. Verleysen, "Class-specific feature selection for one-against-all multiclass svms," *European Symposium on Artificial Neural Networks (ESANN)*, 2011.

[3] H. Deng and G. Runger, "Feature selection via regularized trees," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1–8.

[4] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using Naïve Bayes," *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.

[5] K. Zhong, P. Jain, and I. S. Dhillon, "Mixed linear regression with multiple components," in *Advances in neural information processing systems*, 2016, pp. 2190–2198.

[6] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2003, pp. 986–996.

[7] I. Rish, "An empirical study of the Naïve Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[8] Y. Y. Song and L. Ying, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[9] D. Jung and H. Park, "An iterative algorithm of key feature selection for multi-class classification," *ICUFN*, 2019.

[10] M. Pérez-Enciso and M. Tenenhaus, "Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach," *Human genetics*, vol. 112, no. 5-6, pp. 581–592, 2003.

[11] D. Ballabio and V. Consonni, "Classification tools in chemistry. part 1: linear models. pls-da," *Analytical Methods*, vol. 5, no. 16, pp. 3790–3798, 2013.

[12] D. V. Nguyen and D. M. Rocke, "Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics*, vol. 18, no. 9, pp. 1216–1226, 2002.

[13] V. Fortino, P. Kinaret, N. Fyhrquist, H. Alenius, and D. Greco, "A robust and accurate method for feature selection and prioritization from multi-class omics data," *PloS one*, vol. 9, no. 9, p. e107801, 2014.

[14] 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*, vol. 20, no. 15, pp. 2429–2437, 2004.