American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Mini review



# Multi-classification and Variable Selection Techniques in Cancer Genomic Data Research

**Nan Li*[1] and Nan Zhang[2]**

[1]*Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, US*

[2]*Department of Geology, Beijing University, China*

***Corresponding author:** Nan Li, Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, US.

***To Cite This Article:*** *Nan Li, Nan Zhang, Multi-classification and Variable Selection Techniques in Cancer Genomic Data Research. Am J Biomed Sci & Res. 2021 - 12(2). AJBSR.MS.ID.001725.* DOI: 10.34297/AJBSR.2021.12.001725.

**Received:** February 02, 2021; **Published:** March 03, 2021

**Keywords:** Cancer Classification, LASSO, Logistic Regression, Neural Network, SVM, Variable Selection

## Introduction

In the past two decades, a huge amount of high-throughput -omics data, such as genomics, transcriptomics, metabolomics, and proteomics, have been generated regarding variations in DNA, RNA, or protein features for many cancers. The tremendous volume and complexity of these data bring significant challenges for biostatisticians, biologists, and clinicians. One of the central goal of analyzing these data is disease classification, which is fundamental for us to explore knowledge, formulate diagnosis, and develop personalized treatment. Here, we review the statistical and machine learning techniques studied in cancer classification and the process or difficulties of categorizing cancer subtypes from their genomic features.

Traditionally cancer was classified by organ location, then it is further stratified by the cell type, patient age, or histological grade [1]. Finally, the dramatically wave of genomic data accelerate the trend of classifying cancer subtypes by the clinical outcome or treatment option. Exploration of multi-classification problems is essential for successful application in precision medicine.

## Method

### Multi-classification

A few notes of terminology are introduced at the beginning. Multi-classification is a kind of supervised learning, which aims to predict the value of a class outcome using input variables from a training set of samples with known class labels [2]. Another very popular machine learning technique, clustering, falls into the category of unsupervised learning, which doesn't need outcome label but has the goal to describe the associations and patterns among a set of input variables [2]. We will only talk about classification in this review. Since binary is a special situation for multi-classification, we will focus on multi-classification here.

In literature, there are two main types of classification: soft classification and hard classification. Soft classification rules first estimate the conditional outcome class probabilities and then predict the class label based on the maximum probability. Among them are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression [3]. On the other hand, hard classification rules directly target on the discriminant function without estimating conditional class probabilities, such as support vector machine (SVM, [4]).

Most binary classifiers, such as LDA, QDA and logistic regression, can be extended to multi-classification naturally. However, applying SVM in multi-category problems is not straightforward. One intuitive approach is to reduce a multi-category problem into a series of binary problems through strategies of "one vs. one" or

"one vs. rest". In a "one vs. one" reduction, one trains all pairwise binary classifiers for a K-class problem. For each test point, the predicted class is the one that wins the most pairwise contests. In the "one vs. rest" strategy, a K-class problem is divided into K "one-vs-rest" problems, and each "one-vs-rest" problem is addressed by a different class-specific binary classifier (e.g., "class 1" vs. "not class 1"). Then a new sample takes the class of classifier with the largest real valued output, such as confidence score. These indirectly approaches suffer from unbalanced or reduced sample size and fail to capture correlations among different classes [5].

To overcome these shortcomings, some directly simultaneous multi-classification methods (global models) were proposed to inherit and extend the optimal property of binary SVM to the multi-category case, such as multi-category SVM (MSVM, [6]) and multiclass Proximal SVM (MPSVM, [7]). Another stream of resemble methods, like boosting and random forest, are also very popular due to its high accuracy and strong generalization. Lastly, the famous neural network was developed separately in the field of artificial intelligent. Neural network essentially extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of those features [2]. It is especially effective for complex and hard to interpret input data, and among the most effective general purpose supervised learning methods currently known.

## Variable Selection

In cancer genomic studies, such as microarrays or RNA-seq, the overwhelming number of variables far exceeds the size of training samples even though the underlying model is naturally sparse. Therefore, it is essential to identify important variables for achieving classifiers with higher prediction accuracy and better model interpretability. Variable selection in multi-classification is much challenge than in binary classification or regression, since one needs to consider which variables are important for each individual discriminant functions separately as well as for the whole set of functions. In regression or binary classification, the modern penalized methods, such as LASSO [8], adaptive LASSO [9] or group LASSO [10], outperform the traditional methods of forward/backward/stepwise selection, due to their continuous selection process with smaller variation and better predicting power. Especially in high-dimension genomic analysis, the smaller sample size than the available genomic features make it impossible for practical applications of the traditional subset selection methods. In multi-classification, some existing works are L1-MSVM [11], group L1 multinomial logistic regression [12], supnorm MSVM [13], and supSCAD multinomial logistic regression/MSVM [14].

Alternatively, individual-gene-ranking of discriminant power [15] or filtering through relevance and correlation [16] can achieve good performance with comparatively less computation cost. Some genetic algorithm (such as recursive feature elimination scheme) can be embedded into different multi-classification machines to achieve variable selection in iterative steps. The nearest shrunken centroids classifier [17] was proposed in multiple cancer classification with gene expression and have shown good empirical performance. A hierarchical ensemble model with Error-Correcting Output Codes was studied in [18] on multi-class microarray data.

## Application

A variety of applications of previous reviewed methods have been studied in different kinds of -omics data in terms of cancer classification. SVM with recursive feature elimination was applied in multi-class cancer classification [19] and the performance was compared between microRNA and mRNA expression profiles [20]. Group L1 multinomial regression was applied in a three-class acute leukemia gene expression data [21]. Then an adaptive version of the Group L1 multinomial regression, which was designed to selecting informative gene groups and also important genes within each group, was developed and applied in lung cancer classification [22]. A deep learning model was proposed to classify multiple cancer subtypes using RNA-seq gene expression [23].

## Conclusion

Multi-classification and variable selection have been studied extensively in statistics and machine learning community. Various penalized classifiers have been proposed and examined to achieve good finite sample performance as well as sound asymptotically properties at manageable computing cost. On the other hand, ensemble methods, neural network and other deep learning technique have been applied to process different -omics data and develop complex models or classifiers.

## Acknowledgement

The author thanks the Editor and reviews for constructive comments, careful reading, and guidance of the paper presentation.

## References

1. Song Qingxuan, Sofia D Merajver, Jun Z Li (2015) "Cancer classification in the genomic era: five contemporary problems." Human genomics 9: 27.
2. Friedman, Jerome, Trevor Hastie, Robert Tibshirani (2001) The elements of statistical learning. New York: Springer series in statistics.1(10).
3. McCullagh, Peter, John A Nelder (1989) "Generalized linear models 2nd edition chapman and hall." London, UK.
4. Vapnik, Vladimir (2013) The nature of statistical learning theory. Springer science & business media.
5. Crammer, Koby, Yoram Singer (2001) "On the algorithmic implementation of multiclass kernel-based vector machines." J machine learning research 2: 265-292.
6. Lee, Yoonkyung, Yi Lin, Grace Wahba (2004) "Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data." J American Statistical Association 99(465): 67-81.
7. Tang, Yongqiang, Hao Helen Zhang (2006) "Multiclass proximal support vector machines." J Computational and Graphical Statistics 15(2): 339-355.

8. T ibshirani, Robert (1996) "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58(1): 267-288.

9. Zou, Hui (2006) "The adaptive lasso and its oracle properties." J the American statistical association 101(476): 1418-1429.

10. Yuan, Ming, Yi Lin (2006) "Model selection and estimation in regression with grouped variables." J the Royal Statistical Society: Series B (Statistical Methodology) 68(1): 49-67.

11. Wang, Lifeng, Xiaotong Shen (2007) "On l 1-norm multiclass support vector machines: methodology and theory." Journal of the American Statistical Association 102(478): 583-594.

12. Tutz, Gerhard, Wolfgang Pobnecker, Lorenz Uhlmann (2015) "Variable selection in general multinomial logit models." Computational Statistics & Data Analysis 82: 207-222.

13. Zhang, Hao Helen, Yufeng Liu, Yichao Wu, Ji Zhu (2008) "Variable selection for the multicategory SVM via adaptive sup-norm regularization." Electronic Journal of Statistics 2: 149-167.

14. Li, Nan, Hao Helen Zhang (2021) "Sparse learning with non-convex penalty in multi-classification." Accepted in Journal of Data Science 20(2).

15. Li, Tao, Chengliang Zhang, Mitsunori Ogihara (2004) "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression." Bioinformatics 20(15): 2429-2437.

16. Ooi, Chia Huey, Madhu Chetty, Shyh Wei Teng (2006) "Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data." BMC bioinformatics 7(1): 1-19.

17. Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, Gilbert Chu (2002) "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proc Natl Acad Sci 99(10): 6567-6572.

18. Liu Kun Hong, Zhi Hao Zeng, Vincent To Yee Ng (2016) "A hierarchical ensemble of ECOC for cancer classification based on multi-class microarray data." Information Sciences 349: 102-118.

19. Peng, Sihua, Xiaomin Zeng, Xiaobo Li, Xiaoning Peng, et al. (2009) "Multi-class cancer classification through gene expression profiles: microRNA versus mRNA." J Genetics and Genomics 36(7): 409-416.

20. Cao, Jin, Li Zhang, Bangjun Wang, Fanzhang Li, Jiwen Yang (2015) "A fast gene selection method for multi-cancer classification using multiple support vector data description." J biomedical informatics 53: 381-389.

21. Li Juntao, Yanyan Wang, Tao Jiang, Huimin Xiao, Xuekun Song (2018) "Grouped gene selection and multi-classification of acute leukemia via new regularized multinomial regression." Gene 667: 18-24.

22. Li Juntao, Yanyan Wang, Xuekun Song, Huimin Xiao (2018) "Adaptive multinomial regression with overlapping groups for multi-class classification of lung cancer." Computers in biology and medicine 100: 1-9.

23. Xu, Jing, Peng Wu, Yuehui Chen, Qingfang Meng, Hussain Dawood, et al. (2019) "A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data." IEEE Access 7: 22086-22095.