마이크로어레이 유전자 발현 기반 암 유형 분류를 위한 특징 축소 연구

벨무루간 아레스 발라지(*), 김경백(**)
(*) 전남대학교 인공지능융합학과, arreshvnass@gmail.com
(**) 전남대학교 인공지능융합학과, kyungbaekkim@jnu.ac.kr

A Study of Feature Reduction for Microarray Gene Expression based Cancer Type Classification

Velmurugan Arresh Balaji (*) , Kyungbaek Kim(**)
(*) Chonnam National University, Department of Artificial Intelligence Convergence
(**) Chonnam National University, Department of Artificial Intelligence Convergence

요 약

For efficient and accurate diagnosis and precise outcome in medical treatments in oncology, the Cancer classification area plays a major role. The emergence of DNA microarray technology has enabled researchers to analyze the expression level of thousands of genes simultaneously. We had used well kmown microarray gene expression dataset called 11_tumors dataset. In this study, we had classified different cancer types by using machine Learning and deep Learning algorithms. We tuned the microarray datasets with pre-processing techniques such as principal component analysis and scaling techniques. And, we had analysed the impact on feature reduction over high dimensional microarray cancer gene expression dataset. Also, we have achieved significant improvement in certain algorithms such as support vector classification and multilayer perceptron after tuning the input datasets.

1. Introduction

Cancer-type classification plays a major role in improving patient survival rates. A microarray consists of a solid surface to which biological molecules are arranged in a regular pattern. For many treatments, there are certain genetic molecular factors like signatures or gene alternations with heterogeneous biological characteristics that allow discerning the responses. The carcinogenic cells have some gene expression characteristics like impaired gene expressions. The degree of dataset separability is highly correlated to the performance of classification; Hence, we analysed the performance using techniques such as clustering. For the distribution of microarray gene expression dataset, based on the labels of data, we can obtain good apriori insight over the algorithm.. Even from small quantity of data, the characteristics of the problem given can be learned by techniques such as ML and DL algorithms. And these data were classified into two parts such as training and validation.

For evaluating the performance of the model, validation dataset is used and for the parameter calibration the training dataset is used. To obtain clear knowledge about the dataset,we have done an hierarchical analysis, before applying these supervised learning (classification algorithms). There are distinct distance metrics used by the hierarchical clustering, such as average, complete, ward and median, weighted, centroid and single. At first, we used a dataset without preprocessing it as the input dataset. These distance metrics serve to he differences between the data samples and vary in their capacity to deal with large outliers (i.e., between weighted, centroid, and median metrics) or if they allow choosing the number of clusters to consider.

After this clustering, we tested all of the datasets created in the previous step To find the methodology for best preprocessing . In this study, we compare the performance of the most commonly used ML and DL algorithms in bioinformatics in the task of classifying by supervised and unsupervised techniques. We used the 11_Tumor database and applied different preprocessing strategies.. The datasets used represent measurements of gene expression using cancer microarrays and normal biopsies. This database consists of 174 samples with 12,533 gene expression microarrays for 11 different types of cancer. The 12,533 microarrays of genetic expression are integers withpositive and negative values; these values represent the characteristics that allow the ML and DL algorithms to learn how to classify by cancer type.Considering an Cost efficient and improved cancer diagnosis, these information of gene expression data extracted from the tumour cells in the microarray improves the cancer treatment. For analyzing the huge amount of genes in the microarray dataset, the Machine Learning algorithms are more precise and effective.

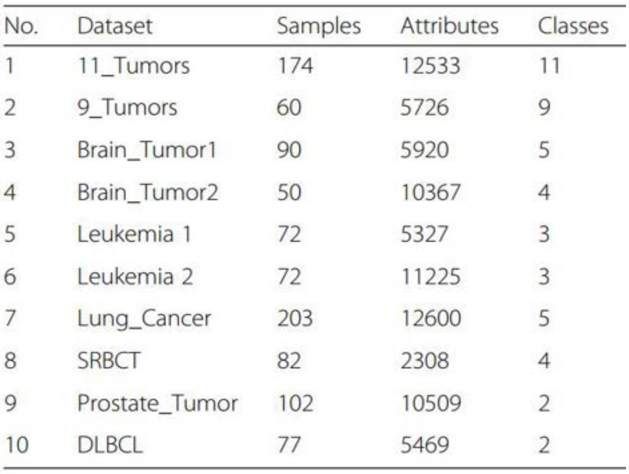
2. Background

2.1 An overview of Microarray Datasets

Based on gene expression, for classifying the cancer types and diagnosing, identifying the gene expression informative subset by means of feature selections plays a crucial role. To infere these targeted gene expressions, Previous articles used both ML and DL algorithms in microarray gene expressions. The 11_Tumors database is an popular Microaarray gene expression dataset related to cancer diagnosis, and for the curse of dimensionality it is one of the best example because, high number of characteristics and few registers of this database. Table 1 presents the details of the experimental datasets in terms of diverse samples, attributes and classes.

2.2 Related Works

The performance of the IG/SGA algorithm [1] is evaluated by considering seven microarray datasets and the results are compared with six techniques. High classification accuracy of 100% is achieved for two datasets (Lung cancer-Michigan and Prostate Cancer datasets). The performance of the IG/SGA algorithm [1] is evaluated by considering seven microarray datasets and the results are compared with six techniques. High classification accuracy of 100% is achieved for two datasets (Lung cancer-Michigan and Prostate Cancer datasets). Innovative gene selection algorithm (GSP) [2] can not only provide a smaller subset of relevant genes for cancer classification but also achieve higher classification accuracies in most cases with shorter processing time compared with GEP.



No.	Dataset	Samples	Attributes	Classes
1	11_Tumors	174	12533	11
2	9_Tumors	60	5726	9
3	Brain_Tumor1	90	5920	5
4	Brain_Tumor2	50	10367	4
5	Leukemia 1	72	5327	3
6	Leukemia 2	72	11225	3
7	Lung_Cancer	203	12600	5
8	SRBCT	82	2308	4
9	Prostate_Tumor	102	10509	2
10	DLBCL	77	5469	2

<Table 1> List of Microarray Datasets with attributes

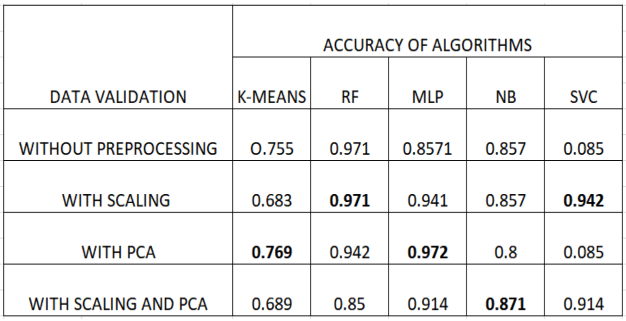
However, the processing time of GSP is still longer than that of PSO and GA models Therefore, Most of the related articles, used the methods for selecting features, an data science technique for testing particular data, like clustering methods, preprocessing techniques, dimension reduction and feature selection. By using some Machine Learning algorithms with one tuning (preprocessing) strategy, and a learning technique unsupervised / supervised learning, [3] this article achieved high accuracy which could add bias to their methodology. Moreover, these Microarray datasets, has the common problem called the curse of dimensionality. Therefore, the data are dispersed and the results are not statistically stable or reliable, directly affecting the accuracy achieved by ML and DL algorithms.

3. Feature Reduction with PCA and Scaling

Hence, We are proposing two Tuning (preprocessing) techniques scaling and principal component analysis (PCA) inspired from [3] to solve this problem. The first technique is used to calibrate the model and to perfectly place the data are in the suitable values. To improve the statistical and to decrease the noise introduced during training the model by irrelevant characteristics, the second technique is used. For each machine learning and deep learning algorithms, there are four datasets to be created for training and validation. No preprocessing techniques to be applied for the first dataset. And only a scaling process for the second dataset; for the third, we applied Principal component analysis with 96% variance retained. Atlast, for the last (fourth) dataset, we applied both PCA and scaling techniques, and obtained a dimensional Features reductionof 90.5% (principal components).

4. Experimental Evaluation

After tuning the raw input datasets into four preprocessed datasets for each dataset, We propose to evaluate the performance of well-known ML classification algorithms, including SVC, LDA, MLP, RF and K-means. The SVC and MLP outperformed the other state-of-art models. These two algorithimic results has improved a lot after tuning with PCA and Scaling techniques.



DATA VALIDATION	ACCURACY OF ALGORITHMS				
	K-MEANS	RF	MLP	NB	SVC
WITHOUT PREPROCESSING	0.755	0.971	0.8571	0.857	0.085
WITH SCALING	0.683	**0.971**	0.941	0.857	**0.942**
WITH PCA	**0.769**	0.942	**0.972**	0.8	0.085
WITH SCALING AND PCA	0.689	0.85	0.914	**0.871**	0.914

(Figure 1) Evaluation comparision with different algorithms for 11_Tumors dataset

5. Conclusion

For Classifying Tumours types with complex microarray cancer datasets, techniques like machine learning and deep learning are highly effective. In this paper, we propose tuning techniques for improving accuracies in classifying the cancer types, which inturn accelarate the precise cancer type predictions for patients with these certain pathologies. And it will also provide new precise treatment or medicine for the cancer patients with these tumor types.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(NRF-2017R1A2B4012559).

참고 문헌

[1] Salem, Hanaa, Gamal Attiya, and Nawal El-Fishawy. "Classification of human cancer diseases by gene expression profiles." Applied Soft Computing 50 (2017): 124-134.

[2] Alanni, Russul, et al. "A novel gene selection algorithm for cancer classification using microarray datasets." BMC medical genomics 12.1 (2019): 10.

[3] Tabares-Soto, Reinel, et al. "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data." PeerJ Computer Science 6 (2020): e270.

[4] Dwivedi, Ashok Kumar. "Artificial neural network model for effective cancer classification using microarray gene expression data." Neural Computing and Applications 29.12 (2018): 1545-1554.

[5] Velmurugan Arresh Balaji, Chulwoong Choi, and Kyungbaek Kim. 2020.Survey on High-Dimensional Medical Data Clustering. In The 9th International Conference on Smart Media and Applications (SMA 2020), September17 – 19, 2020, Jeju, Republic of Korea. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3426020.34260 71.

[6] Rani, M. Jansi, and D. Devaraj. "A combined clustering and ranking based gene selection algorithm for microarray data classification."

2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2017.

[7] Guillen, Pablo, and Jerry Ebalunode. "Cancer classification based on microarray gene expression data using deep learning." 2016 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2016.

[8] Zeebaree, Diyar Qader, Habibollah Haron, and Adnan Mohsin Abdulazeez. "Gene selection and classification of microarray data using convolutional neural network." 2018 International Conference on Advanced Science and Engineering (ICOASE). IEEE, 2018.

[9] Khorshed, Tarek, Mohamed N. Moustafa, and Ahmed Rafea. "Multi-Tissue Cancer Classification of Gene Expressions using Deep Learning." 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 2020.

[10] Saqib, Pakizah, et al. "MF-GARF: Hybridizing Multiple Filters and GA Wrapper for Feature Selection of Microarray Cancer Datasets." 2020 22nd International Conference on Advanced Communication Technology (ICACT). IEEE, 2020.

[11] Gálvez, Juan M., et al. "Towards improving skin cancer diagnosis by integrating microarray and RNA-seq datasets." IEEE journal of biomedical and health informatics 24.7 (2019): 2119-2130.