Machine Learning In Bioinformatics: Gene Expression And Microarray Studies

Beyza BAĞIRÖZ, Emre DORUK
Computer Engineering
Gazi University
Ankara,Turkey
beyzabagirozz@gmail.com, emredoruk@gazi.edu.tr

Oktay YILDIZ
Computer Engineering
Gazi University
Ankara,Turkey
oyildiz@gazi.edu.tr

***Abstract*—Machine learning methods used in the field of bioinformatics are a frequently used solution method in diagnosing, treating and investigating the underlying causes of diseases. In addition, it is an important field of study that allows for the ease of processing, the provision of computational power and the diversity of computational tools specific to the subject, especially in processes that require processing on gene expression and microarray data sets. In this study, an introduction has been made on the use of machine learning methods in the field of bioinformatics gene expression, and the use of machine learning methods has been exemplified by recent studies.**

***Keywords*— machine learning; bioinformatics; gene expression; microarray.**

I. Introduction

Machine learning (ML), which is very popular today and has a wide range of uses, is concerned with the design of programs that can learn the rules from the available data, adapt to changes, and improve the experience and experience gained [1]. The fact that computers find solutions to more complex problems day by day, are more integrated into human life, and the resulting data density is high, making machine learning important.

In addition to these; ML, which has a range of applications from medicine to military, analyzes the past data to develop models to be used in banks, credit applications, telecommunications, stock market and fraud detection. it also manifests itself in many disciplines (physics, astronomy, biology, etc.). ML, which stands out for its in-depth knowledge and understanding of medical data, especially in health and medicine, has an important place in early health applications, prevention of diseases and cost savings.

ML is expanding to various applications by moving to the advanced dimension of computer science and engineering and becoming a basis for interdisciplinary research. ML technologies are also useful for categorizing, processing and integrating data into various systems to reveal meaningful information in bioinformatics applications [6]. Bioinformatics conceptualizes biology in terms of molecules, applies "data-information techniques" derived from fields such as applied mathematics, computer science, and statistics, and enables the processing and discovery of biological data by organizing information about these molecules [7].

Within the scope of bioinformatics research, the size of the data is increasing rapidly. Nowadays, the volume of data is increasing everywhere due to the digitalization of all processes and easier access of high efficiency devices. Biologists or healthcare professionals now prefer anilis and research on larger and ever-growing genomic data rather than using traditional laboratories to study a disease [9]. The trend in increasing data volumes is supported by the increasingly large data technologies and the decrease in information processing costs and the increase in analytical efficiency. With its applications on growing genomic data, ML technology has an effective use in bioinformatic.

Basically, there are five types of data that are frequently used in large volume and bioinformatics research. These data types;

- Gene expression data
- Gene ontology (GO)
- Protein-protein interaction (PPI) data
- Pathway data
- DNA, RNA and protein sequence data

Gene expression analysis among them has come to the forefront recently and has an important position in the diagnosis of diseases. In gene expression analysis, conditions such as expression levels of many genes, stages of treatment and disease progression are analyzed. Today, gene expression data analysis helps to diagnose and examine a disease with a gene by detecting genes affected by viruses or payogenes, gene expression values of infected or uninfected cells [10].

II. Machine Learning in Bioinformatic Era

A. Machine Learning Study Topics in Bioformatics

The basic biological areas where the data are analyzed by applying ML techniques are categorized in six different areas. These are genomics, proteomics, microarrays, systems biology and evolutionary analysis [12].

First of all, Genomics, one of the most important fields of Bioinformatics, is a science that examines all structural and functional aspects of genomes belonging to different species, genomes of organisms, nucleotide sequences by applying

978-1-7281-8073-1/20/$31.00 ©2020 IEEE

chromosome sequencing techniques. It must be processed and analyzed to extract valuable information from genomic data.

The information obtained from the sequencing process is used in fields such as genomics, metagenomics, medical diagnosis, early diagnosis of cancer, regulation of gene expression as well as basic biological research [13]. Firstly, the location and structure of genes can be extracted from genome sequences for data processing. Next, the identification of regulatory elements and non-coding RNA genes is discussed.

In the proteomics field, the main purpose of computational methods is to predict protein structure. Proteins are defined as complex macromolecules containing a large number of atoms.Therefore, the number of possible structures that can occur is high. ML techniques are used for predicting protein function, as in the genomic domain, in the proteomic domain [13].

Another area, microarray, is the best known area where complex experimental data are collected. Two different ways are followed for complex data. The first step is that the data must be preprocessed, that is, the data is changed by applying machine learning algorithms. The second step is to analyze the data to find the desired results. Examples of use of microarray data include pattern identification and data classification classification on data, as well as genetic network induction [14].

System biology; is another area where machine learning technology and biology work together. Since modeling intracellular life processes is a complex process, calculation techniques in machine learning are used for modeling biological and genetic networks, signal transmission networks, and metabolic events [3].

Evolution, which is one of the main topics in biology, and ml techniques are used for phylogenetic tree reconstruction. The phylogenetic tree reconstruction is modeling the evolution of an organism and ml algorithms are applied for this [12].

As the last area, text mining is mentioned due to the application of computational techniques depending on the increasing amount of data. Examining this situation, text mining is becoming increasingly important in computational biology, and functional explanation is applied to topics such as cellular location estimation and protein interaction analysis [15].

B. Importance of Gene Expression to Bioinformatic

Gene expression and gene datasets are important areas of interest frequently used in bioinformatics. It is used to categorize various disease states, existing genetic disorders or personal cellular conditions as a result of extracting gene expression profiles from the data in the field of bioinformatics and analyzing these data. In addition, it has expensive processing steps today.

Detecting gene expression levels and deducing from biological data enables early diagnosis of various diseases. Cancer is one of the most common and most subject to many studies. Cancer causes maximum death worldwide, according to data provided by the World Health Organization [16].

Gene expression data analysis is also important in the diagnosis of cancer diseases. Cancer disease, which has various types, poses a great risk in human life as it affects human identity, behavior and vital functions. Solution to minimize potential risks; it is the early detection of cancer or to reveal the possibility of developing this disease in the future. Gene expression of cancerous cells can be compared to healthy cells to understand cancer pathology. Some cancer conditions in humans are associated with genetics. For this reason, it is necessary to correctly analyze gene expression data and make high accuracy predictions for the factor causing cancer or possible cancer risk. In other words, by using gene expression, candidate can reveal genes to determine a person's sensitivity to cancer.

Some difficulties arise in the detection or structural examinations of various disease types such as cancer diseases. The complexity and high dimensions of genomic data limit the analysis to be made. In the field of microarray research, the increased sample size and feature size of gene expression data requires the use of classification techniques (eg: SVM, Naive Bayes) along with various computational methods (machine learning) for efficient analysis. However, due to the large number of genes and the small number of patients, machine learning techniques and classifiers prevent them from making accurate predictions. Because most of the genes are unnecessary, which reduces prediction performance.

Gene selection is needed to select the genes most relevant to the disease in order to increase performance in the examination and early diagnosis of diseases. If we look at cancer disease, applying gene selection methods for the analysis of gene expression levels and cancer classification is a necessary step. Various gene selection methods are currently being developed for the selection of better gene subsets [17] [18].

III. Machine Learning Techniques for Gene Expression Data Analysis and Releated Works

In this section, in the bioinformatics included in the study, machine learning techniques used for gene expression data analysis and related studies that are created by using these techniques are included. Common used machine learning techniques are classified in Figure 1. Considering the studies on gene expression, it is striking to use machine learning techniques as a solution, especially in cases where computational processes are intense and time consuming. Although these studies are frequently encountered in diagnosis and diagnosis, there are many studies for different purposes with gene and microarray data. For example, in a study by David G. et al, it was stated that it is often difficult to differentiate many different soft tissue sarcoma subtypes based on morphology. Stochastic Neighbor Embedding clustering and a deep neural network methods were analyzed and 3 molecular overlapping tumor groups were determined Again, in the same study, prognostic genes were identified using the K-nearest neighbor algorithm and confirmed by comparing with expression data from the French Sarcoma group [32]. In another study, gene expression data on the basis of cell lines were analyzed via SVM in order to predict the individual activities of cancer drugs. The method developed has been validated for cancer-like diseases (renal carcinoma, lung adenocarcinoma, and chronic myeloid leukemia) treated with target drugs [33].

Machine learning algorithms can be used in gene expression, data generation, data analysis and prediction processes. For example, in a study by Brian K. et al., By integrating 3 different gene expression data to predict Systemic lupus erythematosus (SLE) disease activity, machine learning methods were applied, and the classification process of SLE patients was performed. In this study in which KNN and random forest learning algorithms were used, random forest algorithm was the best method with 83 percent accuracy [34].

In another study by Reinel T. et al., They conducted a comprehensive analysis of ML and Deep Learning (DL) approaches based on 11 different tumor classes. In this study, which includes the application of DL methods such as KNN, K-means, Random Forest, Naive Bayes, Decision Trees together with Evolutionary Neural Networks (CNN) and Multi-layer sensor (MLP), the highest accuracy rate among algorithms was logistic regression with 90.6 percent and 94.43 percent. provided with CNN applications [35].

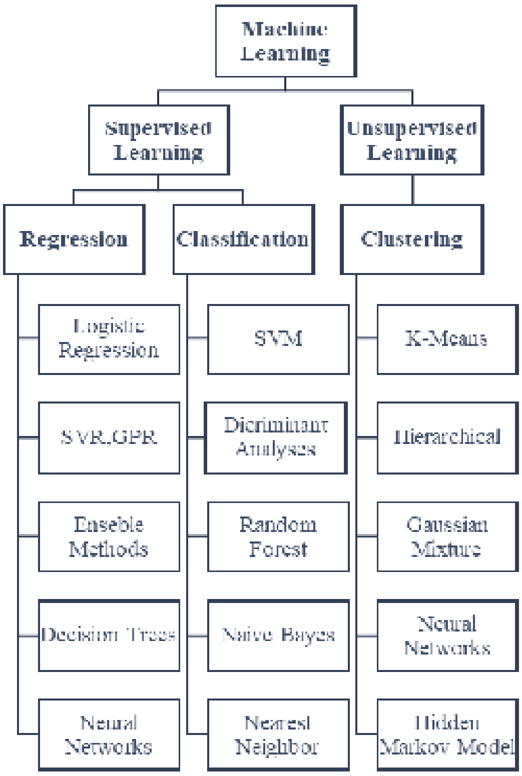

Fig. 1. Machine Learning Techniques

Machine learning techniques are also used in the selection of attributes in order to facilitate the analysis of gene expression or microarray data, as well as the detection of diseases. For example; A recently proposed hybrid trait selection method carried out the state-of-the-art Correlation-based trait selection at the first stage and then implemented the advanced binary particle swarm optimization (iBPSO) application they developed. [36]. The method, which was performed on 11 different microarray datasets and evaluated in a comparative way, obtained remarkable results in terms of both high success of classification accuracy and size reduction performance. In addition, among evolutionary computational techniques, bacterial-inspired algorithms are frequently encountered recently to provide a solution to the feature selection problem. For example in a study, the feature selection method guided by the bacterial colony optimization algorithm aimed to increase the search capacity in discrete optimization problems while reducing computational complexity [37]. In another study, two hybrid feature selection algorithms combining Wrap and a filter based on binary differential evolution were developed and compared with different methods and competitive results were obtained [38]. In yet another study, a hybrid feature selection algorithm, combined with an adaptive genetic algorithm (AGA) with mutual information maximization, has been proposed [39].

Studies that use machine learning methods to create gene-oriented diagnosis and relationship models are also frequently encountered. In specific cases, identifying the underlying causes and gene relationships of diseases is an essential preliminary step in developing the necessary treatments. Hossain et al. In a study by Ovarian Cancer (OC) to determine the disease process, mortality rate and related gene expression patterns; OC mRNA expression levels and OC tissue expression on 41 genes currently included in research, based on Broad Institute Cancer Genome Atlas (TCGA) datasets that include cancer site, stage and subtype, patient age and patient salvage rate, as well as OC gene transcription profiles levels were determined and their effect on survival rate was examined. In the univariate analysis, there is a significant relationship between CDH1, TLR4, BSCL2 genes and survival, although there is a significant relationship in the transcription levels of genes encoded BSCL2, TLR4, ERBB2, CDH1 and SCGB2A1 [40].

IV. Conslusions

Increasing data with the use of technology in bioinformatics has contributed to the development of computer computing tools and the diversity of their usage areas. Information discovery, especially in gene expression and microarrays that create big data, is critical in identifying diseases, detecting diseases and at certain stages of treatment processes. Machine learning methods, which facilitate information discovery processes, are widely used in the field of bioinformatics, increasing its value in this field day by day. In addition, subject-oriented machine learning algorithms enable bioinformatics researchers to focus on their areas of expertise by facilitating their analysis processes.

The study emphasized the importance of machine learning methods in bioinformatics, and presented examples of the use of machine learning methods, and a review was made on the analysis of gene expression and microarray data with ML.

References

[1] Blum, A. (2007). Machine learning theory. Carnegie Melon University, School of Computer Science, 26.

[2] Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (2012). Data mining methods for knowledge discovery (Vol. 458). Springer Science & Business Media.

[3] Zhaoli, "Machine learning in bioinformatics," Proceedings of 2011 International Conference on Computer Science and Network Technology, Harbin, 2011, pp. 582-584, doi: 10.1109/ICCSNT.2011.6182026.

[4] El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. In machine learning in radiation oncology (pp. 3-11). Springer, Cham.

[5] Ranadive, F., Surti, A., & Sharma, P. (2019, March). Comparative Analysis of Machine Learning Classifiers on Bioinformatics and Clinical Datasets. In 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 608-611). IEEE

[6] J. Dale, J. Matta, S. Howard, G. Ercal, W. Qiu and T. Obafemi-Ajayi, "Analysis of grapevine gene expression data using node-based resilience clustering," 2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), St. Louis, MO, 2018, pp. 1-8, doi: 10.1109/CIBCB.2018.8404962.

[7] Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. Methods of information in medicine, 40(04), 346-358.

[8] Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006). Machine learning in bioinformatics. Briefings in bioinformatics, 7(1), 86-112.

[9] Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S., & Bhattacharyya, D. K. (2015). Big data analytics in bioinformatics: A machine learning perspective. arXiv preprint arXiv:1506.05101.

[10] Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S., & Bhattacharyya, D. K. (2016). Big data analytics in bioinformatics: architectures, techniques, tools and issues. Network Modeling Analysis in Health Informatics and Bioinformatics, 5(1), 28.

[11] Bhaskar, H., Hoyle, D. C., & Singh, S. (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. Computers in biology and medicine, 36(10), 1104-1125.

[12] Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006). Machine learning in bioinformatics. Briefings in bioinformatics, 7(1), 86-112.

[13] F. Celesti et al., "Big data analytics in genomics: The point on Deep Learning solutions," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 306-309, doi: 10.1109/ISCC.2017.8024547.

[14] Bharathi, A., & Natarajan, A. M. (2010, December). Microarray gene expression cancer diagnosis using Machine Learning algorithms. In 2010 International Conference on Signal and Image Processing (pp. 275-280). IEEE.

[15] Krallinger, M., Erhardt, R. A. A., & Valencia, A. (2005). Text-mining approaches in molecular biology and biomedicine. Drug discovery today, 10(6), 439-445.

[16] Pati, J. (2018). Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach. IEEE Access, 7, 4232-4238.

[17] Duan, K. B., Rajapakse, J. C., Wang, H., & Azuaje, F. (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE transactions on nanobioscience, 4(3), 228-234.

[18] Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., & Mewes, H. W. (2005). Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry, 29(1), 37-46.

[19] S. Sharma, J. Agrawal, S. Agarwal and S. Sharma, "Machine learning techniques for data mining: A survey," 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, 2013, pp. 1-6, doi: 10.1109/ICCIC.2013.6724149.

[20] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 1310-1315.18

[21] A. C. Lorena et al., "Comparing machine learning classifiers in potential distribution modelling", Expert Systems with Applications, vol. 38, pp. 5268-5275, 2011.

[22] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. International Journal of Computer Trends and Technology (IJCTT), 48(3), 128-138.

[23] Maglogiannis, I. G. (Ed.). (2007). Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies (Vol. 160). Ios Press.

[24] [24] M. Robnik-Sikonja, "Improving Random Forests", Machine Learning ECML, 2004

[25] Z. H. Zhou, "Rule extraction: Using neural networks or for neural networks?", Journal of Computer Science and Technology, vol. 19, no. 2, pp. 249-253, 2004.

[26] Segal, E., Taskar, B., Gasch, A., Friedman, N., & Koller, D. (2001). An Overview of Statistical Learing Theory. Bioinformatics, 17, 243.

[27] Jianjun Ye, Hongxun Yao and Feng Jiang, "Based on HMM and SVM multilayer architecture classifier for Chinese sign language recognition with large vocabulary," Third International Conference on Image and Graphics (ICIG'04), Hong Kong, China, 2004, pp 377-380, doi: 10.1109 / ICIG.2004.44.

[28] Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons.

[29] Hsu, C. C., Huang, Y. P., & Chang, K. W. (2008). Extended Naive Bayes classifier for mixed data. Expert Systems with Applications, 35(3), 1080-1083.

[30] Wang, A., An, N., Chen, G., Li, L., & Alterovitz, G. (2015). Accelerating wrapper-based feature selection with K-nearest-neighbor. Knowledge-Based Systems, 83, 81-91.

[31] Hajmeer, M., & Basheer, I. (2003). Comparison of logistic regression and neural network-based classifiers for bacterial growth. Food Microbiology, 20(1), 43-55.)

[32] David G. P. van IJzendoorn, Szuhai, K., Briaire-de Bruijn, I. H., Kostine, M., Kuijjer, M. L., & Bovée, J. V. (2019). Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLoS computational biology, 15(2), e1006826.

[33] Borisov, N., Tkachev, V., Suntsova, M., Kovalchuk, O., Zhavoronkov, A., Muchnik, I., & Buzdin, A. (2018). A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency. Cell Cycle, 17(4), 486-491.

[34] Kegerreis, B., Catalina, M. D., Bachali, P., Geraci, N. S., Labonte, A. C., Zeng, C., ... & Grammer, A. C. (2019). Machine learning approaches to predict lupus disease activity from gene expression data. Scientific reports, 9(1), 1-12.

[35] Tabares-Soto, R., Orozco-Arias, S., Romero-Cano, V., Bucheli, V. S., Rodríguez-Sotelo, J. L., & Jiménez-Varón, C. F. (2020). A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. PeerJ Computer Science, 6, e270.

[36] Jain, I., Jain, V. K., & Jain, R. (2018). Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing, 62, 203-215.

[37] Wang, H., Jing, X., & Niu, B. (2017). A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. Knowledge-Based Systems, 126, 8-19.

[38] Apolloni, J., Leguizamón, G., & Alba, E. (2016). Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Applied Soft Computing, 38, 922-932.

[39] Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., & Gao, Z. (2017). A hybrid feature selection algorithm for gene expression data classification. Neurocomputing, 256, 56-62.

[40] Hossain, M. A., Islam, S. M. S., Quinn, J. M., Huq, F., & Moni, M. A. (2019). Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. Journal of biomedical informatics, 100, 103313.