COMPARING THE PERFORMANCE OF VARIOUS MACHINE LEARNING TECHNIQUES ON CANCER PHENOTYPE ANALYSIS

ANNOTATED BIBLIOGRAPHY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE DEGREE OF

BACHELOR OF THE SCIENCE OF ENGINEERING

Submitted by:

Nirmani B.G.M. (2017/E/074)

DEPARTMENT OF COMPUTER ENGINEERING

FACULTY OF ENGINEERING

UNIVERSITY OF JAFFNA

[NOVEMBER] [2020]

ARTICLE 1

NIRMANI B.G.M.
2017/E/074
Week 1

Guo, "Transcription: the epicenter of gene expression", *Journal of Zhejiang University SCIENCE B*, vol. 15, no. 5, pp. 409-411, 2014. Available: 10.1631/jzus.b1400113.

(2) The paper analyzed the facts regarding the topics like regulation of transcription, the latest advances in epigenetics, mRNA processing, RNA quality control and HIV. (3) This paper presented some findings according to the studies of different field expert ideas. (4) The behavior of genomic level structures is discussed here with respect to basic molecular biology and biochemistry while providing mechanistic insights. (5) So, this article provided a good understanding of the basic concepts of translations of DNA and gene expression. (6) But it just analyzed the existing studies rather than doing some new experiment, just some theoretical explanation. (7) Finally, the research is combined the latest studies on regulations of transcription which may easy for people to understand the basics of biology. (8) This study is very important, as it is included the basic biological knowledge on gene level structures, which we need in our research. Rather than that, the research is not combined with any machine learning aspects or engineering approaches.

ARTICLE 2

NIRMANI B.G.M.
2017/E/074
Week 1

(1) Y. Hou and J. Linhong, "Gene Transcription and Translation in Design", *Volume 7: 27th International Conference on Design Theory and Methodology*, 2015. Available: 10.1115/detc2015-46128

(2)

(2) The researchers built a biologically inspired design framework by investigating the process from DNA to the protein in embryogenesis, while comparing the design and development processes. (3) Here, the study focused on induction and cellular differentiation, commitment of cell or tissue, and gene transcription and translation. (4)They are describing the process from the information to structure by simulating the translating processes in biology and providing some engineering examples. (5) The article has provided a good approach for explaining the behavior of DNA to the protein translation which is very important in molecular level analyzing. (6) The discussion considers only gene transcription and translation, not other transformations between different stages. (7) A qualitative development framework for gene transcription and translation in design process was built from the research while presenting examples and the concept can be developed to handle the complex situations as well. (8) The research is presenting a genomic level behavior analyzing to take advantage of inspiration of nature in engineering models without paying consideration on contribution of machine learning technologies in molecular level analyzes.

ARTICLE 3

NIRMANI B.G.M.
2017/E/074
Week 2

(1) C. Xu and S. Jackson, "Machine learning and complex biological data", *Genome Biology*, vol. 20, no. 1, 2019. Available: 10.1186/s13059-019-1689-0.

(2) In this article researchers reviewed, various applications of machine learning techniques on large, complex biological data and, handling these complex data to create a better system. (3) Here the authors use data from Center for Applied Genetic Technologies, Institute for Plant Breeding, Genetics and Genomics, The University of Georgia, 111 Riverbend Rd, Athens, GA 30602, USA to study the applications of machine learning on biological problem such as prediction of regulatory regions, solving population and evolutionary genetics questions. (4) Their research focused on the biological meaning of the predictive model than the predictive accuracy of the model from different biological aspects such as genomes, epigenomes, transcriptomes, and metabolomes. (5) The article is useful to our research, as this article discusses how the complex and huge amount of data can be handled using different machine learning techniques. Also the article mentioned that the complex biological phenomena cannot be analyzed using single data type and integrated analysis is necessary (6) Chunming Xu and Scott A. Jackson also stated that hidden biological effects can affect the accuracy of the system and they did not consider this information in their model which is a drawback of their model. Heterogeneity of different data types can occur and biological data are highly dimensional. (7) As novel techniques will provide biological insights from the large and often heterogeneous data, applying machine learning to complex biological data will be increasingly in demand in the future. (8) Though the article does not directly explain the cancer phenotype data, it shows how the behavior of biological data can be analyzed with machine learning.

ARTICLE 4

NIRMANI B.G.M.
2017/E/074
Week 3

(1) A. LG and E. AT, "Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence", *Journal of Health & Medical Informatics*, vol. 04, no. 02, 2013. Available: 10.4172/2157-7420.1000124.

(2) Advanced data mining techniques can be used to discover hidden patterns and relationships. Here, researchers compared Decision Tree, Support Vector Machine (SVM), and Artificial Neural Network (ANN) through sensitivity, specificity, and accuracy on breast cancer dataset. (3) The researchers' objective was predicting the recurrence of breast cancer to make right decisions on treatments using data mining technologies. They used patients registered in the Iranian Center for Breast Cancer (ICBC) program from 1997 to 2008 for their study. And they had 1189 records with 22 predictor (population characteristics) variables. (4) This study focused on clearly defined scope as it analyzed breast cancer data while comparing three different machine learning technologies. (5) Each machine learning techniques have its own limitations and strengths. So the researchers have tried to find the best technology in this case. (6) However their data consists of missing value which may decrease the performance of the model. (7) According to their results, SVM is the best classifier predictor followed by ANN and DT. Also they stated that the future studies may improve the performance using more genomic data and a longer follow-up duration. (8) This research had used similar comparing methods on the same dataset as our research. But our project is using more than three machine learning methods to predict more than one breast cancer predictions.

ARTICLE 5

NIRMANI B.G.M.
2017/E/074
Week 3

(1) G. Dubourg-Felonneau et al., "A Framework for Implementing Machine Learning on Omics DataML4H/2018/102", *Machine Learning for Health (ML4H) Workshop at NeurIPS*, no. 42018102, p. 5, 2018.

(2) This article provided a framework for combining omics data sets and handling high dimensional data to predict cancer patient's survival for individuals with high accuracy and low variance. (3) Here the authors analyzed and integrated for a set of 3,533 breast cancers with 15233 observed genes. Both unsupervised and semi supervised machine learning methods have been came up with a better method of combining highly heterogeneous data. (4) The research combined three different datasets related to breast cancer. (5) This research was related with our research as it analyzed data using some different machine learning technologies to reveal a better methodology while combining different datasets. (6) Here they tried to predict breast cancer patient survival for individuals. So they tracked clinical data of patients. But the problem is, not every patient is tracked until death. Only 55.4% of patients were tracked until death. (7) However, this study could manage the problem of high dimensional –omics data by pipelining method. Also they provided a prediction model for short term survival of breast cancer patients. (8) Even though this study differ from ours in term of combining different dataset rather than using single dataset for the prediction, they also used cancer genomic data.

ARTICLE 6

NIRMANI B.G.M.
2017/E/074
Week 4

(1) M. Ding, L. Chen, G. Cooper, J. Young and X. Lu, "Precision Oncology beyond Targeted Therapy: Combining Omics Data with Machine Learning Matches the Majority of Cancer Cells to Effective Therapeutics", *Molecular Cancer Research*, vol. 16, no. 2, pp. 269-278, 2017. Available: 10.1158/1541-7786.mcr-17-0378.

(2) This researchers provided a better approach to study the hidden information of omic data and associate them for a drug sensitivity prediction. (3) Their aim was avoiding or minimizing the use of ineffective drugs which are having more side effects on patients, and suggesting the most suitable drugs for a specific situation. So that they used machine learning methods and classification models on cancer data, which are gained from another large pharmacogenomics studies (Genomics of Drug Sensitivity in Cancer Project, and the Cancer Cell Line Encyclopedia). (4) Here, their task was finding out the features regarding sensitivity and specificity. They did not consider predicting any phenotype data of those patients. (5) However, their research has some relevance to our study as they also analyzed the hidden features of omic data using various machine learning technologies to reach their goal. (6) Still they mentioned that the techniques they used to reduce dimensionality may lead to lose of important predictive variables, which is a drawback of their study. Also they haven't considered important variables such as tumor size, stage and other factors that are using in clinical level. (7) As a whole, they built a system for target therapy, and it's performance was tested on a different data set. (8) This article is related to our study as both of them are using the genomic level feature handling for the predictions.

ARTICLE 7

NIRMANI B.G.M.
2017/E/074
Week 5

(1) C. Boeri et al., "Machine Learning techniques in breast cancer prognosis prediction: A primary evaluation", *Cancer Medicine*, vol. 9, no. 9, pp. 3234-3243, 2020. Available: 10.1002/cam4.2811.

(2) In this article, they studied the cancer recurrence (both loco-regional and systemic) and death within 32 months from the disease. (3) Their aim was evaluating the application of machine learning methods on breast cancer prognosis. They used Artificial Neural Network and Support Vector Machines, for each case and altogether they had 6 models. (4) Their data was collected from their institute and they considered breast cancer recurrence predictions. (5) They used two machine learning models per each study (Loco-regional recurrence, Systemic recurrence and Death from disease) on breast cancer predictions and reported high accuracy and specificity. (6) The used data with low number of samples, which has followed up for a short term. They paid less attention on molecular data by using clinical data for their model. On other hand, they are collecting data from the patients who had undergone different surgical techniques and who had taken different drug dosages. But effect of those different situations haven't considered in the study. (7) Even though, they suggested ML models for improvise the clinical practices using larger data set for this study may increase the accuracy with high confident. (8) Comparing to our study, they paid more attention on clinical data, used only two ML methods and only three targets were used in their study, where we are using more machine learning methods for more different targets.

ARTICLE 8

NIRMANI B.G.M.
2017/E/074
Week 6

(1) M. Khawar, N. Aslam, R. Mahtab Mahboob, M. Mirza, H. Jahangir and A. Mughal, "Comparative Study of Machine Learning Algorithms in Breast Cancer Prognosis and Prediction", *Comparative Study of Machine Learning Algorithms in Breast Cancer Prognosis and Prediction*, no. 20, pp. 125-133, 2020.

(2) The research is to present a comparative study on machine learning tools with breast cancer and lung cancer data. (3) Both Support Vector Machine and Decision tree methods are used to model the data and their evaluations had done in terms of accuracy, confusion matrix, sensitivity, specificity and precision. The used tools are WEKA, R studio (based on R language), Spyder (Python), Jupyter Notebook (Python). (4) As mentioned earlier, two machine learning methods had used to predict cancer susceptibility, recurrence and survival. (5) Even their main purpose is to compare machine learning tools, also they used two machine learning methods. On other hand they are providing effective factors that can be used to compare among machine learning situations and some of them can be used for our study also. (6) CPU time and Memory usage are used as factors of performance in the study. But it is difficult to provide the same environment every time. Because the background process may affect. (7) Anyway, comparisons on the above scenarios have presented successfully and they have mentioned that Jupyter Notebook tool had given better results in almost all the situations. (8) Rather than paying much attention to the comparison between machine learning methods their aim is to compare the performance of different machine learning tools. So that, only two machine learning methods had been used. And two datasets had used for the research.

ARTICLE 9

NIRMANI B.G.M.
2017/E/074
Week 7

(1) M. Khawar, N. Aslam, R. Mahtab Mahboob, M. Mirza, H. Jahangir and A. Mughal, "Comparative Study of Machine Learning Algorithms in Breast Cancer Prognosis and Prediction", *Comparative Study of Machine Learning Algorithms in Breast Cancer Prognosis and Prediction*, no. 20, pp. 125-133, 2020.

(2) The research is to present a comparative study on machine learning tools with breast cancer and lung cancer data. (3) Both Support Vector Machine and Decision tree methods are used to model the data and their evaluations had done in terms of accuracy, confusion matrix, sensitivity, specificity and precision. The used tools are WEKA, R studio (based on R language), Spyder (Python), Jupyter Notebook (Python). (4) As mentioned earlier, two machine learning methods had used to predict cancer susceptibility, recurrence and survival. (5) Even their main purpose is to compare machine learning tools, also they used two machine learning methods. On other hand they are providing effective factors that can be used to compare among machine learning situations and some of them can be used for our study also. (6) CPU time and Memory usage are used as factors of performance in the study. But it is difficult to provide the same environment every time. Because the background process may affect. (7) Anyway, comparisons on the above scenarios have presented successfully and they have mentioned that Jupyter Notebook tool had given better results in almost all the situations. (8) Rather than paying much attention to the comparison between machine learning methods their aim is to compare the performance of different machine learning tools. So that, only two machine learning methods had been used. And two datasets had used for the research.