International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Cancer Prediction Using Machine Learning Algorithm

Shweta[1], Riya[2], Abhay Kumar[3]

Galgotias University, Greater Noida, India
shwetaverma322001[at]gmail.com

Galgotias University, Greater Noida, India
riya2000jaat[at]gmail.com

Galgotias University, Greater Noida, India
abhaykumar[at]galgotiasuniversity.edu.in

Abstract: ***Cancer has been characterized as a heterogeneous disease consisting of many sub types. The importance of classifying cancer patients into high or low risk groups has lead many research teams to study applications of machine learning techniques. In this work, we present a review of recent ML approaches employed in the modeling of cancer progression.***

Keywords: Machine Learning, Supervised learning, Cancer prediction

1. Introduction

Over the past decades, a continuous evolution related to cancer research has been performed. Scientists applied different methods, such as screening in early stage, in order to find types of cancer before they cause symptoms. With the advent of new technologies in the field of medicine, large amounts of cancer data have been collected and are available to the medical research community. As a result, ML methods have become a popular tool for medical researchers. These techniques can discover and identify patterns and relationships between them, from complex datasets, while they are able to effectively predict future outcomes of a cancer type.

The fundamental goals of cancer prediction and prognosis are distinct from the goals of cancer detection and diagnosis. In cancer prediction/prognosis one is concerned with three predictive foci:

1) The prediction of cancer susceptibility (i.e. risk assessment);
2) The prediction of cancer recurrence and 3) the prediction of cancer survivability. In the first case, one is trying to predict the likelihood of developing a type of cancer prior to the occurrence of the disease. In the second case one is trying to predict the likelihood of redeveloping cancer after to the apparent resolution of the disease. In the third case one is trying to predict an outcome (life expectancy, survivability, progression, tumordrug sensitivity) after the diagnosis of the disease.

In order to predict the various types of diseases, different deep learning & machine learning algorithms are used , such as Support vector machine (SVM), Neural Network (NN), LR, Naive bayes (NB), Fuzzy logic, clustering , transfer learning, ensemble learning, Random forest, Transduction learning, KNN, and Adaboost are mostly utilized in diverse contributions. As per the above discussion, it would be nice to have such a system that would allow to detect and prevent the cancer at an early stage. This can increase the survival rates for those who are going to effect of cancer.

2. Objective

There is a need to study and make a system which will compare the symptoms of cancer for early notification. This system will save the cost of further treatment. It will aid also aid the physicians analyse the pattern of commonly occurring cancer and cause of recurrence.

3. Technology Used

Machine Learning (ML) is one of the core branches of Artificial Intelligence. It's a system which takes in data, finds patterns, trains itself using the data and outputs an outcome**.** There are many applications in the area of biomedical research where ML fits suitably. ML uses different techniques and algorithm to generalize the biological sample of n-dimensional spaces for a given set of datasets.

a) KNN (K-Nearest Neighbour)

kNN is a data classification algorithm that detects a new case to be with the existing case within a defined area by calculating nearest neighbor with similar features of the case. The value of k would find all the similar existing features case with the new case and surround all the case so that it could be possible to identify the new case for the similar category.

b) Decision Tree

A Decision Tree is to find the possible solutions of a given problem based on certain conditions that takes decision and the solutions are presented graphically for better understanding. It gives a systematic solution and documentation process at each step.

c) Support Vector Machine

SVM is a concept that is used to classify the labelled data and to apply regression analysis on the given dataset.There is a hyperplane to have sets of input data where SVM divides the dataset into two classes as the classification is done on the basis of labelled data in the best possible way. The distant margin is measured between the hyperplane and the nearest data point from either set of classified datasets.

d) Random Forest

A Random Forest Algorithm takes the decision tree concept a step further by producing a big number of decision trees to make a forest. These trees are reformed on the basis of selection of data and variables randomly.

e) Naïve Bayes Classifier

The Bayes' theorem describes the NB classifier with independence assumption between predictors. Bayesian classifiers are statistical classifiers based on probability. This is particularly used for large datasets where decision may be taken for filtration on the basis of data points of different class and attributes.

4. Data Set Used

A. List of types of cancer being predicted

Ovarian cancer	Breast cancer
Liver cancer	Oral cancer
Colon cancer	Prostate

B. Algorithms used

Naïve Bayes	Random Forest
Support Vector Machine	Fuzzy Neural Network
Decision Tree	Simple Logistic
ANN	kNN

REF	ML Technique	Type of cancer	Sample	Accuracy
[2], 2008	SVM-REF	Breast cancer	84	100%
		Colon cancer	45	
		Lung cancer	72	
		Ovarian cancer	39	
[4], 2011	SVM	Liver cancer	156	100%
[8], 2012	PSO-KNN	Breast cancer	97	100%
	PSO-SVM			
[10], 2014	SFLLF	Colon	62	100%
		Prostate	136	

C. Various ML Techniques applied on cancer for 100% accuracy

The above table shows the list of respective classifier and feature selection techniques that are used in several research papers for the prediction of cancer in different aspects. This table also shows that the dataset samples taken to find the accuracy depend upon the instance of data and k-fold cross validation.

5. Results

The motive of this study was to understand and improve the process of cancer prediction and increase its efficacy and also to do a comparative study between algorithms and find the best suited algorithm. The most recent works relevant to cancer prediction/prognosis by means of ML techniques are presented. Among the most common applied ML algorithms relevant to the prediction outcomes of cancer patients, we found that SVM and KNN classifiers were widely used. On the basis of research for ovarian cancer prediction K- means with Harmony search proved to be more effective with the accuracy rate of 97%, for liver cancer prediction Fuzzy Neural Network proved to be more effective with the accuracy rate of 95.45%, for colon cancer prediction Fuzzy Granular Support Vector Machine proved to be more effective with the accuracy rate of 99.71% and for Breast cancer prediction Random Forest proved to be more effective with the accuracy rate of 99.24%.

6. Related Work

Sung-Bae Cho et al [2] explored many features and classifiers to select extracted genes from microarray which have many noises. They have taken three datasets: Leukemia cancer, Colon cancer and Lymphoma cancer which has the sample 72, 62 and 47 respectively. They have used Pearson's and Spearman's correlation coefficients, Euclidean distance, information gain, mutual information and signal to noise ratio for feature selection. For classification, they used MLP, kNN, SVM and SOM. They performed experimental results with all the dataset given and shown the best result for accuracy is 97.1% on Leukemia dataset with all the classifier shown above.

Rui Xu et al. [3] used PSO for the prediction of patient survival using gene expression data. PSO reduces the dimensionality by implementing Probabilistic NN. The experimental results of PSO/PNN on B-cell Lymphoma dataset of 240 sample was more effective up to 80% accuracy in survival prediction.

Mehdi Pirooznia et al. [6] studied many classification methods and feature selection methods for expressed genes in microarray data. They were able to find the efficiency of the various classification methods like: SVM, Radial Basic Function, Mult-Layer Perceptron, DT and RF. The 10-fold cross validation had been applied to calculate the accuracy of the classifier includes: Kmeans. Further the efficiency of the feature selection methods was measured by SVM-RFE, Chi-Squared and Correlation based feature selection (CFS). In the conclusion, the authors got the best efficient result by SVM-RFE feature selection methods with 100% accuracy to identify the significant genes.

7. Conclusion

This paper presented an extensive review of various ML classification techniques for the prediction of cancer and standard datasets have been used in wide variety of cancer such as ovarian cancer, breast cancer. A detailed list of results found by many researchers has been tabulated to

solve the problems by various computational intelligence techniques. The most successful approach is SVM and combination of SVM technique which gave up to 100% accuracy on a smaller number of training datasets which is not a good prediction in case with large datasets. However, options are available for the possibilities of improvement of predicting the cancer at an early stage. There are many datasets available to explore more for the same. There are large numbers of cancer types available with unknown functions.

References

[1] Y. Hu, K. Ashenayi, R. Veltri, G. O'Dowd, G. Miller, , "A Comparison of Neural Network and Fuzzy c-Means Methods in Bladder Cancer CellClassification",http://ieee xplore.ieee.org/lpdocs/epic03 wrapper.htm?arnumber=37 4891

[2] Mehdi Pirooznia, Jack Y Yang, Mary Qu Yang, Youping Deng, "A comparative study of different machine learning methods on microarray gene expression data", BMC Genomics, Open Access BioMed Central, 2008, International Conference on Bioinformatics & Computational Biology (BIOCOMP'07) Las Vegas, NV, USA. 25-28 June 2007, DOI: 10.1186/1471-2164-9-S1-S13

[3] Rui Xu, Xindi Cai, Donald C. Wunsch II, "Gene Expression Data for DLBCL Cancer Survival Prediction with A Combination of Machine Learning Technologies", 27th Annual Conference, 2005, pp 894-897, ISBN 0780387406

[4] P. Rajeswari, G. Sophia Reena, "Human Liver Cancer Classification using Microarray Gene Expression Data", International Journal of Computer Applications (0975–8887) Volume 34–No.6, November 2011

[5] V. Bevilacqua G. Mastronardi, F. Menolascina P. Pannarale, A. Pedone, "A Novel MultiObjective Genetic Algorithm Approach to Artificial Neural NetworkTopologyOptimisatio n: The Breast Cancer Classification Problem", Vancouver, BC, Canada July 16-21, IEEE 2006, pp 1958- 1965, ISBN 0780394909

[6] Hiro Takahashi, Yasuyuki Murase, Takeshi Kobayashi, Hiroyuki Hondaa, "New cancer diagnosis modeling using boosting and projective adaptive resonance theory with improved reliable index", ELSEVIER Biochemical Engineering Journal 33 (2007)

[7] Mehdi Pirooznia, Jack Y Yang "A comparative study of different machine learning methods on microarray gene expressiondata",(BIOCOMP'07) Las Vegas, NV, USA. 25

[8] Barnali Sahu, Debahuti Mishra, "A Novel Feature Selection Algorithm using Particle Swarm Optimization for Cancer Microarray Data", International Conference on Modeling Optimization and Computing (ICMOC-2012), ELSEVIER Procedia Engineering 38 (2012) pp 27–31

[9] Jayashree Dev, Sanjit K Dash, Swet Dash, Madhusmita Swain, "A Classification Technique for Microarray Gene Expression Data using PSO-FLANN", International Journal on Computer Science and Engineering (IJCSE)

[10] C Gunavathi, K Premalatha, "A Comparative Analysis of Swarm Intelligence Techniques for Feature Selection in Cancer

[11] Classification", Hindawi Publishing Corporation The Scientific World Journal, Volume 2014, Article ID 693831, pp 1-12