International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V67I3P119

Volume 67 Issue 3, 99-101, March 2019
© 2019 Seventh Sense Research Group®

Original Article

Classification of Cancerous Profiles using Machine Learning Algorithms

Yaramala Sushma[1], Vagolu S Prasad Babu[2], Vanitha Kakollu[3]

[1,2] MCA Student & Department of Computer Science & GITAM (Deemed to be University), Visakhapatnam, India
[3] Assistant Professor & Department of Computer Science & GITAM (Deemed to be University), Visakhapatnam, India

Abstract* - Many existing methods are available for lung cancer identification. This type of treatment recommended for an individual is influenced by factors such as cancer type, the severity of cancer (stage), and, most importantly, genetic heterogeneity. In such a complex environment, the targeted drug treatments are likely to be irresponsive or respond differently. To study anticancer drug response, we need to understand cancerous profiles. These cancerous profiles carry information to explore the underlying factors responsible for cancer growth. Hence, there is a need to analyze cancer data to predict optimal treatment options. Analysis of such profiles can help predict and discover potential drug targets and drugs. In this paper, the main aim is to provide a machine learning-based classification technique for cancerous profiles.

Keywords* - Cancer, Machine Learning, SVM, Random Forest, k NN, Genes, Drug prediction.

I. INTRODUCTION

All living organisms are made up of a basic unit of life, called Cells. Individual cells describe a completely complex functionality. Genes are the carrier of genetic information within the Cell. The information about the inherited phenotypic traits in living organisms is determined by genes[3]. Genetics is a branch of science that has evolved ever since the study of genes started. Advancement in bioinformatics has raised the patient's life expectancy and boosted the treatment procedure for various chronic diseases. Screening of various diseases like diabetes, cancer and Cancer attack is no more a tedious task. Chip technology in healthcare has provided laboratory on-a-chip devices. These chips help predict the drug responses corresponding to the patient's genetic profile. All these technological advancements in the healthcare industry are helping in the earlier diagnosis and prognosis of stringent diseases like cancer. Genetics identifies which features are inherited and explains how these features pass from generation to generation. Genetics also studies the expression level of the genes to determine the up and down state of the gene. These gene expression data lays the foundation for various kinds of analysis that can perform using statistics and computations[10]. These gene expressions help in pathway analysis, drug target discovery, identifying disease biomarkers. Many researchers are trying their hard to reveal the hidden aspects and networks that can help diagnose and treat diseases like cancer [12]. Machine learning approaches are giving a powerful hand in such a data-driven analysis. Gene expression involves the overall process of information retrieval from the gene, hence helping to synthesize functional products called protein.

II. PROPOSED SYSTEM

The proposed system represents a recent method that improved the algorithm's performance and accuracy in a distributed environment. In this paper, we have analyzed using SVM, Random Forest, Decision Tree, and k-Nearest Neighbour algorithms by applying validation measures. The paper sets itself apart by harnessing the powers of Machine Learning techniques. The paper proposes a system with a strong prediction algorithm, which implements powerful classification steps with a comprehensive report generation module. This paper aims to implement a self-learning protocol such that the past inputs of the disease outcomes determine the future possibilities of cancer to a particular use. The proposed algorithm is divided into two sections. One is Dataset Pre-processing and Classification using Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbour.

A. Support Vector Machine algorithm

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification challenges. However, it is mostly used in classification problems. This algorithm plots each data item as a point in n-dimensional space (where n is several features you have), with the value of each feature being the value of a particular coordinate. Support Vectors are simply the coordinates of individual observation. In this paper, we will mainly consider the input based upon Support Vector Machine as training data[4], and testing data is decision value. In this method, we consider the following steps like Load Dataset, after loading the dataset will Classify Features (Attributes) based on class labels then estimate Candidate Support Value[9], like the condition is While (instances!=null), Do condition if Support Value=Similarity between each instance in the attribute then finding the Total Error Value. Suppose if any instance < 0, then the estimated decision value = Support Value\Total Error, repeated for all points until it is empty. Therefore mainly, we have calculated the entropy and Gini index.

B. Random Forest Algorithm

The Random Forest algorithm is a learning method that operates by constructing a fixed number of decision trees at training time. Random forests have an implementation of the bagging technique. Some samples are repeatedly selected from the training set to fit the chosen models; then, the classification is made using a majority voting scheme between all the models. In this paper, we will consider a Bootstrap method for data Set. The bootstrapping estimation method is used to make predictions on a data set by re-sampling it and creating decision trees. We repeat this until we get the Predictive outcome of a new data point and evaluate the model.

C. Decision Tree Algorithm

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics and machine learning. Tree models where the target variable, which can take a discrete set of values, are called classification trees[5][1]; in these tree structures, leaves represent class labels, and branches represent conjunctions of features that lead to those class labels. This paper deals with decision trees in machine learning.

This paper mainly considers the decision tree based on training data, testing data, and decision value. This method will consider the following steps to load the train and test data. Then begin Tree= {} If (D is' pure") ‖ (other stopping criteria met) then terminate all the attributes a, a belongs to D. Then Compute criteria to impurity function then we will split on Abest= Best attribute according to above-computed criteria. Then Tree= Create a decision node that tests Abest in the root. Dv= Induced sub-datasets from D based on Abest. For all Dv do TREEV = J48(Dv). Attach Tree v to the corresponding branch of Tree.

D. k-Nearest Neighbour Algorithm

The model representation for k NN is the entire training dataset. It is as simple as that. k NN has no model other than storing the entire dataset; Efficient implementations can store the data using complex data structures like k-d trees to make lookup and match new patterns during prediction efficient. Because the entire training dataset is stored, you may want to think carefully about the consistency of your training data. It might be a good idea to accurate it, update it often as new data becomes available, and remove erroneous and outlier data. Then consider the following steps Load the training and test data and choose the value of K[10]. Find the Euclidean distance to all training data points for each point in test data. And store the Euclidean distances in a list and sort it. Later choose the first k points, then assign a class to the test point based on most classes.

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

III. METHODOLOGY

We used a dataset from kaggle.com and UCI repository for various disease-based datasets we tested. The diseases that are lower in progress but higher in duration are termed non-communicable diseases. There are mainly four classifications among no communicable diseases. The collected data were used to create a structured database system. The fields were identified, duplications were extracted, missing values were filled, and the data were coded according to attribute domain value. After data cleaning, the number of cases was reduced mainly due to the unavailability of clinical results [14]. Attribute domain values are provided by practicing physicians. Data preparation requires approximately 80% of the time. Once data is gathered, it needs to be pre-processed, cleaned, constructed, and formatted in a style that SVM comprehends and can work with. The dataset of the proposed work has been taken from the universal genomics of drug sensitivity repository (cancer X-gene org.).In this paper, we have used four classification algorithms. The algorithms are support vector machine, random forest, Decision tree, and kNN. The proposed model needs to be trained and tested under various conditions by altering SVM parameters to obtain correctness. In addition, we consider that the model's accuracy is maximum. From the collected data, 70:30 will be considered train and test data for the model, respectively. In case of necessity, there must be a need to improve the algorithms being used. Results, analysis, and predictions have been evaluated using precision, recall, F-measure, and accuracy. Precision is the fraction of retrieved data that is useful for the query. The recall is the fraction of relevant data for the query that is effectively retrieved. F-measure is a measure that sums up precision and recall, and accuracy is the proximity of a computation to the true value, which is calculated by taking true positive and true negative with a fraction of true positive, true negative, and false-positive with a false negative.

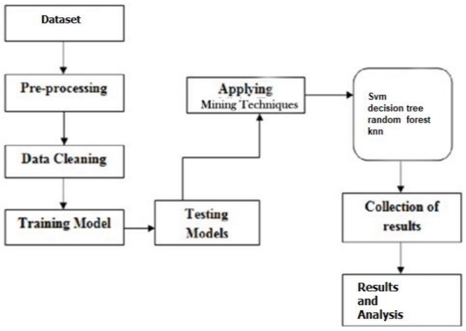

Fig. 1 Integrated architecture for lung cancer

We took the lung dataset and pre-processed it after that data cleaning was done. Then used the training model, and we tested the models[12]. We have applied mining techniques like SVM, decision tree, random forest, and kNN. We have collected the results from this, and finally, an analysis has been done.

IV. RESULT

In this bar chart, we got comparison models between algorithms. SVM got 98% accuracy. SVM got the highest accuracy. Next, Random Forest got 96% accuracy.

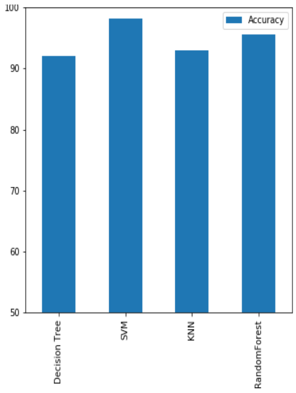

Fig. 2 Comparative models with accuracy

V. CONCLUSION & FUTURE WORK

We have seen that clinical data mining is a recent research field that aims to utilize data mining and machine learning capabilities to reveal biological patterns. Moreover, the oncogenomics research domain aims at identifying and analyzing cancer-related genes and thus helps in diagnosis at the genotype level. Although various approaches have been proposed in the literature for classification, gene selection remains a major curse. Cancer is a heterogeneous disease that consists of various subtypes[8]. Hence, there is an urgent need to develop systems or methods that can help in the early diagnosis and prognosis of cancer type. The past decade has evolved various new approaches related to cancer research. Scientists have used various biological and computational techniques to detect cancer types early. The collection of large cancer data repositories has hiked the research in this domain. Various machine learning approaches have been used to predict cancer.

The proposed technique is to solve the classification problem for cancerous genomic profiles. Our technique is based on the concept of utilizing SVM, kNN[13], Decision Tree, and Random Forest machine learning algorithm[7]. Result provides a comparative analysis of model performance when the sample size is varied. As the sample size increases, the model performance also increases, which shows a positive aspect of the robustness and adaptivity of the model. In the future, this approach can be extended to implement an integrative framework for anticancer drug prediction.

REFERENCES

[1] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O., Akinjobi J. “Supervised Machine Learning Algorithms: Classification and Comparison," Volume-48, pp. 128-138, 2017.

[2] Y, Mangasarian OL, Wolberg WH. 2000. Breast cancer survival and chemotherapy: A support vector machine analysis. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, 55:1- 20.

[3] F.Bray, P. McCarron, and D. M. Parkin, “The changing global patterns of female breast cancer incidence and mortality” Breast Cancer Res. vol. 6, pp. 229-239, 2004.

[4] R.L.Birdwell, D. M. Ikeda, K. D. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computeraided detection" Radiology. Vol. 219, pp. 192-202, 2001.

[5] N.Guler, E. Ubeyli, and I. Guler, "Recurrent neural network employing Lyapunov exponents for EEG signals classifications" Expert systems with Applications. Vol. 29, pp. 506-514, 2005.

[6] J.Abonyi and F. Szeifert "Supervised fuzzy clustering for the identification of fuzzy classifiers" Pattern Recognition Letters. Vol. 24, pp. 2195-2207, 2003.

[7] R.Setiono, “Generating concise and accurate classification rules for breast cancer diagnosis” Artificial Intelligence in Medicine. vol. 18, pp. 205-219, 2000

[8] L.Hadjiiski and B. Sahiner, “Advances in computer-aided diagnosis for breast cancer. Curr. Opin. Obstet. Gynecol. vol. 18,

[9] J.R.Duda and P. Hart, Pattern Classification and Scene Analysis. John-Wiley, 1973.

[10] D.Specht, “Probabilistic neural networks for classification, mapping or associative memory” in Proc. IEEE Int. Conf. Neural Network, pp. 525- 532, 1988.

[11] D.Delen and G. Walker, “Predicting breast cancer survivability: a comparison of three data mining methods” Artificial Intelligence in Medicine, vol. 34, pp. 113-127, 2005.

[12] A.Hong and S. Cho, "Lymphoma cancer classification using genetic programming with SNR features" Lecture Notes on Computer Science. Vol. 3003, pp. 78-88, 2004.

[13] D.Specht, “Probabilistic neural networks for classification, mapping or associative memory” in Proc. IEEE Int. Conf. Neural Network, pp. 525- 532, 1988.

[14] V.Vapnik, The Nature of Statistical Learning Theory. Springer, 1995.