Cancer Profiling using Machine Learning

Prof. Veena K R[1] Aayush Shrivastava[2] Dawan P Swamy[3] Jithendra C[4] Mohit Kumar[5]
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Sapthagiri College of Engineering (VTU), Bengaluru, India

Abstract— There are choices accessible for tumor treatment. The sort of treatment suggested for an individual is impacted by different factors, for example, disease compose, the seriousness of malignancy (arrange) and most essential the hereditary heterogeneity. In such a perplexing domain, the focused-on medication medicines are probably going to be unmoved or react in an unexpected way. To examine anticancer medication reaction, we must comprehend cancerous profiles. These cancerous profiles have data which can uncover the fundamental elements in charge of tumor development. Consequently, there is having to dissect malignancy information for anticipating ideal treatment choices. Analysis of such profiles can help to predict and discover potential drug targets and drugs. In this paper the principle point is to give machine learning based order method for cancerous profiles.

Keywords: Cancer Profiling Using Machine Learning, Neural Network

I. Introduction

We all living organisms are made up of basic unit of life, called Cells. Individual cells describe a completely complex functionality. What makes them more interesting, are genes. Genes are the carrier of genetic information within the Cell. The information about the inherited phenotypic traits in living organisms is determined by genes. Genetics is a branch of science that has evolved ever since study of genes started. Advancement in bioinformatics has raised the patient's life expectancy and boosted the treatment procedure of various chronic diseases. Screening of various diseases like diabetics, cancer and heart attack is no more a tedious task. Chip innovation in healthcare has provided laboratory on-a- chip devices. These chips help in predicting the drug responses corresponding to patient's genetic profile. All these technological advancement in healthcare industry are helping in earlier diagnosis and prognosis of stringent diseases like cancer. Genetics explains and identifies which features are inherited and how these features pass from generation to generation.

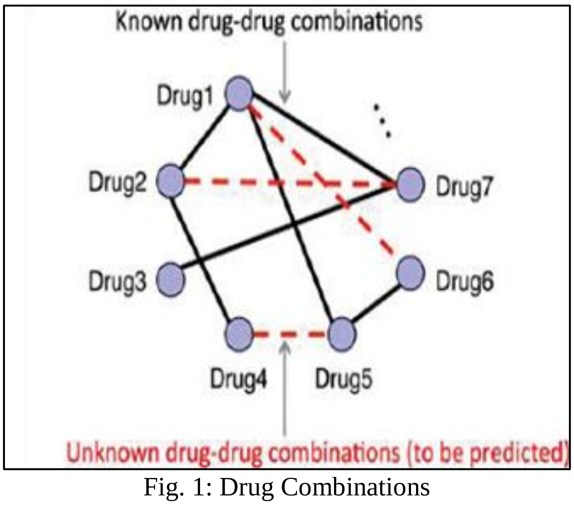

Known drug-drug combinations
Drug1
Drug2
Drug3
Drug4
Drug5
Drug6
Drug7
Unknown drug-drug combinations (to be predicted)

Fig. 1: Drug Combinations

Genetics also studies about the expression level of the genes, to determine the up and down state of the gene. These genes expression data lay the foundation of various kinds of analysis that we can perform using statistics and computations. This expression helps in pathway analysis, drug target discovery, identifying disease biomarkers. Researchers and Scientists are trying their hard to reveal the hidden aspects and networks, which can help in proper diagnosis and treatment of diseases like Cancer. Data Mining and Machine learning approaches are giving a powerful hand in such a data driven analysis. Gene expression helps in the synthesis of functional products called protein because it involves the overall process of information retrieval from the gene. The amount of mRNA produced by the gene at an instant of time, corresponds to the gene expression value. These expression values may alter depending upon the environment, any biological regulator and biological pathways involved. The process of mapping information from genes to proteins synthesis is carried by agent called mRNA. Transcription and Translation are the two sub processes that are involved in this process. Transcription involves duplication of gene sequence in the form of RNA. Once the genetic coding is copied on messenger RNA (mRNA) then it exists the nucleus and enters cytoplasm and eventually encoded protein is synthesized. Translation involves interpretation of mRNA sequence of amino acids to synthesize proteins.

II. Literature survey

The study of profiles of genes helps in predicting the gene functionality and the identification of tumor. Further it helps in cancer classification and also in drug discovery [1]. Active research on classifying cancer and the subtypes of it has gained popularity in recent years. Vandana et al. has proposed an approach for fuzzy clustering using large graphs [10]. Traditionally, it was believed that cancer can be curable based on its anatomical origin. However, it was found out that if two persons are diagnosed with the same cancer type, the drug that works well in the case of one person might not work well in the case of others. The genetics of the human body play an important role in the case of such heterogeneity [2]. Hence optimal drug prediction based on genomic profiles is one of the most emerging topics in the case of cancer bioinformatics [7].

The process of micro array data classification is used to predict the category of a given sample[3]. It builds a classifier and classify given data points into predefined diseases classes. A number of statistical learning approaches have been defined like nearest neighbor classification [4], least square and regression modelling [5], discriminatory methods [6] and weighted voting [8]. In order to attain successful diagnosis of cancer, a lot of work still has to be undergone in the process but the cure of it still remains a mystery as so far no proper diagnosis method has been found. Various research projects have been undertaken in order to strengthen the platform of fighting the disease. Pan-Cancer project [9] is one of the project which analyzed the molecular instability in wide range of tumor cells. The data from each tumor type was combined in order to contribute to the field of cancer research.

III. Existing System

It has been noticed that the drugs have largely been predicted based on anatomical origin. Hence there has been heterogeneous behavior towards the drug target therapy due to the genomic variations. As a result, the drugs respond differently even if two persons are diagnosed with the same cancer type. The molecular classification of cancer and predicting category of the microarray data is a tedious task. What algorithms must be used so that the classification process becomes more effective and what drugs should be administered in order to cure the disease is a major concern. The molecular classification of cancer and predicting category of the microarray data is a tedious task. What algorithms must be used so that the classification process becomes more effective and what drugs should be administered in order to cure the disease is a major concern.

IV. Proposed System

The proposed system consists of a hybrid algorithm which consists of the following three sections:

- Dataset Pre-processing
- Classification using neural network
- Classification using Support Vector Machine

A. *Dataset Pre-processing*

Data preprocessing transforms raw data into an understandable format which is also a data mining technique. A column of each dataset is a particular attribute on the basis of which preprocessing has to be performed. The steps involved are Data Cleaning, Data Integration, Data Transformation and Data Reduction.

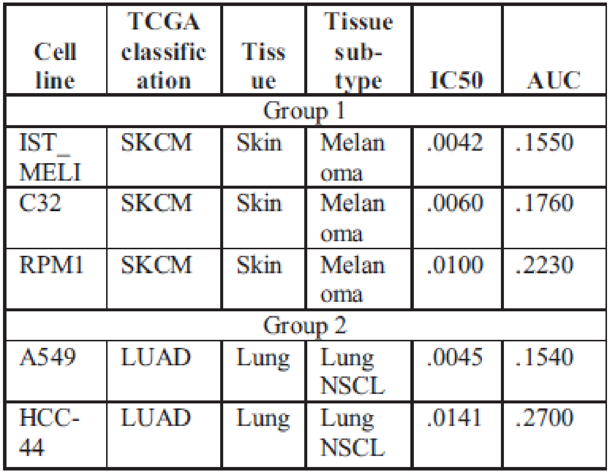
A series of steps during preprocessing that data undergoes through are:

- Data Cleaning: A process of filling in missing values, resolving the inconsistencies in the data and smoothing the noisy data etc.
- Data Integration: Putting together of data with different representations and resolving conflicts within the data.
- Data Transformation: Data is normalized, aggregated and generalized.
- Data Reduction: A reduced representation of the data in a data warehouse is presented in this step.
- Data Discretization: By dividing the range of attribute intervals it reduces the number of values of a continuous attribute.

B. *Classification using Neural Network*

Neural networks are a large number of highly inter connected processing elements working in multi-layered structures that receive inputs, processes the inputs and produces the outputs. The output of one layer acts as the input to the next intermediate layer. It follows that:

1) Weighted sum of first hidden layers(say n1 and n2)
2) Apply the activation function
3) Calculate the weighted sum of node 3
4) Derive the final output



Cell line	TCGA classification	Tissue	Tissue sub-type	IC50	AUC
Group 1					
IST_MELI	SKCM	Skin	Melanoma	.0042	.1550
C32	SKCM	Skin	Melanoma	.0060	.1760
RPM1	SKCM	Skin	Melanoma	.0100	.2230
Group 2					
A549	LUAD	Lung	Lung NSCL	.0045	.1540
HCC-44	LUAD	Lung	Lung NSCL	.0141	.2700

Table 1: Grouping of Tissue types

The first four records belong to ‘melanoma’ and has been put together in Group 1 and the next two records belong to ‘Lung NSCL’ and has been put together in Group 2 as per the neural network architecture. The data at input layer is converted to the data at hidden layer taking 30 neurons into account. If the number of rows in training data and group is same, then the neural network will support true training. The neural network is a trained architecture and can be used for classification of the cancerous profiles at the tissue level.

C. *Classification using Support Vector Machine*

Support-vector machines are supervised machine learning techniques with associated machine learning algorithms that analyze data used for binary classification and regression analysis. It also uses a kernel vector used for separating the classes. The SVM in the proposed work is used for target classification and it maps the drug to a particular class and in our proposed work the target drug is mapped to a class grouped by neural network algorithm.

V. Result and Analysis

A. *Prediction of Drug:*

After applying the necessary algorithms and taking the required inputs like TCGA Classification, tissue, tissue subtype, IC50 and AUC, the target drug can be predicted.

B. *Accuracy:*

The accuracy can be obtained with the help of a confusion matrix. A confusion matrix is used to describe the performance of a classification model on a set of test data for which the true values are known. It is calculated by taking true positive and true negative with a fraction of true positive, true negative and false positive with false negative.

Accuracy=(Tp+Tn)/(Tp+Tn+Fp+Fn), whereTp=True positive, Tn=True negative, Fp=False positive, Fn=False negative.

VI. Conclusion and Future Work

Dissecting growth-related qualities and subsequently helps in analysis at genotype level. Different methodologies have been proposed in writing for order however quality choice still remains a noteworthy major curse. Cancer is a heterogeneous infection which comprises of different subtypes. Subsequently, there is critical need to create identify techniques that can help in early determination and anticipation of growth compose. Past decade has developed different new methodologies identified with cancer research. Different machine learning approaches have been utilized to anticipate if tumor is dangerous or not. The proposed procedure is an endeavor to take care of order issue for dangerous genomic profiles. Our strategy depends on idea of using SVM and NN machine learning calculation. Result gives near investigation of model execution when the example measure is differed. As the example estimate increment display execution additionally expands, which indicates positive angle towards the strength and adaptivity of the model. In future, this methodology can be stretched out to actualize integrative system for anti- cancer drug prediction.

References

[1] SantanuGhorai, Anirban Mukherjee, Sanghamitra Sengupta and Pranab K Dutta, “Cancer classification from gene expression data by nppc ensemble,” IEEE Transactions on Computational Biology and Bioinformatics (TCBB), Vol. 8, No. 3, pp. 659–671, 2011.

[2] Alexandre R Zlotta, “Genome sequencing identifies a basis for everolimus sensitivity,” European urology, Vol. 64, No. 3, pp. 29-33, 2013.

[3] P Ganesh Kumar, T Aruldoss Albert Victoire, P Renukadevi and Durairaj Devaraj, “Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm,” Expert Systems with Applications, Vol. 39, No. 2, pp. 1811–1821, 2012.

[4] Liwei Fan, Kim-LengPoh, and Peng Zhou “A sequential feature extraction approach for naive bayes classification of microarray data,” Expert Systems with Applications, Vol. 36, No. 6, pp. 9919–9923, 2009.

[5] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing and Mark A Caligiuri “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” Science, Vol. 286, No. 5439, pp. 531–537, 1999.

[6] Gersende Fort and Sophie Lambert-Lacroix, “Classification using partial least squares with penalized logistic regression,” Bioinformatics, Vol. 21, No. 7, pp. 1104–1111, 2005.

[7] Jianting Sheng, Fuhai Li, and Stephen TC Wong, “Optimal drug prediction from personal genomics profiles,” Biomedical and Health Informatics, Vol. 19, No. 4, pp. 1264–1270, 2015.

[8] Leping Li, Clarice R Weinberg, Thomas A Darden and Lee G Pedersen, “Gene selection for sample classification based on gene expression data study of sensitivity to choice of parameters of the GA/KNN method,” Bioinformatics, Vol. 17, No. 12, pp. 1131–1142, 2001..

[9] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander and Joshua M Stuart, “The cancer genome atlas pan-cancer analysis project,” Nature Genetics, Vol. 45, No. 10, pp. 1113–1120, 2013

[10] Vandana Bhatia and Rinkle Rani, “A parallel fuzzy clustering algorithm for large graphs using Pregel,” Expert Systems with Applications, Vol. 78, pp-135-144, 2017.