Classification of Cancerous Profiles using Machine Learning

Aman Sharma
Department of computer Sc. & Engg.,
Thapar University, Patiala, INDIA
amans.3008@gmail.com

Rinkle Rani
Department of computer Sc. & Engg.,
Thapar University, Patiala, INDIA
raggarwal@thapar.edu

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

Keywords: Cancer, Clustering, Machine Learning, Genes, Drug Prediction

I. Introduction

We all living organisms are made up of basic unit of life, called Cells. Individual cells describes 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 technology in healthcare has provided laboratory on-a- chip devices. These chips helps 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 identifies which features are inherited, and explains how these features pass from generation to generation. Genetics also studies about the expression level of the genes, to determine the up and down state of the gene. These genes expression data lays the foundation of various kinds of analysis that we can perform using statistics and computations. These 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 involves the overall process of information retrieval from the gene, hence helps in the synthesis of functional products called protein. The amount of mRNA produced by the gene at a particular 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 so as to synthesize proteins.

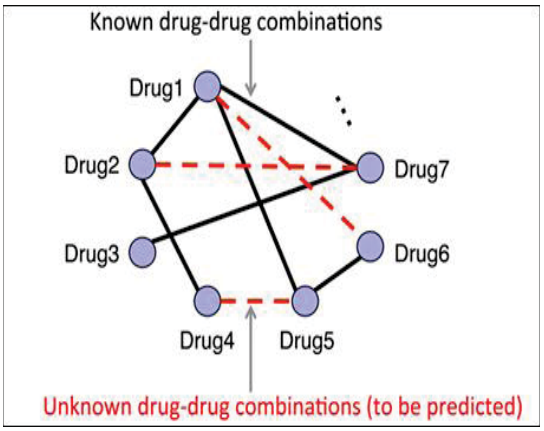

Figure 1: Drug Combinations

Gene's classification and clustering methods are the integral part of any analysis in the microarray data. Recently various classification algorithms have been proposed such as Neural Nets, MLP Neural Nets, Bayesian, Decision Tree and Random Forest. Validation of these algorithms is done to check the robustness of trained model by K-fold validation.

978-1-5386-3446-2/17 $31.00 © 2017 IEEE
DOI 10.1109/MLDS.2017.6

Cancer patients often show heterogeneous drug responses such that only a small subset of patients is sensitive to a given anticancer drug. With the advent of next generation sequencing, huge amount of genome profiling data has driven the research on uncovering the genetic mutations and expressed genes of individual patients. These genetic mutations and expression are responsible for such a wide spread heterogeneity in drug responses. Drug related sensitivity is determined by evaluating the changes at genetic level for each specific drug. Optimal drug prediction is a key component for oncology precision medication. Traditionally all cancers are classified based on their anatomical origin. But recent research shows that cancer patients of same type can show heterogeneous behavior towards target drug therapy. It has been seen that computational approaches such as machine learning can help to predict the potential drugs and their targets. Machine learning algorithms facilitates adaptive learning and hence helps in predicting anti-cancer drug treatment and classifying cancer patients. So, this paper will review how machine learning algorithms are effective in classification of cancerous profiles. We are classifying cancer cell-lines based on their genetic similarity and the type of cancer.

II. Literature Survey

Expression profiles of thousands of genes are used for predicting the functionality of genes, finding gene regulatory systems, tumour sub-type identification. Further it helps in drug discovery and also helps in cancer classification [1]. Active research on classifying cancer and its subtype has gained momentum in recent years because of the availability of various open data sources. Bharti Saneja et al. has proposed an approach for outlier detection in heathcare domain [12]. Likewise, Vandana et al. has proposed an approach for fuzzy clustering using large graphs [11].

Traditionally, cancer sub-types were determined based on their anatomical origin. But recent research shows that cancer patients of same type can show heterogeneous behaviour towards target drug therapy. Genomic variations and instability are the core end responsible factors for such heterogeneity [2-3]. Hence optimal drug prediction using genomic profiles is an active research topic in the field of cancer bioinformatics [10].

Classification of microarray data is a supervised learning which helps in predicting the category of a given sample [4]. It builds a classifier model from the labelled gene expression data and hence classify given data points into predefined diseases classes. Various statistical approaches have been defined in literature like nearest neighbour classification [5], least square and regression modelling [6], discriminatory methods [7] and weighted voting [8]. Successful diagnosis of disease like cancer is a tedious issue to look upon and hence raises a serious issue for future therapy. Although various classification techniques haven been proposed for cancer diagnosis, still no proper diagnosis methodology has been developed.

Traditionally all cancers are classified based on their anatomical origin. But recent research shows that cancer patients of same type can show heterogeneous behaviour towards target drug therapy. Genomic variations and instability are the core end responsible factors for such heterogeneity [2-3]. Various cancer research projects have contributed to the present research in this field. Pan-Cancer project [9] is one of the project which analysed the molecular instability in wide range of tumour cells and combined the data from each tumour type to foster the cancer related research.

III. Proposed Methodology

The proposed methodology encompasses of hybrid algorithm which contains inner and outer classification. The proposed algorithm is divided into three sections:

a) Dataset Pre-processing

b) Clustering using Neural Network

c) Classification using Support Vector Machine

Pseudo-Code-1

```
1.  createclusters(alldata)
2.  Classes{1} = data(1,4)  // The first subtissue is
                               placed in the first class.
3.  rcount = 2
4.  root_count = 0
5.  for k=1:row_count_data
6.        edata =data(K,4)
7.        for i=1 : row_count_data
9.              if edata == data(K,4)
10.                   C1 = data(K,2)
11.                   C2 = data(K,3)
12.                   S = find classes {:,2-3}
13.                   Is isempty (S)
14.                   Classes {rcount,:}=data(K,:}
15.             End if
16.       End for
17. End for
```

The dataset of the proposed work has been taken from universal genomics of drug sensitivity repository (cancer X-gene org.). The compound values of four targets are, namely, TAKI (MAP 3K7), HSP70 [PARP1, PARP2] and FLT3. On an average, each

compound contains 900 AUC values. Each tissue has different IC_{50} and AUC value. The tissue values have been utilized as inner clustering in the proposed algorithm.

A. Pre-processing

The fourth column of each dataset is the tissue on the basis of which preprocessing has to be performed. The Pseudo-Code description of the preprocessing is given above.

Pseudo-Code-2

```
1.  Neuralarch = function createcluster(org_data)
2.  [row,cols] = Size (org_data)
3.  Group= [ ]    // Initializing the group parameters
4.  Groupname = [ ]
5.  Grpcount = 0  Traindata = [ ] record_count = 1
6.  Group[0] = ord_data(1).issuenumber
7.  Groupname[0] = org_data(1).issuename
8.  Currentgroup = groupname[0]
9.  for i = 1:rows
10.     if (ord_data (i).Tissue name==Currentgroup)
11.             Traindata[recordcount,0:cols]
12.              Record_count = record_count+1
13.             Group(record_count) = grpcount
14.     Else
15.              Groupcount = groupcount+1
16.             Traindata[recordcount,0:colsRecord_count]
17.             Group(record_count)=grpcount
18.      End if
19. End for
```

B. Feed forward Neural Network

In feed forward neural networks output of one layer act as input to intermediate next layer without making any acyclic dependency. Output is generally obtained from final network layer. It can be written as below:

$[x, t] = simplefit\_dataset;$
$net = feedforwardnet(10);$
$net = train(net, x, t);$
$view(net);$
$y = net(x);$
$perf = perform(net, y, t);$

Algorithmic description of Back propagation feed forward network

There can be multiple input variables ($x_1, \ldots, x_m$) corresponding to input layer with multiple nodes. Input layer just passes the input data to intermediate processing nodes (nodes 1, 2 and 3) and mathematically can be written as below:

1. $weighted\ sums\ of\ first\ hidden\ layer$:

$$n_3 = w_{13}x_1 + w_{23}x_2$$
$$n_4 = w_{14}x_1 + w_{24}x_2$$

2. $Apply\ the\ activation\ function$:

$$\tanh()$$

3. $Calculate\ the\ weighted\ sum\ of\ node\ 5$:

$$n_5 = w_{35}y_3 + w_{45}y_4$$

4. $Then\ Final\ output$

For each set of target based on the inner tissue types, group will be created.

TABLE I. Grouping of Tissue Types

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

The first four records belong to same tissue ‘melanoma’ and hence for neutral, they will be put in group1 and the last two records belong to tissue ‘Lung_NSCL’ & hence they will be put in group 2 as per neural architecture. The neural network object is initialized with the train set and an equal number of group values. There would be 30 hidden neurons for the conversion of input data at input layer to the transformed data at the hidden layer. Now, the certainty of taking neurons is random. It is just an estimate and it is completely dependent upon the data size. The neural network will support true training only if the number of rows in training data and group is same.

The epoch is the total number of iterations which the neural network can run to completely understand the data pattern of the training set. 50 is the maximum number of iteration here but it is not necessary that the network will run all those 50 iterations. There are certain performance measures for the neural network.

If any of the performance measures are satisfied then it would stop the training at that moment only and the network will roll back. The network will choose that value where the MSE would be least.

Here, a trained architecture will be identified and can be used for the tissue classification as described in Pseudo-Code-2.

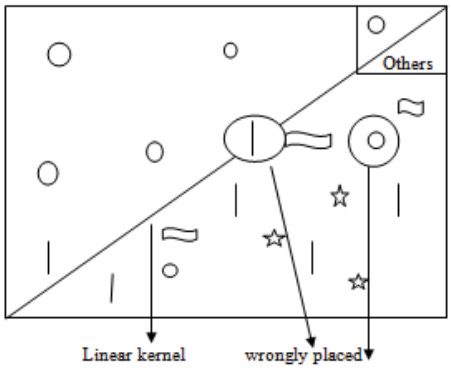

Fig. 2. SVM representation

C. Significance of the Support Vector Machine in proposed algorithm

Support vector machine (SVM) is a binary classification which can identify objects in two categories, first is by itself and secondly, by others. The SVM in the proposed algorithm is utilized by the target [TAK1, HSPO, PARP1-2, FLT] classification. Statistical Learning Theory forms the basis of support vector machine. It helps to learn patterns, predict labels and cluster data points. The mapping is written as below:

$$X \rightarrow Y,$$
$$where\ x \in X\ is\ some\ object\ and\ y \in Y\ is\ a\ class\ label.$$

D. The Classification process

The classification process utilized the trained set for both SVM by utilizing the output of Neural Network. First of all, SVM will be utilized for the target classification of tissues. The SVM Classification uses the following attributes:

a) SVM train set

b) Classified Kernel vector

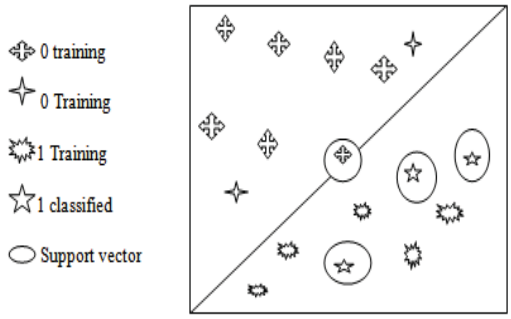

Fig. 3. Classified sample series

Above Figure. 3 represents the classified sample series. At the end of all training, a rule set will be created to predict the drug for the target.

IV. Results and Analysis

The prediction results have been evaluated using following parameters:

Precision: It is the fraction of retrieved data that are useful for the query. It can be described as follows:

$$Precision = \frac{Relevant_{Data} - Retrived_{Data}}{Retrieved_{Data}}$$

Recall: It is the fraction of data that are relevant for the query which is effectively retrieved. It can be described as follows:

$$Recall = \frac{Relevant_{Data} - Retrived_{Data}}{Relevant_{Data}}$$

F-measure: It is measure that sums up precision and recall, and describes as follows:

$$F - measure = 2.\frac{Precision.recall}{Precision + Recall}$$

Accuracy: It 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 false negative.

$$Accuracy = \frac{t_{p+}t_n}{t_{p+}t_n + f_{p+}f_n}$$

Where, $t_p = Truepositive, t_n = Truenegative,\ f_p = falsepositive\ and\ t_n = Truenegative$

Table II: Target Class (SVM)

No. of Samples	Precision	Recall	F-measure	Accuracy
10	0.9266	0.9154	0.9210	92.10
20	0.9383	0.9278	0.9330	93.30
50	0.9397	0.9283	0.9340	93.40
100	0.9418	0.9297	0.9357	93.57

Table. 1 shows the obtained values of precision, recall, F-measure and accuracy by means of number of sample for the target class of SVM. Graphical representation for the same is shown below.

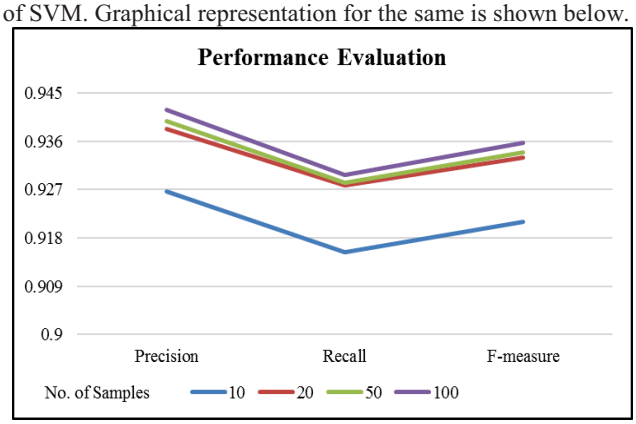

Fig. 4. Performance Evaluation of SVM Classification

Above Fig. 4 shows the precision, recall, and F-measure for 10, 20, 50 and 100 samples using SVM. Numbers of samples are described by different color coded lines. x-axis defines performance parameters and the y-axis shows their corresponding values when the sample size is varied. An average precision is 93.66. With the increase in the number of samples, the precision rate is also increasing. So, it can be said that the appropriate precision is obtained in the proposed work.

Table III: Tissue Class (NN)

No. of Samples	Precision	Recall	F-measure	Accuracy
10	0.8876	0.8454	0.8660	86.60
20	0.8883	0.8476	0.8675	86.75
50	0.8895	0.8565	0.8727	87.27
100	0.8915	0.8593	0.8751	87.51

Precision rate (p) is defined as the ratio of correctly detected samples to the sum of correctly detected samples plus false positives. Its value must be high in order to get good proposed algorithm. The obtained value of (p) for 100 samples is 89.2%.

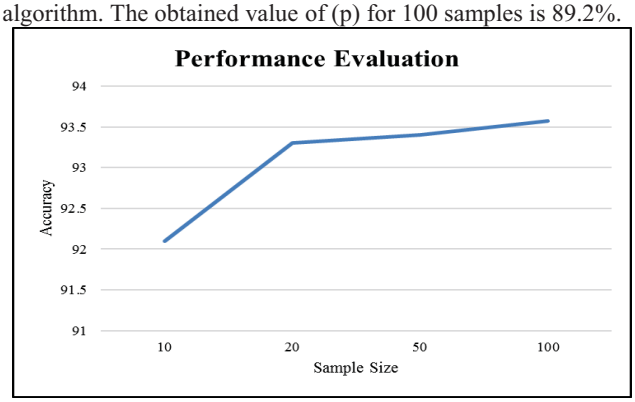

Fig. 5. Accuracy of SVM Classification

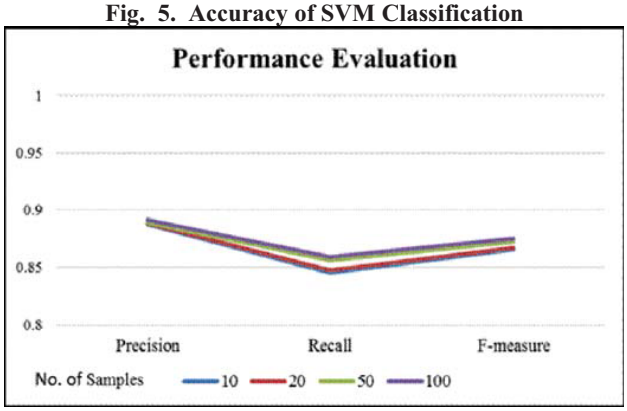

Fig. 6. Performance Evaluation of NN clustering

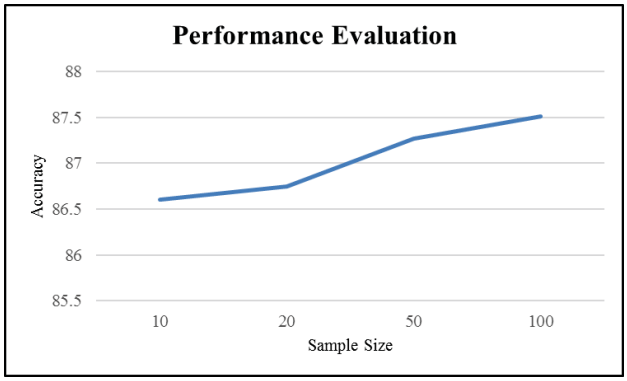

Fig. 7. Accuracy of NN clustering

V. Conclusion

We have seen that clinical data mining is a recent research field that aims to utilize data mining and machine learning capabilities for revealing the biological patterns. Moreover, oncogenomics research domain aims at identifying and analyzing cancer related genes and thus helps in diagnosis at genotype level.

Although various approaches have been proposed in literature for classification but gene selection still remains a major curse. Cancer is a heterogeneous disease which consists of various subtypes. Hence, there is urgent need to develop systems or methods that can help in early diagnosis and prognosis of cancer type. Past decade has evolved various new approaches related to cancer research. Various biological and computational techniques have been used by scientists to early detect cancer type. Collection of large cancer data repositories has hiked the research in this domain. Various machine learning approaches have been used to predict if tumor is malignant or not.

So, in order to address aforementioned challenges the proposed technique is an attempt to solve classification problem for cancerous genomic profiles. Our technique is based on concept of utilizing SVM and NN machine learning algorithm. Result provides comparative analysis of model performance when the sample size is varied. As the sample size increase model performance also increases, which shows positive aspect towards the robustness and adaptivity of the model. In future, this approach can be extended to implement integrative framework for anti-cancer drug prediction.

VI. References

[1] Santanu Ghorai, 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] GopaIyer, Aphrothiti J Hanrahan, Matthew I Milowsky, Hikmat Al-Ahmadie, Sasinya N Scott, Manickam Janakiraman, Mono Pirun, Chris Sander, Nicholas D Socci and Irina Ostrovnaya, "Genome sequencing identifies a basis for everolimus sensitivity," Science, Vol. 338, No. 6104, pp. 221–229, 2012.

[4] 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.

[5] 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.

[6] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasen beek, 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.

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

[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] 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.

[11] 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.

[12] Bharti Saneja, and Rinkle Rani, "An efficient approach for outlier detection in big sensor data of health care," International Journal of Communication Systems, DOI: 10.1002/dac.3352, 2017