Journal of
Computer Science & Systems Biology

Mandal et al., J Comput Sci Syst Biol 2013, 6:6
DOI: 10.4172/jcsb.1000130

Review Article Open Access

An Approach towards Automated Disease Diagnosis & Drug Design Using Hybrid Rough-Decision Tree from Microarray Dataset

Sudip Mandal[1]*, Goutam Saha[2] and Rajat K. Pal[3]

[1]ECE Department, GIMT, Krishna Nagar, India
[2]IT Department, NEHU, Shilong, India
[3]CSE Department, University of Calcutta, Kolkata, India

Abstract

Biological databases related to medical science, containing pathological, radiological and genetic information of patients is undergoing tremendous growth, beyond our analyzing capability. However such analysis can reveal new findings about the cause and subsequent treatment of any disease. Here the genetic information of Lung Adenocarcinoma, in the form of microarray dataset has been investigated which have five different stages. Rough Set Theory (RST) has been used in analysis with an aim to effectively extract biologically relevant information, as RST is a tool that works well in an environment, heavy with inconsistent and ambiguous data, or with missing data and provides efficient algorithms for finding hidden patterns in data. The investigation has been carried out on the publicly available microarray dataset obtained from the GEO profiles at National Centre for Biotechnology Information (NCBI) website. Cross validation of the generated rule sets shows 100% accuracy. Now to extract the hidden biological dependencies between responsible genes, Decision Tree is used at consecutive two stages of cancer development to identify the main culprit genes for cancer development from one stage to another and that may lead to the drug design. The analysis revealed that hybrid Rough- Decision Tree is able to extract hidden relationships among the various genes which play an important role in causing the disease and also able to provide a unique rule set for automated medical diagnosis. Moreover at the end, the functions of the identified genes are studied and validated from Gene Ontology website DAVID which clearly shows the direct or indirect relation of genes with the cancer. This study highlights the usefulness and efficiency of RST and Decision Tree in the disease diagnosis process and its potential use in inductive learning and as a valuable aid for building more biologically significant expert systems in medical sciences.

Keywords: Rough Set Theory; Decision Tree; Microarray Data; Cancer Diagnosis; Drug Design

Introduction

The growth in the size and the number of existing pathological databases far exceeds the ability of humans to analyze this data. Such databases contain large amounts of information regarding patients and their medical conditions from the clinical, pathological and microbiological aspects. Hidden relationships or patterns within this data can provide new and vital medical knowledge, if possible to extract. This knowledge can serve as a diagnostic tool to identify the culprit gene(s) which may eventually be helpful in finding the most accurate treatment process for a particular disease. DNA Microarray [1] is an experimental procedure which indicates whether a gene is active or not, and if active, how much are their activation profile. They are represented as a dataset in public domain websites. Medical knowledge can be efficiently extracted by analyzing the data using suitable soft computing tools like Neural Network [2], Genetic Algorithms [3], Decision Trees [4-5] and fuzzy theory [6-8] etc. Medical databases usually contain certain amount of inconsistent or incomplete data. Suitable soft computing tools needs to be identified which can work in such an environment and still be able to extract the biological significant hidden interdependencies from among the various levels of representation of data. Existing intelligent data analysis techniques such as neural networks, genetic algorithms, decision trees, etc. are usually based on quite strong assumptions (some knowledge about dependencies, probability distributions, large number of experiments, etc. are required), and are unable to derive conclusions from incomplete knowledge or manage inconsistent pieces of information. Drug discovery for genetic disease in term of identifying and controlling the culprit genes form large medical database is a very crucial problem in this era.

Rough Set Theory [9-14]is an approximation tool that works well when the environment is heavy with inconsistent and ambiguous data or involves missing data The rough set approach of data analysis has many important advantages such as providing efficient algorithms for finding hidden patterns in data from database finding minimal set of data (data reduction), evaluating significance of data from the reduced data, generating set of decision rules from data and offering a straightforward interpretation of obtained results. Along the years, RST has earned a well-deserved reputation as a sound methodology for dealing with imperfect knowledge in a simple though mathematically sound way. This paper presents the basic idea of utilization of RST for the automated diagnosis of Lung Adenocarcinoma and the prediction of the responsible genes for the same. For this purpose, a microarray dataset [15], obtained from NCBI website [http://www.ncbi.nlm.nih.gov/], has been taken that contains data related to two groups of lung cells of people: those have been diagnosed with different stages of cancer like Stage I, Stage II, Stage III Stage IV and those lung tissues which are Normal and healthy. Classification Rule Sets have been generated for the dataset and the possible biological relevance of these rules generated has been predicted.

***Corresponding author:** Sudip Mandal, ECE Department, GIMT, Krishna Nagar, India, Tel: 09933320422; E-mail: sudip.mandal007@gmail.com

Received November 01, 2013; **Accepted** November 20, 2013; **Published** December 28, 2013

Citation: Mandal S, Saha G, Pal RK (2013) An Approach towards Automated Disease Diagnosis & Drug Design Using Hybrid Rough-Decision Tree from Microarray Dataset. J Comput Sci Syst Biol 6: 337-343. doi:10.4172/jcsb.1000130

Now to observe the regulatory path in which way the cancer develop from one stage to another stage, Decision Tree is used and applied independently at two consecutive stages of cancer. Several methods have been proposed for estimating gene networks from microarray data using mathematical models such as Boolean networks differential equations and Bayesian networks [16-23].Sometimes different decisions will ultimately end up at the same outcome even though a different path was chosen. This idea is the basis for using data to predict the category of a particular individual based on measured or categorical variables. These decisions can be formed into what is more formally known as a Decision Tree [24-27] which are used with a categorical response variable. By superimposing these graphs that are obtained using J48 Decision Algorithm, the departure in regulatory path at different stages and most culprit genes can be observed which are most biologically relevant.

The rest of the paper has been organized in the following manner. Section 2 discusses the automated disease diagnosis problem that has been investigated here along with the concept of RST to identify the minimal set of responsible genes. In Section 3 the application of Decision Tree on the given microarray database is discussed along with experimental results and Validations of the generated rule sets and Genes. In the concluding section, discussions have been done on this novel approach followed by list of references.

The Automated Disease Diagnostic Problem

Preliminaries

Lung Adenocarcinoma has been increasing in recent years, often beginning in the outer parts of the lungs and as such well-known symptoms of Lung Cancer such as chronic cough and coughing up blood may be less common until later stages in the disease. Early symptoms of Adenocarcinoma that may be overlooked include fatigue, mild shortness of breath, backache, shoulder ache, or chest pain.

Microarrays are used in the medical domain to produce molecular profiles of diseased and normal tissues of patients. Such profiles are useful for understanding various diseases and aid in more accurate diagnosis, prognosis, treatment planning, as well as drug discovery. DNA microarrays (gene arrays or gene chips) usually consist of thin glass or nylon substrates containing specific DNA gene samples spotted in an array by a robotic printing device. Researchers spread fluorescently labelled m-RNA from an experimental condition onto the DNA gene samples in the array. This m-RNA binds (hybridizes) strongly with some DNA gene samples and weakly with others, depending on the inherent double helical characteristics. A laser scans the array and sensors detect the fluorescence levels (using red and green dyes), indicating the strength with which the sample expresses each gene. The logarithmic ratio between the two intensities of each dye is used as the gene expression data. The relative abundance of the spotted DNA sequences in a pair of DNA or RNA samples is assessed by evaluating the differential hybridization of the two samples to the sequences on the array. Gene expression levels can be determined for samples taken: 1) at multiple time instants of a biological process (different phases of cell division) or 2) under various conditions (tumor samples with different histopathological diagnosis). Each sample corresponds to a high-dimensional row vector of its gene expression profile.

The information obtained from the above procedure is archived as microarray datasets and usually publicly maintained for further research on them. One such microarray data has been obtained by us from the NCBI site with the GEO series accession number GSE10072 which has 22284 genes, 107 no of samples with 5 type of samples characteristic (Normal, Stage I, Stage II, Stage III Stage IV). This dataset, like all other medical datasets, contains vital information hidden in the huge amount of seemingly unrelated and ambiguous data, which if possible to extract, may be significant in the automated diagnosis of diseases and may eventually lead to the discovery of new forms of treatment of complex diseases or altogether lead to their prevention if diagnosis is made possible at a very early stage. Table 1 shows a truncated Micro Array dataset.

Rough set theory

In RST, Rough set is defined in the following way: Let X ⊆ U be a target set that we wish to represent using an attribute subset *P*, i.e., an arbitrary set of objects *X* comprises a single class, and we wish to express this class (i.e., this subset) using the equivalence classes induced by attribute subset *P*. In general, *X* cannot be expressed exactly, because the set may include and exclude objects which are indistinguishable on the basis of attributes *P*. The target set *X* can be approximated using only the information contained within *P* by constructing the P-lower and P-upper approximations of *X*:

$$p_X = \{X | [X]_p \subseteq X\}$$

$$p^X = \{X | [X]_p \subseteq X \neq \Phi\}$$

The P-lower approximation, or positive region, is the union of all equivalence classes in $[x]_p$ which are the subsets and contained by the target set. The P-upper approximation is the union of all equivalence classes in $[x]_p$ which have non-empty intersection with the target set. The lower approximation of a target set is a conservative approximation consisting of only those objects, which can positively be

IDENTIFIER	GENE ID	SAMPLE ID				
		GSM254625	GSM254626	GSM254629	GSM254631	GSM254637
DDR1	1007_s_at	11	10	11	11	11
RFC2	1053_at	7	7	7	7	7
HSPA6	117_at	8	8	8	8	8
PAX8	121_at	9	10	10	10	10
GUCA1A	1255_g_at	5	5	5	5	5
UBA7	1294_at	9	9	9	9	9
THRA	1316_at	6	6	6	6	6
PTPN21	1320_at	6	6	6	6	6
CCL5	1405_i_at	9	9	8	8	8
CYP2E1	1431_at	5	5	5	5	5
SAMPLE CHARCTERISTIC		Stage: II	Normal	Stage: I	Stage: III	Stage: IV

Table 1: Truncated list of Genes and their values (Micro-Array Data).

identified as members of the set. The upper approximation is a liberal approximation, which includes all objects that might be members of target set. The accuracy of the rough-set representation of the set *X* can be given by the following:

$$\alpha_P(X) = \frac{|P_X|}{|P^X|}$$

In order to utilize the concept of rough set in predicting the dominant genes with their expression values, responsible for Lung Adenocarcinoma, the concepts of information table, decision table, reducts and core, and rule extraction have been used. An information table consists of the different variables called attributes and cases called objects. Attributes are presented in columns and the objects in rows. The attributes contained in the information table are the gene expression values (Pawlak).

Assessing the dependence of Lung Adenocarcinoma on particular genes based on this dataset, is computationally very difficult because of the size of the information table. For this reason reduction of the number of attributes to a manageable order has been done using Rough Set Theory. Then extraction of hidden relationships among this reduced data is done. In an information system there often exist some condition attributes that do not provide any additional information about the objects in *U*. So, we should remove those attributes since the complexity and cost of decision process can be reduced if those condition attributes are eliminated. Given a classification task mapping a set of variables *C* to a set of labeling *D*, a reduct is defined as any $R \subseteq C$, such that γ(C,D) = γ(R,D). The set of attributes which are common to all reduct is called core. The core is the set of attributes which is possessed by every legitimate reduct, and therefore consists of attributes which cannot be removed from the information system without causing collapse of the equivalence-class structure. It is possible for the core to be empty, which means that there is no indispensable attribute. Thus 'reducts' are formed and the extracted rules help us in assessing the dominant genes responsible for Lung Cancer. Let *S = (U, C, D)* be a reduced decision table where *C* denotes the reduced no. of attributes i.e. reduct. Every $x \in U$ determines a sequence $c_1(x) \ldots c_n(x)$; $d_1(x) \ldots d_m(x)$, where $\{c_1, \ldots, c_n\} = C$ and $\{d_1, \ldots, d_m\} = D$. The sequence will be called a decision rule induced by *x (in S)* and will be denoted by $c_1(x) \ldots c_n(x) \rightarrow d_1(x) \ldots d_m(x)$, or in short $C \rightarrow x\ D$. The number $supp_x(C, D)$ will be called a support of the decision rule $C \rightarrow x\ D$ and the number is given by $supp_x(C, D) = |A(x)| = |C(x) \cap D(x)|$.

For the ease of understanding, we have presented only 3 objects with 4 attributes in Table 2 as a sample of information table.

Reduct and core calculation

Here, the genes are considered as attributes and about 90 patient's data are considered as objects. These are used to generate the required decision table. Based on different conditions or different expression values of genes the status of the sample may be either normal or cancerous. This decision table is used as a training dataset which is used to calculate hidden dependency amongst different genes which are responsible for Adenocarcinoma. Here, number of attributes in the dataset are 22284 and the no of objects are 90. The values of the different attributes are real but for RST analysis it is considered as integers.

To find out the minimal subset of attributes or reduct & core from a huge no of attributes, normal heuristic search algorithm is very time consuming and memory complexity is very high. Therefore Genetic Algorithm (GA) is used as alternative search algorithm which is quite effective for rapid search of large, nonlinear spaces. Unlike classical feature selection strategies where one solution is optimized, a population of solutions can be modified at the same time. This can result in several optimal feature subsets as output.

A feature subset is typically represented by a binary string with length equal to the number of feature presents in the dataset. A zero or one in the j^{th} position in the chromosome denotes the absence or presence of the j^{th} feature in this subset. An initial population of chromosomes is created; the size of population and how they are connected are important issues. From this pool of feature subset, the typical genetic operators (crossover and mutation) are applied. Again, the choice of which types of crossover and mutation used must be carefully considered, as well as their probabilities of application. This generates a new feature subset pool but a suitable stopping criterion must be chosen. This is typically achieved by limiting the number of generations that take place or by setting some suitable threshold which must be exceeded by the fitness function. If the stopping criterion is not satisfied then individuals are selected from the current subset pools and the process described above repeats. As with all feature selection approaches, GA can get caught in local minima i.e. missing a truly minimal feature subset. In spite of the problem, GA is wide accepted for efficient search algorithm to calculate the reducts and core from decision table as calculation of all reducts is very exhaustive

CASES		A	ATTRIBUTES OR GENE EXPRESSION ExpressionEXExpEXPRESSION			ION	
	1007_s_at	1053_at	..	AFFX- pnX-5_at	AFFX-TrpnX-M_at		Decision
1	10	7	..	4		4	Stage: I
2	10	6	..	4		4	NORMAL
3	10	6	..	4		4	Stage: II

Table 2: Truncated Decision Table.

No	Size	Positive Region	Reducts (minimal subset)	Core
1	5	1	{203015_s_at,204351_at, 208747_s_at, 211735_x_at,212880_at }	211735_x_at
2	5	1	{200647_x_at, 204351_at, 207985_at, 211735_x_at, 212880_at}	204351_at 212880_at

Table 3: Reduct and Core of the decision table.

Stages	Responsible Genes For Cancer Stage Development Direct Regulation	Indirect Regulation
Normal- Stage:I	211735_x_at	None
Stage:I- Stage:II	211735_x_at,203015_s_at,208747_s_at,204351_at,2006 47_x_at,	207985_at 21800_at
Stage:II- Stage:III	211735_x_at ,208747_s_at,204351_at,200647_x_at, 207985_at	203015_s_at
Stage:III- Stage:IV	211735_x_at ,203015_s_at, 212880_at	None

Table 4: Regulation path obtained from Decision Tree.

Gene Name	Functions obtained from DAVID
211735_x_at	This gene involves intracellular accumulation of a structurally defective SFTPC protein. Pulmonary surfactant associated proteins promote alveolar stability by lowering the surface tension at the air-liquid interface in the peripheral air spaces. This gene directly influenced the lung disease like respiratory problem, breathless situation, bronchopulmonary and adinocarcinoma of human body.
204351_at	This protein binds two calcium ions, Co-localizes with S100PBP in the nucleus, Interacts with S100PBP and S100Z and it belongs to the S-100 family. This gene directly responsible for lung adenocarcinoma, cervical cancer, prostate cancer, pancreatic cancer.
203015_s_at	This nuclear protein interacts with cancer-related protein **SSX2.** May connect the nectin-afadin and E-cadherin-catenin system through alpha-actinin and may be involved in organization of the actin cytoskeleton at AJs through afadin and alpha- actinin. Its function directly related to cancer and leukemia. It has high genetic expressed value in the lung.
212880_at	Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Gene-disease associate with it (HuGE Navigator).
208747_s_at	There are splice variants of C1s mRNA transcripts in normal human cells for this gene and catalytic activity due to this gene. Autoimmune diseases, selective C1s deficiency, Gene-disease associate with it.
200647_x_at	Component of the eukaryotic translation initiation factor 3 (eIF-3) complexes, which is required for several steps in the initiation of protein synthesis. Disease: Unknown
207985_at	Function and disease are unknown

Table 5: Functional Classification of Responsible Genes for Lung Adenocarcinoma.

and complex in nature. In GA approaches, reducts candidates are encoded as bit strings, with the value in position i^{th} set if i^{th} attribute is present the fitness function depends on two parameters. The first is the number of bit set. The function penalizes those strings which have large numbers of bit set, driving process to find the smaller reducts. The second is the number of classifiable objects given by this candidate. The reduct should discern between as many objects as possible (ideally all of them). Although this approach is not guarantee to find the minimal subset of attributes, it may find many subsets for any given dataset. The main drawback is the large time taken to compute each string's fitness but it can handle large complexity.

Here, for the calculation of all minimal reducts, Genetic Algorithm with full indiscernibility and modulo decision technique has been used with the constraint maximum number of reduct is two. From the huge database or decision table, two reducts of 5 different attributes are generated, each of which have the positive region 1 and Stability Coefficient equal to 1. Following table shows the calculated reducts and cores for above decision table.

So RST can select few dominant genes from large number of genes by considering only two reducts which is implemented using RSES software package. These dominant genes can be considered as the responsible genes for lung Adenocarcinoma. The lists of responsible genes are as 200647_x_at, 204351_at, 207985_at, 211735_x_at, 212880_at, 203015_s_at and 208747_s_at among of which 211735_x_at, 204351_at and 212880_at are most responsible genes or indispensible genes because these genes are the core i.e. these genes are common in all reducts. RST successfully reduced the decision table and able to extract only 7 responsible genes from 22284 genes though the biological relevancies of these genes are needed to be studied yet. But observing the different combination of values of minimal number of attributes of reducts, it is very easy to predict the current status of lungs. So Rough Set can be used to design an automated disease diagnosis system from microarray data. There are five different status of Lung Adenocarcinoma and the contribution of these genes in cancerous stage development can be observed with the help of decision tree which is discussed in next section.

Application of Decision Tree in Drug Design Process

Preliminaries

Decision tree structures are a common way to organize classification schemes. In classifying tasks, decision trees visualize what steps are taken to arrive at a classification. Every decision tree begins with what is termed a root node, considered to be the "parent" of every other node. Each node in the tree evaluates an attribute in the data and determines which path it should follow. Typically, the decision test is based on comparing a value against some constant. Classification using a decision tree is performed by routing from the root node until arriving at a leaf node. Decision trees can represent diverse types of data. The simplest and most familiar is numerical data. It is often desirable to organize nominal data as well. Nominal quantities are formally described by a discrete set of symbols. The type of data organized by a tree is important for understanding how the tree works at the node level. Recalling that each node is effectively a test, numeric data is often evaluated in terms of simple mathematical inequality. Decision tree induction algorithms are functionally recursive in nature. First, an attribute must be selected as the root node. In order to create the most efficient (i.e., smallest) tree, the root node must effectively split the data. Each split attempts to pare down a set of instances (the actual data) until they all have the same classification. The best split is the one that provides what is termed the most information gain. J48 is a version of an earlier algorithm developed by J. Ross

Quinlan, the very popular C4.5. In this study J48 Algorithm is used which gives several options related to tree pruning. Many algorithms attempt to "prune", or simplify, their results. Pruning produces fewer, more easily interpreted results. More importantly, pruning can be used as a tool to correct for potential over fitting. J48 employs two pruning methods. The first is known as sub tree replacement. This means that nodes in a decision tree may be replaced with a leaf -- basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed sub tree rising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Sub tree rising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub tree rising can be somewhat computationally complex.

Error rates are used to make actual decisions about which parts of the tree to replace or rise. There are multiple ways to do this. The

simplest is to reserve a portion of the training data to test on the decision tree. The reserved portion can then be used as test data for the decision tree, helping to overcome potential over-fitting. This approach is known as reduced-error pruning. Though the method is straight-forward, it also reduces the overall amount of data available for training the model. For particularly small datasets, it may be advisable to avoid using reduced error pruning. Other error rate methods statistically analyze the training data and estimate the amount of error inherent in it. The mathematics is somewhat complex, but this approach seeks to forecast the natural variance of the data, and to account for that variance in the decision tree. This approach requires a confidence threshold, which by default is set to 25 percent. This option is important for determining how specific or general the model should be. If the training data is expected to conform fairly closely to the data you'd like to test the model on, this figure can be lowered. The C4.5 algorithm generates a classification-decision tree for the given data-set by recursive partitioning of data. The decision is grown using Depth-first strategy. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each continuous attributes.

Construction of decision tree (J48)

After reduct calculation, the whole decision table can be replaced by few genes or attributes. Now this reduced decision table can be used as the training data for constructing the Decision Tree which is applied using WEKA 3.6. To identify the druggable gene at each stage of cancer the decision is drawn for two successive stages corresponding training dataset to the consecutive stages. So there are 4 no of decision tree Normal-Stage:I, Stage:I, Stage:II, Stage: II- Stage: III and Stage: III, Stage: IV which are given below.

Nodes are related to the genes and edges are related to the gene expression value of that gene and the last nodes denote the status of lungs. Edges from one node to another denote the regulation. The edges which go from one gene to another gene can be considered as the indirect regulation path that means these genes indirectly affect the status of disease via some others genes regulation. Few edges directly came from node to the last leaf or node or status of disease. This type of path can be considered as direct regulatory path i.e. these genes directly affect the status of disease. Following table shows the different most responsible genes at each stage of cancer, which are directly or indirectly involved or regulated in the development in different cancerous stages in lungs. Following table shows the different genes with different regulation which are mainly responsible for the moving one cancer stage to another state among all others genes.

From the above Figures and Table that 211735_x_at is the most culprit gene for development in cancer as it remains in all decision trees for different stage classification. It directly influenced each state of the cancer in human body.

204351_at is also another important gene which directly influenced or regulates the status of lung either Stage: I, Stage: II, Stage: III.

The 203015_s_at gene also directly or indirectly influenced or regulate the status of lung either Stage: I, Stage: II, Stage: III, Stage: IV.

208747_s_at gene also directly influenced to determine the status of lung in Stage: I, Stage: II, Stage: III.

207985_at gene indirectly influenced to determine Stage: I, Stage: II.

200647_x_at gene also directly influenced the status of lung in Stage: I, Stage: II, Stage: III.

212880_at gene also directly influenced the status of lung in Stage: III and Stage: IV.

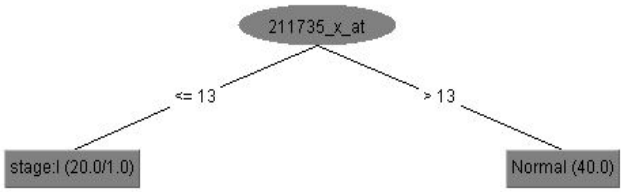

Figure 1: Decision Tree for Normal-Stage: I.

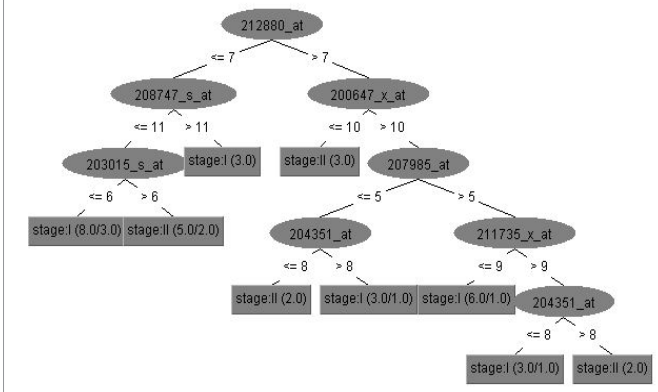

Figure 2: Decision Tree for Stage: I-Stage II.

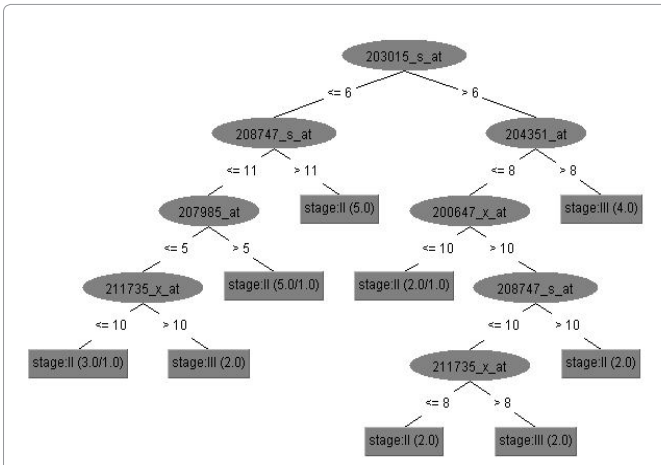

Figure 3: Decision Tree for Stage: II- Stage: III.

Validations of obtained results

Above described novel approach is only a mathematical model to determine classification and to find out small scale dominant regulation in case of Lung Adenocarcinoma. After applying RST the responsible genes are found. Now this results need to validate whether those genes are actually affect the state of disease or not in real life. The validations consist of two parts; one is mathematical validation and another real life functional classification of Genes.

The mathematical validation is based on rule generation from obtained reducts using rough set theory and classification of whole dataset on the basis of that generated rules which have only 7 responsible genes. The entire training dataset consist of five parts on the basis of decision attribute (comprising of 49 cancerous among which 19 Cases are Stage: I, 16 Cases are Stage: II, 11 Cases are Stage: III & 3 Cases are Stage: IV patients and 41 normal persons). The same dataset has been applied on the generated rule sets, to verify the accuracy of prediction in terms of diagnosis of diseases. The cross validated result, as shown in the Figure 1 indicates that these rules can accurately predict with 100% accuracy and total coverage is 1. In spite of using 22284 genes, 7 responsible genes can easily classify the entire dataset which validate our results. The confusion matrix for the cross validation which shows actual and predicted cases, is given below (Figure 5).

Though the mathematical validation give satisfactory result, still we can't say that these 7 Genes are actually affect or regulate the status of lung in human body. So to biologically relevancy of the process, next part of validation is consisting of actual functional classification of those genes in human body. This has been obtained from a Gene Ontology website called DAVID [http://david.abcc.ncifcrf.gov/] where it is possible to find the functional classification of these genes. If the lists of responsible genes are given as input with appropriate gene identifier, the website shows the function of these genes or proteins in human body. Moreover it is also possible to find the particular genetic disease which occurred due to variation in these gene expressions. These functions of genes are obtained and stored from different cancer hospitals, gene research centers around the world where scientist actually observe the function in wet lab. However, the functions of these are given below.

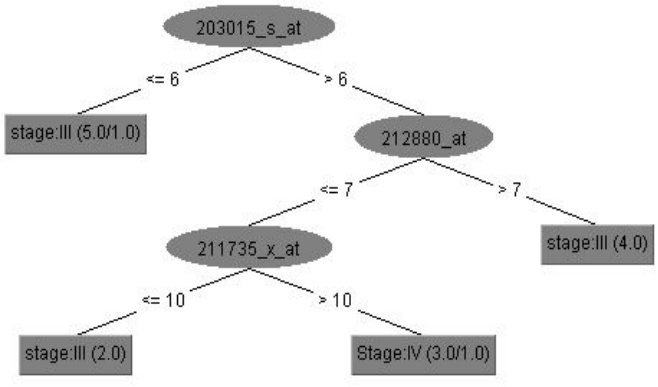

Figure 4: Decision Tree for Stage: III-Stage: IV.

Results of experiments by train&test method: "ADINOCARCINOMA_4 STAGES_TRAINING.ros"

		Predicted							
		stage:II	Normal	stage:I	stage:III	Stage:IV	No. of obj.	Accuracy	Coverage
Actual	stage:II	16	0	0	0	0	16	1	1
	Normal	0	41	0	0	0	41	1	1
	stage:I	0	0	19	0	0	19	1	1
	stage:III	0	0	0	11	0	11	1	1
	Stage:IV	0	0	0	0	3	3	1	1
	True positive rate	1	1	1	1	1			

Total number of tested objects: 90
Total accuracy: 1
Total coverage: 1

Figure 5

From above table of survey it is clearly shown that 211735_x_at, 204351_at, 203015_s_at genes are directly involved with cancer and 212880_at, 208747_s_at, 200647_x_at are responsible for different gene disease. This proves the biological relevancy of the proposed method.

So for drug design, if we can able to control the genetic value or feature of (or these protein) 211735_x_at, 204351_at, 203015_s_at, 212880_at, 208747_s_at and 200647_x_at by some process, it will directly influenced and regulate the status of Lung Adenocarcinoma such that concerned person will be instantly cured. The drug design must carry on the laboratory (Wet Lab) after identification of responsible genes to control the cancer. So the rough set and decision tree describe a novel method for inferring a biologically significant small dominant genetic regulatory network for drug design which is also validated by the DAVID website.

Discussions

This study has been carried out on human disease diagnosis process using genetic information in the form of microarray data of the disease Lung Adenocarcinoma. It has been assumed that the set of genes will more or less act in the same general way in a particular species e.g. human beings in the present investigation. With this assumption in mind, a procedure has been proposed here for developing two distinct purposes, first, diagnosis of disease using the microarray data and second, the isolation of the most important genes responsible for causing the disease. The algorithm is developed using a soft computing tool Rough Set Theory (RST). The results obtained can be used for quite dependable predictions. The validation of predicted results has also been carried out here. The result shows 100% accuracy in prediction in the form of diagnosis of disease. The result also predicted responsible or affected genes for causing Lung Adenocarcinoma.

Then using these responsible genes, decision tree is obtained at each two consecutive stages to determine the responsible genes at each stage of cancer that affect the progress of cancer. This may lead toward the drug design in future for cancer with regulating the activity of the culprit genes or proteins by injecting proper drugs which will prevent further development of cancer in human body. The main disadvantages of this process is that we only consider first two reduct to identify the responsible genes and try to construct small decision table from it. The real life validation for this can be carried out in the DAVID website.

All the validation results strengthen our proposal that by using only a mathematical tool like RST and Decision Tree, it is possible to isolate those genes which are relatively more active at different stages in causing or inhibiting Lung Adenocarcinoma and that will lead a one step ahead towards drug design.

References

1. Masys DR (2001) Linking microarray data to the literature. Nat Genet 28: 9-10.
2. Wang Z, Palade V, Xu Y (2006) Neuro-Fuzzy Ensemble Approach for Microarray Cancer Gene Expression Data Analysis. International Symposium on Evolving Fuzzy Systems: 241-246.

3. Huerta EB, Duval B, Hao J (2006) A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data. EvoWorkshops 3907: 34-44.

4. David JM, Balakrishnan K (2010) Machine Learning Approach for Prediction of Learning Disabilities in School-Age Children. Int J Comput Appl 9: 7-12.

5. David JM, Balakrishnan K (2011) Prediction of Key Symptoms of Learning Disabilities in School-Age Children using Rough Sets. International Journal of Computer and Electrical Engineering 3: 163-168.

6. Bezdek JC (1993) Editorial: Fuzzy Models- What are they and Why. IEEE Transactions on Fuzzy Systems 1: 1-12.

7. Vinterbo SA, Kim EY, Ohno-Machado L (2005) Small, fuzzy and interpretable gene expression based classifiers. Bioinformatics 21: 1964-1970.

8. Ho SY, Hsieh CH, Chen HM, Huang HL (2006) Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis. Biosystems 85: 165-176.

9. Walczak B, Massart DL (1999) Rough sets theory. Chemometrics and Intelligent Laboratory Systems 47: 1-16.

10. Tsumoto S (2001) Medical diagnostic rules as upper approximation of rough sets. Fuzzy Systems, 2001. The 10th IEEE International Conference 1551-1554.

11. Pawlak Z (2002) Rough set theory and its applications. Journal of Telecommunications and Information Technology 7-10.

12. Midelfart H, Komorowsk J, Norsett K, Yadetie F, Sandvik AK, et al. (2002) Learning Rough Set Classifiers from Gene Expression and Clinical Data. Fundamenta Informaticae 53: 155-183

13. Banerjee M, Mitra S, Banka H (2007) Evolutionary rough feature selection in gene expression data. Systems, Man, and Cybernetics, Part C: Applications and reviews 37: 622-632.

14. Hassanien AE, Ali JMH (2004) Rough set approach for generation of classification rules of breast cancer data. Informatica 15: 23-38.

15. Midelfart H, Laegreid A, Komorowski J (2001) Classification of gene expression data in an ontology. Medical Data Analysis 2199: 186-194.

16. Akutsu T, Miyano S, Kuhara S (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput .

17. Akutsu T, Miyano S, Kuhara S (2000) Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics 16: 727-734.

18. Chen T, He HL, Church GM (1999) Modeling gene expression with differential equations. Pac Symp Biocomput .

19. de Hoon M, Imoto S, Miyano S (2002) Inferring gene regulatory networks from time-ordered gene expression data using differential equations. Discovery Science 2534: 267-274.

20. Friedman N, Goldszmidt M (1998) Learning Bayesian networks with local structure. 252-262.

21. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. J Comput Biol 7: 601-620.

22. Grzegorczyk M (2010) An introduction to Gaussian Bayesian networks. Methods Mol Biol 662: 121-147.

23. Mandal S, Saha G, Pal RK (2013) Reconstruction of Dominant Gene Regulatory Network from Microarray DataUsing Rough Set and Bayesian Approach. J Comput Sci Syst Biol 6: 262-270.

24. Mandal S, Saha G (2013) Rough Set Theory based Automated Disease Diagnosis using Lung Adenocarcinoma as a Test Case. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 1: 59-66.

25. Podgorelec V, Kokol P, Stiglic B, Rozman I (2002) Decision trees: an overview and their use in medicine. J Med Syst 26: 445-463.

26. Rajput A, Aharwla PR et al (2011) J48 and JRIP Rules for E-Governance Data. International Journal of Computer Science and Security 5: 201-208.

27. Lavanya D, Rani KU (2011) Performance Evaluation of Decision Tree Classifiers on Medical Datasets. International Journal of Computer Applications (0975 – 8887) 26: 1-4.

28. Khan MU, Choi JP, Shin H, Kim M (2008) Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. Conf Proc IEEE Eng Med Biol Soc 2008: 5148-5151.