Prediction of Dominant Genes Responsible for Lung Adenocarcinoma using Rough Set Theory

Abhinandan Khan, Jadavpur University, India
Dr. Goutam Saha, GCELT, India
Srirupa Dasgupta, GCELT, India
Soumya Kanti Datta, EURECOM, France
khan.abhinandan@gmail.com, dr_goutamsaha@yahoo.com, srirupadasgupta@rediffmail.com, dattas@eurecom.fr

***Abstract*-This paper presents an efficient approach of predicting the dominant genes responsible for Lung Adenocarcinoma using Rough Set Theory. The work takes a microarray dataset containing data of diseased, suspected and healthy patients and characterizes them in terms of objects and attributes. Using rough set theory, redundant attributes are then determined and eliminated. The core attributes are worked out by analyzing the relationship among the remaining attributes. Then Johnson's reduction algorithm has been used to extract underlying important rules from the remaining dataset. The paper reports three sets of rules, one each for diseased, suspected and healthy persons. The dominant genes can be accurately predicted by investigating the genes appearing in the generated Rule Sets. Microarray data obtained from a patient is analyzed in accordance with the Rule Sets generated. If any match is found with any one of the mentioned three cases, the patient will be diagnosed accordingly.**

Keywords: Lung Adenocarcinoma, Indiscernibility Relation, Reduct, Core, Microarray Dataset.

I. INTRODUCTION

Lung Adenocarcinoma is a malignant proliferation of lung epithelial cells involving glandular differentiation or mucin production by the tumor cells. It is the most common type of lung cancer in women & non-smokers. Lung adenocarcinoma often begins in the outer parts of the lungs and develops slowly but metastasis takes place early & widely. Causes of adenocarcinoma include radon gas exposure, air pollution, smoking (incidence is about 75%). Heredity is surprisingly another causal factor in triggering lung tumors. Thus it is very important to investigate the genetic basis of the disease and establish relation between the genetic changes and their outcomes predisposing to this disease which will help in further prevention and screening of the disease.

This paper highlights a mathematical approach using Rough Set Theory for the automated prediction of dominant genes responsible for Lung Adenocarcinoma using the microarray dataset. Rough set works well when the environment is heavy with inconsistent and ambiguous data or involves missing data [2]. The rough set approach of data analysis efficiently investigates hidden patterns in data. The process also allows clear interpretation of obtained results [1].

Initially microarray datasets have been collected containing data related to diagnosed, suspected and healthy persons. Decision Rule Sets have been generated from this dataset using rough set. Based on these rule sets, the dominant genes can be identified. At a later stage of the research, a patient's microarray data can be analyzed in a similar fashion to generate decision rule sets and investigating the result with the mentioned rule sets of the above three groups, the condition of the person can be validated accordingly.

The rest of the paper is organized as follows. Section II describes the concept of Rough Set Theory. Section III explains the proposed approach and determination of 'reduct' and 'core' in details. Sections IV and V present the methodology behind the decision table and the experimental results respectively. Then the paper concludes with the inferences drawn in section VI.

II. ROUGH SET CONCEPT

The rough set [4] concept is based on indiscernibility relations. A set of all indiscernible or similar objects form an elementary set. Rough set is defined in the following way. Let $X \subseteq U$ be a target set that is represented 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 [9].

$$P_X = \{x | [x]_p \subseteq X\} \quad (1)$$

$$P^X = \{x \mid [x]_p \subseteq X \neq \Phi\} \quad (2)$$

The accuracy of the rough-set representation of the set X can be given by the following:

$$\alpha_P(X) = \frac{|P_X|}{|P^X|} \quad (3)$$

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 identified as members of the set. The upper approximation is a liberal approximation, which includes all objects that might be members of target set [5].

ISBN: 978-606-544-078-4

III. PROPOSED APPROACH

The rough set based study enables us to reduce superfluous data, generate rules of categorization showing hidden relationships between the description of objects and their assignment to classes [3].

Application of rough set in determination of dominant genes

In this work, concepts of information table, decision table, reducts & core and rule extraction has been used to identify the dominant genes (responsible for Lung Adenocarcinoma) with their expression values. Each information table consists of attributes and objects. In this work, the attributes represent the gene expression values obtained from microarray dataset. This paper deals with a huge collection of 22215 numbers of genes for diagnosis of Lung Adenocarcinoma.

This paper investigates a collection of 192 cases from the GEO Dataset maintained at the mentioned website. 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. Thus 'reducts' are formed. The extracted rules help us in assessing the dominant genes responsible for Lung Cancer in smokers.

Determination of Reducts and Core

There exists a subset of attributes, which can, by itself, fully characterize the knowledge (information) in the database. Such an attribute set is called a reduct. The reduct of an information system is not unique. There may be many such subsets of attributes, which preserve the equivalence-class structure expressed in the information system [6].

Formally, a reduct (RED) is a subset of attributes RED $\subseteq$ P such that [9]

- $[x]_{RED} = [x]_P$, that is, the equivalence classes induced by the reduced attribute set RED are the same as the equivalence class structure induced by the full attribute set P.
- The attribute set RED is minimal, in the sense that $[x]_{(RED - \{a\})} \neq [x]_P$ for any attribute a $\in$ RED ; in other words, no attribute can be removed from set RED without changing the equivalence classes $[x]_P$.

A reduct can be thought of as a sufficient set of features so as to represent the category structure. In the Table I, attribute set {200598_s_at,201884_at} is a reduct i.e. the information system projected on just these attributes possesses the same equivalence class structure as that expressed by the full attribute set {1},{2},{3},{4},{5}, {6}.

The set of attributes common to all reducts is called the 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. The core may be thought of as the set of necessary attributes – necessary, that is, for the category structure to be represented.

It is possible that there is no indispensable attribute and core is empty. Any single attribute in such an information system can be deleted without altering the equivalence-class structure. In such cases, there is no essential or necessary attribute which is required for the class structure to be represented.

Table I
Sample Information Table

Objects	Attributes					
	200007_at	200598_s_at	200795_at	201041_s_at	201884_at	202917_s_at
1	11	6	9	9	11	6
2	11	8	9	8	11	5
3	11	6	9	8	10	7
4	11	7	9	8	10	7
5	9	5	7	6	10	8

In this paper, the core attributes of Lung Adenocarcinoma has been determined. Different reducts may be created but, the reduct with minimum size has to be determined. From the sample information table shown as Table I, we take various attributes to compare with the set of all attributes, viz. A*= {1}, {2}, {3}, {4} and {5}, as shown below:

- {200598_s_at , 201884_at}* = {1}, {2}, {3}, {4}, {5} comparing with A*, {200598_s_at, 201884_at}*=A*. Thus {200598_s_at, 201884_at}* is a reduct.
- {200598_s_at, 200795_at}* = {1, 3}, {2}, {4}, {5}; comparing with A*, {200598_s_at, 201884_at}* ≠ A*. Thus {200598_s_at, 201884_at}* is not a reduct.
- {200598_s_at , 200795_at , 201041_s_at, 201884_at}* = {1}, {2}, {3}, {4}, {5}; comparing with A*, {200598_s_at, 200795_at, 201041_s_at, 201884_at}* = A*. Thus {200598_s_at, 200795_at, 201041_s_at, 201884_at}* is a reduct.

This method of obtaining all the reducts is very time consuming. So the task is to find out just a single reduct using a heuristic algorithm LEM1. The first step of this algorithm is to eliminate the leftmost attribute from the set and check whether the remaining set is a reduct or not. If the set is not a reduct, the attribute is put back into that set and the next attribute is eliminated for similar checking. Likewise, this elimination procedure is continued until the last attribute is attained.

Here the left most attribute, 200007_at is first eliminated and then whether the remaining combined set forms a reduct or not is checked. As {200598_s_at, 200795_at, 201041_s_at, 201884_at, 202917_s_at}* = A* is obtained, {200598_s_at, 200795_at, 201041_s_at, 201884_at, 202917_s_at} is a reduct. Eliminating the next left most attribute, 200598_s_at, it is checked whether the remaining set forms a reduct or not.

{200795_at, 201041_s_at, 201884_at, 202917_s_at}* = A* is obtained in this case therefore {200795_at, 201041_s_at, 201884_at, 202917_s_at} is also a reduct. Similarly, after eliminating 200795_at, {201041_s_at, 201884_at, 202917_s_at}*≠ A* is obtained which shows that {201041_s_at, 201884_at, 202917_s_at}* is not a reduct. Proceeding in this manner it is found that eliminating 201041_s_at, 201884_at and 20291_s_at one by one do not yield reducts. So it can be safely concluded that the LEM1 algorithm forms the set of attributes {200795_at, 201041_s_at, 201884_at, 202917_s_at} as the core reduct [7]. The determination of reducts from huge datasets using LEM1 algorithm is time consuming and very complex. Hence, we use the Johnson's Reduction Algorithm [10]. This algorithm is very fast as it uses greedy search to find one reduct.

IV. Decision Table

One of the important aspects in the analysis of decision tables is the extraction and elimination of redundant attributes. The identification of the most important attributes from the data set is also an equally important aspect. Redundant attributes are attributes that could be eliminated without affecting the degree of dependency between the remaining attributes and decision or the equivalence class structure. The degree of dependence is a measure that conveys the ability of the attributes to discern objects from each other. In a decision table, variables are presented in columns in two categories- control attributes and decision attributes. Decision table has only 3 possible outcomes: Smokers diagnosed with cancer (LA = 1), Smokers not diagnosed with cancer (LA = 2) And smokers suspected of cancer (LA = 3). Rows of the decision table, like a simple information table, are filled with the cases.

A list of the gene expression values i.e. attributes, is used for identifying the dominant genes responsible for Lung Cancer. Table 3 i.e. the Sample Decision Table, portrays three elementary sets - {LA}:{1,2} for smokers diagnosed with cancer, {LA}:{3,4} for smokers not diagnosed with cancer, and {LA}:{5}for smokers suspected of cancer. Elementary sets of decisions are known as concepts. Decision tables are crucial for rule extraction. Based on the concept of Rough Sets, we determine the relationship between the decisions attributes and the control attributes.

A decision table may contain more than one reduct and any reduct can be used to replace the original table without affecting the equivalence structure. We can define the number of reducts from the decision table. Selecting the best reduct amongst all reducts in a decision table, is important. Here, we have adopted an eligibility criterion for the reducts: the best are those with minimum number of attributes. We are getting such type of reduct for our required predictions using the Johnson's Algorithm. Hence, based on the sample decision table, we are striving to find a solution such that, a single attribute or a single dominant gene is responsible for the outbreak of Lung Adenocarcinoma in a smoker [8].

TABLE II
SAMPLE DECISION TABLE

Objects	Attributes						Decision
	200007_at	201884_at	207913_at	217022_s_at	217478_s_at	36711_at	
1	10	10	8	8	8	7	1
2	11	10	8	7	8	7	1
3	9	11	6	7	9	6	3
4	11	12	8	8	9	8	3
5	10	8	7	7	8	6	2

V. Results

A part of the extracted rules from the above mentioned microarray data of diseased, suspected and healthy persons have been shown in Table III. Based on Table III, another table, Table IV, has been formed describing the activity of each gene with respect to the activity level in different cases as found in the rules generated. Thus the dominant gene detection process can be considered as a decision making process and the rules generated by considering the original data set give a strong platform for making decisions. At a later phase these rules can be used for diagnosing a patient accurately and finding responsible gene for the disease.

VI. Discussion

By investigating Table III and Table IV, we arrive at the following conclusions. It can be safely inferred that Table III can serve as a look up table which can be used for categorizing a diseased, suspected and unaffected person. After analyzing some patient's microarray data in a way, similar to that used in this paper, if analogous rule set is found, the patient can then be diagnosed accordingly.

It is found that the core reduct contains 23 genes of which only 20 take part in the generation of Rule Sets. Thus it can be safely inferred that these 20 genes play an important part in the decision making process i.e. diagnosis of a patient. It can be seen from Table IV that two genes denoted by their ID_REF, '214414_x_at' and '217478_s_at' respectively do not play any part in the rules generated (as shown in Table III) for the case of unaffected or suspected patients. Thus we can predict with some accuracy that these two genes play a significant role in producing Lung Cancer in smokers. Out of these two genes, the gene marked with ID_REF, '214414_x_at', plays a more significant role in producing cancer. This is due to the fact that this gene appears in a rule singularly in case of patients diagnosed with cancer. This gene is not found in any other rules developed. Thus it can be safely predicted that this is a cancer causing gene.

It is seen from Table IV that the gene identified with ID_REF, '36711_at' is found in the Rule Sets generated for smokers not diagnosed with cancer. Thus it can be safely predicted that this gene may be cancer preventing.

TABLE III
EXTRACTED RULE SETS FROM THE DECISION TABLE

Rule	LHS Support	RHS Support	LHS Coverage	RHS Coverage
200007_at(10) AND 207808_s_at(4) => Decision(1)	10	10	0.052083	0.111111
200007_at(10) AND 201041_s_at(8) => Decision(1)	15	15	0.078125	0.166667
201884_at(12) AND 36711_at(6) => Decision(1)	2	2	0.010417	0.022222
200007_at(9) AND 201041_s_at(8) AND 207808_s_at(6) => Decision(1)	1	1	0.005208	0.011111
200598_s_at(6) AND 212671_s_at(6) => Decision(2)	1	1	0.005208	0.2
214414_x_at(6) => Decision(3)	1	1	0.005208	0.010309
207808_s_at(9) AND 217478_s_at(9) => Decision(3)	4	4	0.020833	0.041237

Another gene identified with ID_REF, '212671_s_at' does not take part in generating any Rule Sets for either the healthy smokers or those with Lung Cancer as shown in Table IV. However it appears in the Rule Sets generated for the smokers suspected of Lung Cancer. So no firm conclusion can be drawn about the activity of this gene which may be investigated in wet lab.

Another important observation that can be made from Table IV is that six genes identified with ID_REF, '200007_at', '200598_s_at', '200795_at', '201041_s_at', '202917_s_at' and '207808_s_at' appear in the Rule Sets generated for all the three cases. However their expression levels differ in the three cases. It means that the regulation of these genes play an important role in producing Lung Cancer in smokers.

TABLE IV
EXPRESSION VALUES OF GENES APPEARING IN THE EXTRACTED RULE SETS

Gene	Expression Values In The Rule Sets Obtained		
	Healthy	Suspected	Diagnosed
200007_at	8,9,10,11	8,9	7,8,9,10,11
200598_s_at	5,6,7,8	6	4,5,6,7,8
200795_at	7,8,9,10,11	8	5(S),6,7,8,9,10
201041_s_at	5,7,8,10,11	8,9	5,6,7,8,10,11,12(S)
202917_s_at	5,6,7,8,9,11,12,13	5	7,8,9,10,11,12
207808_s_at	4,6,9	4	6,9
212671_s_at	NA	6	NA
214414_x_at	NA	NA	6(S)
217478_s_at	NA	NA	9
36711_at	6	NA	NA

These are the predictions which can be clearly drawn using mathematical tool, Rough Set Theory, which has got some biological significance.

Further experiments can also be carried out on these genes appearing in the Rule Sets generated (Table III) in the wet lab for further biological investigation with respect to the relevance of these genes in relation to diagnosis of Lung Cancer. Investigation can also be carried out in the light of metabolic pathway engineering. This gives us future scope of research. The contribution of the paper lies in the proposed approach for diagnosis of the disease and genes responsible. This work can further pave way for detection of cancer and other diseases and consequently impact the way diseases are diagnosed.

REFERENCES

[1] H. Midelfart, J. Komorowski, K. Nørsett, F. Yadetie, A.K. Sandvik and A. Lægreid, "Learning Rough Set Classifiers from Gene Expression and Clinical Data," In *Fundamenta Informaticae* 53(2), 155-183, 2002.

[2] H. Midelfart, A. Lægreid and J. Komorowski, "Classification of Gene Expression Data in Ontology," in *Second International Symposium on Medical Data Analysis,* LNCS 2199, pages 186-194, 2001.

[3] A. Lægreid, T.R. Hvidsten, H. Midelfart, J. Komorowski and A.K. Sandvik, "Predicting Gene Ontology Biological Process from Temporal Gene Expression Patterns," in *Genome Res.* 2003 May 13(5):965-79.

[4] Z. Pawlak, Rough Sets. University of Information Technology and Management ul. Newelska 6, 01-447 Warsaw, Poland.

[5] A. Øhrn, and T. Rowland, "Rough Sets: A Knowledge Discovery Technique for Multifactorial Medical Outcomes." In *American Journal of Physical Medicine and Rehabilitation,* vol 79, no. 1, pp. 100-108, 2000.

[6] P. Maji and S.K. Pal, "Fuzzy–Rough Sets for Information Measures and Selection of Relevant Genes from Microarray Data," In *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics,* Vol. 40, No. 3, June 2010.

[7] P. Maji and S.K. Pal, "Measures Approximation Spaces in IEEE Feature Selection Using f-Information," In *Fuzzy Transactions on Knowledge and Data Engineering,* Vol. 22, No. 6, June 2010.

[8] P. Mitra, S. Mitra and S.K. Pal, "Staging of Cervical Cancer with Soft Computing," In *IEEE Transactions on Biomedical Engineering,* Vol. 47, No. 7, July 2000.

[9] http://en.wikipedia.org/wiki/Rough_set.

[10] F. Li, T. Guan, X. Zhang, and X. Zhu, "An Aggressive Feature Selection Method based on Rough Set Theory," In *Second International conference on Innovative Computing, Information and Control,* September 2007.