Prediction of Prominent Genes for Insulin Effect using Rough Set Theory

Srirupa Dasgupta
Department of IT
GCELT, Kolkata, India
srirupadasgupta@rediffmail.com

Ritwik Mondal
Department of IT
GCELT, Kolkata, India
ritwik.mondal@gmail.com

Goutam Saha
Department of IT
North Eastern Hill University, Shillong, India
dr_goutamsaha@yahoo.com

Rajat Kumar Pal
Department of CSE
University of Calcutta, Kolkata, India
pal.rajatk@gmail.com

***Abstract*— This paper presents an efficient approach for determining the dominant genes responsible for predicting the insulin effect on human muscles using Rough Set theory. The work takes a microarray dataset containing data of insulin sensitive, insulin resistant, and diabetic patients, and characterizes them in terms of objects and attributes. Using Rough Set theory, redundant attributes are eliminated and reducts are generated. From the reducts, rules are generated. These rules are further verified with test sets and ontology. The paper reports three sets of rules, one each for insulin-sensitive, insulin-resistant, and diabetic 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 is diagnosed accordingly.**

Keywords*— *Insulin-effect; Insulin-resistive; Insulin-sensitive; Indiscernibility Relation; Reduct; Core; Microarray data set.

I. INTRODUCTION

Insulin-effect is a condition, which notes how the body responds to amounts of insulin. Broadly there are three types of effect of insulin on human body. Insulin-sensitive is a term, which describes those people who require relatively low and normal levels of insulin to process blood sugar. Insulin-resistant is a term for people who need a lot of insulin to process glucose. Finally, diabetes occurs when the body fails to keep blood glucose under control even after producing high levels of insulin. Insulin-resistance is the early stage that later on turns into diabetes if no corrective action is taken. An array of other problems associated with this disease emerges out in the body functioning due to this condition. There are many negative health effects that come forward before the full-blown diabetes sets in.

The standard diet of most of the world population is composed of a large percentage of simple carbohydrates such as white bread, breakfast rolls, candy bars, pasta, potatoes, and carbonated soft drinks. Anyone can become insulin resistant. In fact, most of us are likely to have some level of insulin resistance. It is just a matter of degree. The more processed and refined food that we eat, the more insulin is required to metabolize it. The more insulin injected in our blood, our cells become the less responsive. With aging, this continual exposure to high levels of insulin wears out our tolerance for refined carbohydrates and reduces our sensitivity to insulin.

This paper highlights a mathematical approach using Rough Set theory for the automated prediction of dominant genes that are responsible for detecting the insulin-effect on human muscles using the microarray dataset for the purpose.

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]. That is why Rough Set theory has been adopted for analysis in the present investigation.

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 theory. 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 investigate the results with the mentioned rule sets for confirming the patient's clinical condition within the above mentioned three groups of people. The accuracy of predicted result about the condition of the person can serve as validation of proposed algorithm. The results are further verified from the Gene Ontology site DAVID.

The rest of the paper is organized as follows. Section II describes the concept of Rough Set theory. Section III explains the proposed approach for incorporating Rough Set theory for this investigation purpose and determination of 'reduct' and 'core' in details. Section IV presents the results of the experiments and methodology behind developing the decision table. Section V is based on discussion, inferences, and verification of the results. In Section VI, we conclude the paper with a few remarks.

II. ROUGH SET CONCEPT

The Rough Set [4] concept is based on indiscernible relations. A set of all indiscernible or similar objects form an elementary set. Rough set is defined in the following way. Let $X \in U$ 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 of P. The target set X can be approximated using only the

ISBN: 978-81-923777-9-7

information contained within P by constructing the P-lower and P-upper approximations of X [9],[7].

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

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

The accuracy of the Rough Set representation of 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],[11].

III. Proposed Approach

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

A. Application of rough set in determination of dominant genes

In this work, concepts of information table, decision table, reducts and core and rule extraction have been used to identify the dominant genes (responsible for insulin effect) with their expression values. Each information table consists of attributes and objects. The attributes represent the gene expression values obtained from microarray dataset. This paper deals with a huge collection of 12626 features or numbers of genes for diagnosis of insulin effect on human muscles.

This paper investigates a collection of 110 cases from the GEO data sets obtained from the NCBI (National Center for Biotechnology Information) website.

Assessing the dependence of insulin-effect 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 insulin-effect on humans.

TABLE I. SAMPLE INFORMATION TABLE

Objects	Attributes					
	1000_at	1001_at	1002_at	1003_at	1004_at	1011_at
1	10	8	9	9	11	9
2	10	8	9	8	11	8
3	10	8	9	8	10	10
4	9	8	9	8	10	10
5	10	8	7	6	10	11

B. 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 P such that [9]

- $[x]_{RED} \subseteq [x]_P$, that is, the equivalence classes induced by the reduced attribute set $[x]_{RED}$ 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 Table I, attribute set {1000_at, 1011_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, i.e., $[x_P]=\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\}$ as resolved by {1000_at, 1001_at, 1002_at, 1003_at, 1004_at, 1011_at} .

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. 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.

ISBN: 978-81-923777-9-7

In this paper, the significant attributes of insulin effect has been determined using Rough Set Exploration System (RSES). The determination of reducts from huge data sets is time consuming and very complex. The RSES uses two algorithms to calculate the reducts.

C. Algorithmic Approach to Find Reduct

• Exhaustive algorithm: This algorithm realizes the need of object-oriented reducts. It has been shown that any minimal consistent decision rule set for a given decision table |S| can be obtained from objects by reduction of redundant descriptors. This method is based on Boolean reasoning approach. It is very time consuming and also requires a huge amount of memory.

• Genetic algorithm: This algorithm[10], is comparatively faster. Using permutation encoding and special crossover operator one can calculate a predefined number of minimal consistent rules. The later has thus been used to calculate the reducts for the data set of 110 persons with 12626 attributes or genes.

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 dataset 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.

TABLE II. SAMPLE DECISION TABLE

Objects	Attributes						Decision
	1000_at	1001_at	1004_at	1011_s_at	1015_s_at	1016_s_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

In a decision table, variables are presented in columns in two categories— control attributes and decision attributes. Decision table has only three possible outcomes: people diagnosed as insulin-sensitive (Decision value 1), people diagnosed as insulin-resistive (Decision value 2), and people diagnosed as diabetic (Decision value 3). Rows of the decision table, like a simple information table, are filled with the cases. Table II, i.e., the Sample Decision Table, portrays three elementary sets— rows:{1,2} for people diagnosed as insulin-sensitive, row:{5} for people diagnosed as insulin-resistive, and rows:{3,4} for people diagnosed as diabetic. 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. We select those reducts from the reduct set, which match with a maximum number of objects in the data set.

IV. EXPERIMENTAL RESULTS

The snapshot of the reduct set that have been extracted from the above mentioned microarray data of insulin-sensitive, insulin-resistive, and diabetic persons have been shown in Table III. Based on it Table IV, Table V have been formed describing the rules thus generated.

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. We are interested in applying these rules for making a very important decision— diagnosing a patient accurately and finding responsible gene for the disease.

The following Table III shows a glimpse of the reducts from the reduct set.

TABLE III. REDUCT SET

Reduct set: insulin_effect85

(1-96)	Size	Pos.Reg.	SC	Reducts
1	8	1	1	{1080_s_at,1916_s_at,33760_at,34544_at,37324_at,37481_at,38074_at,39655_at}
2	10	1	1	{1019_g_at,1594_at,31567_at,32526_at,33591_at,34532_at,34869_at,35527_at,41066_at,612_s_at}
3	11	1	1	{1010_at,1101_at,31488_s_at,32848_at,34375_at,35066_g_at,36542_at,36812_at,38394_at,40210_at,41124_r_at}
4	13	1	1	{1019_g_at,1494_f_at,1803_at,258_at,31983_at,34320_at,35035_at,36231_at,37332_r_at,38631_at,39174_at,39475_at,40650_r_at}
5	14	1	1	{1012_at,2091_at,32936_at,33933_at,34001_at,35504_at,35830_at,372_f_at,37798_at,37992_s_at,38175_at,38700_at,39125_at,695_at}
6	12	1	1	{1055_g_at,303_at,31590_g_at,33276_at,33463_at,35747_at,38348_at,39581_at,39973_at,40322_at,40776_at,844_at}
7	12	1	1	{1052_s_at,1273_r_at,1295_at,35340_at,35365_at,37483_at,38086_f_at,38547_at,39634_at,39932_at,40077_at,40195_at}
8	14	1	1	{1029_s_at,1083_s_at,303_at,31925_s_at,33169_at,34763_at,37785_at,38625_g_at,38790_at,39248_at,39895_r_at,39939_at,472_at,510_g_at}
9	12	1	1	{1059_at,1187_at,1524_at,33001_s_at,35514_at,36237_at,36463_at,36540_at,39731_at,40754_at,41124_r_at,41343_at}
10	13	1	1	{1069_at,1087_at,31564_at,31652_at,32567_at,32917_at,32988_at,33500_i_at,33841_at,34023_at,37521_s_at,38982_at,39907_at}
11	10	1	1	{1288_s_at,35828_at,37399_at,39788_at,40079_at,40142_at,40873_at,41260_at,41663_at,643_at}
12	11	1	1	{1171_s_at,127_at,211_at,31430_at,33590_at,33753_at,36797_at,37758_s_at,38098_at,38507_at,41556_s_at}
13	11	1	1	{1081_at,1559_at,243_g_at,32170_r_at,33219_at,34422_r_at,36102_at,36767_at,37920_at,40813_at,41000_at}
14	12	1	1	{1107_s_at,32622_at,32769_at,33720_at,34255_at,35772_at,36012_at,40143_at,40457_at,40863_r_at,41445_at,41812_s_at}
15	13	1	1	{1091_at,185_at,32314_g_at,32780_at,33012_at,36524_at,37967_at,39392_at,40673_at,40869_at,40958_at,535_s_at,991_g_at}
16	13	1	1	{1170_at,2068_s_at,31333_at,33430_at,34544_at,36165_at,38196_at,37563_at,39124_r_at,39147_g_at,40004_at,40889_at,868_at}

From the reducts thus generated, we generate rules to predict the genes or combination of genes that are responsible for a particular disease stage. Once the rules are generated we verify the accuracy of prediction using various test sets. In most of the cases the accuracy is found to be 1. However, there are some (very few, though) cases where the accuracy is 0. These data are considered as spurious data and are thus eliminated.

The portion of the huge rule set is shown in Table IV. This huge rule set is verified using test samples of data, which are further shown below for each class.

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

ISBN: 978-81-923777-9-7

TABLE IV. PORTION OF THE RULE SET

Rule set: insulin_effect85

(1-854...	Match	Decision rules
1	1	(1080_s_at=8)&(1916_s_at=7)&(33760_at=8)&(34544_at=6)&(37324_at=8)&(37481_at=6)&(38074_at=8)&(39655_at=8)=>(Decision={1|1})
2	1	(1080_s_at=8)&(1916_s_at=7)&(33760_at=8)&(34544_at=6)&(37324_at=9)&(37481_at=6)&(38074_at=8)&(39655_at=8)=>(Decision={1|1})
3	1	(1080_s_at=8)&(1916_s_at=7)&(33760_at=7)&(34544_at=6)&(37324_at=9)&(37481_at=7)&(38074_at=8)&(39655_at=7)=>(Decision={1|1})
4	1	(1080_s_at=8)&(1916_s_at=6)&(33760_at=7)&(34544_at=7)&(37324_at=8)&(37481_at=7)&(38074_at=7)&(39655_at=8)=>(Decision={1|1})
5	1	(1080_s_at=8)&(1916_s_at=7)&(33760_at=8)&(34544_at=7)&(37324_at=11)&(37481_at=7)&(38074_at=8)&(39655_at=7)=>(Decision={1|1})
6	1	(1080_s_at=8)&(1916_s_at=7)&(33760_at=7)&(34544_at=7)&(37324_at=11)&(37481_at=7)&(38074_at=8)&(39655_at=7)=>(Decision={1|1})
7	1	(1080_s_at=8)&(1916_s_at=6)&(33760_at=8)&(34544_at=7)&(37324_at=9)&(37481_at=7)&(38074_at=8)&(39655_at=8)=>(Decision={1|1})
8	1	(1080_s_at=8)&(1916_s_at=7)&(33760_at=8)&(34544_at=6)&(37324_at=10)&(37481_at=7)&(38074_at=8)&(39655_at=8)=>(Decision={1|1})
9	1	(1080_s_at=8)&(1916_s_at=6)&(33760_at=8)&(34544_at=6)&(37324_at=9)&(37481_at=6)&(38074_at=8)&(39655_at=8)=>(Decision={1|1})
10	1	(1080_s_at=8)&(1916_s_at=6)&(33760_at=8)&(34544_at=7)&(37324_at=10)&(37481_at=7)&(38074_at=7)&(39655_at=8)=>(Decision={1|1})

Results of experiments by train&test method: insulin_effect

		Predicted					
		1	2	3	No. of obj.	Accuracy	Coverage
Actual	1	1	0	0	1	1	1
	2	0	0	0	0	0	0
	3	0	0	0	0	0	0
	True positive rate	1	0	0			

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

Fig. 1. Test result for 1 sample of class1 predicted in class1 accurately

Results of experiments by train&test method: insulin_effect_3

		Predicted					
		1	2	3	No. of obj.	Accuracy	Coverage
Actual	1	0	0	0	0	0	0
	2	0	1	0	1	1	1
	3	0	0	0	0	0	0
	True positive rate	0	1	0			

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

Fig. 2. Test result for 1 sample of class 2 predicted in class2 accurately

Results of experiments by train&test method: insulin_effect_3_1

		Predicted					
		1	2	3	No. of obj.	Accuracy	Coverage
Actual	1	1	0	0	1	1	1
	2	0	0	0	1	0	0
	3	0	0	1	1	1	1
	True positive rate	1	0	1			

Total number of tested objects: 3
Total accuracy: 1
Total coverage: 0.667

Fig. 3. Test result for 2 out of 3 samples each for class 1 and 3 predicted

Results of experiments by train&test method: insulin_effect_4_2

		Predicted					
		1	2	3	No. of obj.	Accuracy	Coverage
Actual	1	1	0	0	1	1	1
	2	0	1	0	1	1	1
	3	0	0	2	2	1	1
	True positive rate	1	1	1			

Total number of tested objects: 4
Total accuracy: 1
Total coverage: 1

Fig. 4. Test results for 4 samples out of which 2 samples for class 3 and one each for class 1 and 2 predicted accurately

Results of experiments by train&test method: insulin_effect_10_3

		Predicted					
		1	2	3	No. of obj.	Accuracy	Coverage
Actual	1	5	0	0	5	1	1
	2	0	2	0	2	1	1
	3	0	0	3	3	1	1
	True positive rate	1	1	1			

Total number of tested objects: 10
Total accuracy: 1
Total coverage: 1

Fig. 5. Test results for 5, 2 and 3 objects repectively for classes 1,2 and 3 predicted accurately

The above results in Fig. 1, Fig. 2, Fig. 3, Fig. 4 and Fig. 5 verify the accuracy of classification using RSES.

V. DISCUSSION

The microarray data set contains 12626 attributes and 110 samples. The last attribute is the decision attribute, which takes on three values 1, 2, and 3 representing diabetes-sensitive, diabetes-resistive, and diabetic conditions, respectively. The entire data set is split up into a training set with 85 objects and the remaining objects as test set. We have used RSES2.2 to generate reducts using the training data set.

Some of the reducts thus generated from the entire data set shown in Table III. The rules generated from these reducts are shown in Table IV. Finally, the most important inferences drawn from this table are shown in Table V, which lists the expression values of the particular genes that are present in the rules extracted. 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 be diagnosed accordingly.

Note that no core reduct could be generated from the given data set. Even filtering of the reducts does not give good accuracy.

TABLE V. EXPRESSION VALUES OF GENES APPEARING IN THE EXTRACTED RULE SETS

Gene	Expression Values in the Rule Sets (rule number-wise) Obtained		
	Insulin-sensitive	Insulin-resistive	Diabetic
1080_at	7,8	6,7	7
1916_at	6,7,8,9,10	6	4,5,6,7,8
33760_at	7,8	7,8	7,8
33454_at	6,7	6,7	6,7,8
37324_at	5,6,7,8,9,11,12,13	8,9,10,11	7,8,9,10,11,12

ISBN: 978-81-923777-9-7

VI. Conclusion

Further experiments can also be carried out on these genes appearing in the rule sets generated (Table V) in the wet lab for further biological investigation with respect to the relevance of these genes in relation to diagnosis of insulin-effect. Investigation can also be carried out in the light of metabolic pathway engineering. This gives us future scope of research. In a nutshell, we have presented an approach based on rough set theory to diagnose and identify the genes responsible for insulin-effect. Determination of reducts, derivation of rule sets for insulin-sensitive, insulin-resistive, and diabetic people from the decision table are done. The contribution of the paper lies in the proposed approach for diagnosis of the disease and prominent genes responsible. This work can further pave way for detection of different insulin-effect 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," *Fundamenta Informaticae* 53(2), pp. 155-183, 2002.

[2] H. Midelfart, A. Lægreid and J. Komorowski, "Classification of Gene Expression Data in Ontology," *Proceedings of Second International Symposium on Medical Data Analysis*, LNCS 2199, pp. 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," *Genome Res.* 2003 May 13(5):965-79.

[4] K.G. Norsett, A. Laegreid, H. Midelfart, F. Yadetie, S.E. Erlandsen, S. Falkmer, J.E. Gronbech, H.L. Waldum, J. Komorowski, and A.K. Sandvik, "Gene expression based classification of gastric carcinoma," *Cancer Lett.* 2004 Jul 16; 210(2):227-237.

[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;2001.

[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] Maji and S.K. Pal, " Feature Selection Using f-Information Measures in Fuzzy Approximation Spaces" 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] Z. Pawlak, "Rough Sets." University of Information Technology and Management ul. Newelska 6, 01-447 Warsaw, Poland;1999.

[10] Wr´oblewski, J.: "Finding minimal reducts using genetic algorithms", *Proceedings of Second International Joint Conference on Information Sciences*, 1995.

[11] M. Banerjee, S. Mitra, and H. Banka "Evolutionary Rough Feature Selection in Gene Expression Data" *IEEE Transactions on Systems Man, and Cybernetics—PART C, APPLICATIONS AND REVIEWS, VOL. 37, NO. 4, JULY 2007.*

ISBN: 978-81-923777-9-7