IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Special Issue-1 E-ISSN: 2347-2693

Classification Rule Generation for Diabetic Patients using Rough Set Approach

Kasturi Ghosh*
University Institute of Technology
The University of Burdwan
Burdwan, India
kasturi.dikpati@gmail.com

Sripati Mukhopadhyay
Department of Computer Science
The University of Burdwan
Burdwan, India
dr.sripatim@gmail.com

Abstract— **Classification rule-generation is a Data Mining activity. A supervised process uses a training data set to generate the rules. The objective is to predict a predefined class or goal attribute, which can never appear in the antecedent part of a rule. The generated rules are used to predict the class attribute of an unknown test data set. In this paper we have tried to generate classification rule for diabetic patients using Rough set. This present research work relates Data mining to Health Informatics. The proposed algorithm generates the different classification rules related to predict insulin dose depending upon blood glucose measurement and helps in diabetes monitoring.**

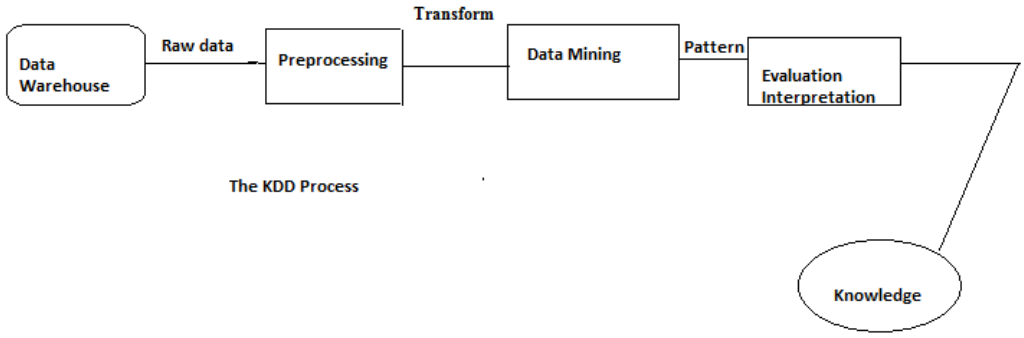
Keywords-Data Mining, Rough Set, Indiscernibility, Reduct, Decision tables and decision algorithms, Classification Rule

I. INTRODUCTION

Raw data are rarely of direct benefit. Its true value is predicted on i) the ability to extract information useful for decision support or exploration and ii) understanding the phenomenon governing the data source. In most domains data analysis was traditionally a manual process. One or more analysts would become intimately familiar with data and with the help of statistical techniques, provide summaries and generate reports. When the scale of data manipulation, exploration and inference goes beyond human capacities, people need the aid of computing technology for automating the process.

All these have prompted the need for intelligent data analysis methodologies, which could discover in database while some people treat Data Mining as a synonym for KDD [1].

In Data Mining interesting patterns or relationships among different data are discovered. Data mining tasks are also predictive since they classify the behaviour of the model based on available data. Data Mining uses automated tools that use different algorithms to discover hidden patterns, associations, redundancy, and structure from large amount of data stored in Data warehouses [1]. Data Mining also filters necessary information from a large data warehouse and infers the rules.



The KDD Process

Figure 1: KDD process

Rough Set Theory, proposed in 1982 by Zdzislaw Pawlak, is in a state of constant development. Its methodology is concerned with the classification and analysis of imprecise, uncertain or incomplete information and knowledge, and is considered as one of the first non-statistical approaches in data analysis [4, 5].

Rough Set theory [2] constitutes a consistency base for Data Mining; it offers useful tools for discovering patterns hidden in data in many aspects. The main feature of Rough Set data analysis in both non-invasive and notable ability is to handle qualitative data.

Rough Set can be used in different phases of the knowledge discovery process, as attribute selection, attribute extraction, data reduction, decision rule generation and pattern extraction.

Data Mining Technology provides a new thought for organizing and managing huge volume of data. Rough Set theory is one the important methods for knowledge discovery. This method can analyse intact data, obtain uncertain knowledge and offer an effective tool by reasoning.

Data mining with Rough Set is a multi-phase process consisting of mainly discretization deducts and rule generation on training set, classification on test set.

Rough set theory, since it has been widely used in Data Mining and has important functions in the expression, study and conclusion of uncertain knowledge, it is a powerful tool, which sets up the intelligent decision system.

II. FUNDAMENTAL CONCEPTS:

A. *Information System or Information Table:*

An Information System or Information Table [2] can be viewed as a table consisting of objects (rows) and attributes (columns). It is used in the representation of data that will be utilized by Rough Set, where each object has a given amount of attributes. These objects are described in accordance with the format of the data table, in which rows are considered objects for analysis and columns as attributes.

B. *Indiscernibility Relation:*

Indiscernibility Relation [2] is a central concept in Rough Set Theory, and is considered as a relation between two objects or more, where all the values are identical in relation to a subset of considered attributes. Indiscernibility relation is an equivalence relation, where all identical objects of set are considered as elementary.

C. *Decision tables and decision algorithms:*

A decision table [2] contains two types of attributes designated as the condition attribute and decision attribute. Each row of a decision table determines a decision rule, which specifies the decisions (actions) that must be taken when conditions are indicated by condition attributes are satisfied, determines the decision. A set of decision rules is designated as decision algorithms, because for each decision table it can be associated with the decision algorithm, consisting of all the decision rules that occur in the respective decision table. A may be made distinction between decision algorithm and decision table. A decision table is a data set, whereas a decision algorithm is a collection of implications, that is, a logical expression [5].

D. *Reduction attributes in information system:*

For many application problems, it is often necessary to maintain a concise form of the information system, but there exist data that can be removed, without altering the basic.

The process of reducing[2] an information system such that the set of attributes of the reduced information system is independent and no attribute can be eliminated further without losing some information from the system, the result is known as reduct. If an attribute from the subset B of A preserves the indiscernibility relation RA, then the attributes A - B are dispensable. Reducts are such subsets minimal, i.e., that do not contain any dispensable attributes. Therefore, the reduction should have the capacity to classify objects, without altering the form of representing the knowledge.

E. *Construction of Classification Rule:*

Classification Rule [2] construction for Data classification is based on a common principal divide and conquers. The algorithm attempts to divide a training set T into multiple disjoint subsets, so that each subset belongs to a

single target class. In its simplest form, a training set consisting of N records could be divided into N subsets, {T1, T2... TN}, such that each subset is associated with a single record and target class.

F. Health Informatics:

A definition about Health Informatics is provided by the National Library of Medicine, which defines Health Informatics as "the field of information science concerned with the analysis and dissemination of medical data through the application of computers to various aspects of health care and medicine"[8]. Zaiane provides an even more specific definition, which divides Health Informatics into four subfields:

"Health Informatics is the computerization of health information to support and optimize (1) administration of health services; (2) clinical care; (3) medical research; and (4) training. It is the application of computing and communication technologies to optimize health information processing by collection, storage, effective retrieval (in due time and place), analysis and decision support for administrators, clinicians, researchers, and educators of medicine."

III. PROPOSED ALGORITHM :

From the Diabetes database [6, 9] 4 to18 data rows are taken arbitrarily, only for code 58 (pre breakfast blood glucose measurement) and 60 (pre lunch blood glucose measurement) for patient id 01 to patient id 12. Other fields of the Information Table are patient id date, time and blood glucose measurement. Another field Insulin Dose [7] is added to Table1. Table1 is used as the input for the CRG (Classification Rule Generation) Algorithm.

Algorithm: CRG (time, code, blood glucose value, insulin dose)

***Input*:** time, code, blood glucose value, insulin dose

***Output*:** classification rules (related with blood glucose measurement and insulin dose) for pre breakfast and pre lunch

Step 1: Use indiscernibility relation for the conditional attribute "code" having the decision attribute "insulin dose" the same value.

Step 2: Apply data reduction on the conditional attribute "blood glucose value" for each output of step 1 and obtain ranges of "blood glucose value" for distinct insulin doses against each code.

Step 3: Find out missing ranges of the conditional attribute "blood glucose value".

Step 4: Use a suitable interpolation method (like Lagrange Interpolation) to find out the insulin dose for all the missing values of the conditional attribute "blood glucose value".

Step 5: Infer the classification rules.

Table-1(a): Output of Step 1 when insulin dose is 0 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
04/25/1991	7:29	58	67	0
04/26/1991	5:52	58	77	0
07/26/1990	05:17	58	70	0
07/28/1990	05:59	58	53	0
08/20/1990	06:19	58	83	0
08/21/1990	05:56	58	80	0
09/02/1990	08:01	58	66	0
09/04/1990	12:12	58	63	0
09/06/1990	06:12	58	60	0
09/06/1990	11:47	58	85	0

Table-1(b): Output of Step 1 when insulin dose is 0 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
07/23/1990	11:37	60	65	0
07/25/1990	10:13	60	84	0
08/23/1990	11:00	60	41	0
08/24/1990	11:26	60	38	0
08/25/1990	11:43	60	46	0
09/04/1990	12:12	60	63	0
09/06/1990	11:47	60	85	0

If code is 58 and blood glucose measurement is in the range 40-85 then insulin dose will be 0.

If code is 60 and blood glucose measurement is in the range 40-85 then insulin dose will be 0.

Table-2: Output of Step 1 when insulin dose is 10 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
04/21/1991	9:09	58	100	10
04/28/1991	8:42	58	109	10
04/29/1991	7:39	58	128	10
07/22/1990	05:58	58	92	10
07/24/1990	05:31	58	124	10
07/25/1990	06:25	58	92	10
07/27/1990	06:01	58	96	10
08/23/1990	07:15	58	102	10
09/04/1990	06:04	58	123	10
09/05/1990	06:59	58	118	10
04/06/1989	08:00	58	110	10

Table–3: Output of Step 1 when insulin dose is 11 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
10/10/1989	08:00	58	149	11
08/22/1990	07:05	58	149	11
08/25/1990	07:47	58	158	11
04/07/1989	08:00	58	159	11
07/20/1990	09:21	58	134	11
08/01/1990	08:48	58	132	11

If code is 58 and blood glucose measurement is in the range 92-128 then insulin dose will be 10.
If code is 58 and blood glucose measurement is in the range 132-158 then insulin dose will be 11.

Table -4 (a): Output of Step 1 when insulin dose is 12 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
10/11/1989	08:00	58	171	12
07/30/1990	05:26	58	171	12
04/29/1989	08:00	58	161	12
04/30/1989	08:00	58	170	12
05/03/1989	08:00	58	168	12
08/23/1990	07:27	58	185	12

Table-4(b): Output of Step 1 when insulin dose is 12 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
10/10/1989	12:00	60	116	12
07/28/1990	10:40	60	94	12
09/03/1990	12:38	60	115	12
09/05/1990	11:52	60	100	12
04/30/1989	12:00	60	126	12
05/03/1989	12:00	60	101	12
04/09/1989	12:00	60	129	12
04/12/1989	12:00	60	095	12
04/14/1989	12:00	60	103	12

If code is 58 and blood glucose measurement is in the range 161-185 then insulin dose will be 12.
If code is 60 and blood glucose measurement is in the range 94-129 then insulin dose will be 12.

Table -5(a): Output of Step 1 when insulin dose is 13 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
04/22/199	7:35	58	216	13

Table -5(b): Output of Step 1 when insulin dose is o13 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
10/11/198	12:00	60	148	13

1				
10/12/1989	08:00	58	220	13
10/14/1989	08:00	58	201	13
10/15/1989	08:00	58	206	13
07/21/1990	06:43	58	202	13
09/03/1990	08:04	58	192	13
05/01/1989	08:00	58	210	13

9				
10/14/1989	12:00	60	141	13
07/21/1990	10:57	60	134	13
07/27/1990	10:35	60	135	13
09/02/1990	11:40	60	136	13
05/01/1989	12:00	60	155	13
07/28/1990	14:17	60	134	13

If code is 58 and blood glucose measurement is in the range 192-220 then insulin dose will be 13.
If code is 60 and blood glucose measurement is in the range 134-155 then insulin dose will be 13.

Table -6(a): Output of Step 1 when insulin dose is 14 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
04/24/1991	7:52	58	239	14
08/02/1990	06:41	58	240	14
08/06/1990	06:51	58	236	14
05/01/1990	07:39	58	241	14
07/27/1990	10:50	58	232	14
09/04/1990	05:52	58	238	14

Table -6(b): Output of Step 1 when insulin dose is 14 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
05/02/1989	12:00	60	182	14
04/08/1989	12:00	60	175	14
05/03/1990	11:35	60	177	14
07/29/1990	13:27	60	181	14

If code is 58 and blood glucose measurement is in the range 232-241 then insulin dose will be 14 (Table 7(a)).
If code is 60 and blood glucose measurement is in the range 175-182 then insulin dose will be 14 (Table 7(b)).

Table -7(a): Output of Step 1 when insulin dose is 15 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
04/23/1991	7:25	58	257	15
04/27/1991	10:03	58	259	15
04/12/1989	08:00	58	256	15
08/10/1990	06:26	58	268	15
08/11/1990	06:50	58	253	15
08/24/1990	07:14	58	265	15

If code is 58 and blood glucose measurement is in the range 253-268 then insulin dose will be 15.

Table -7(b): Output of Step 1 when insulin dose is 15 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
04/29/1991	13:38	60	192	15
08/22/1990	11:38	60	194	15
04/29/1989	12:00	60	210	15

If code is 60 and blood glucose measurement is in the range 192-210 then insulin dose will be 15.

Table -8(a): Output of Step 1 when insulin dose is 16 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
10/13/1989	08:00	58	288	16
08/24/1990	07:46	58	298	16
05/02/1989	08:00	58	288	16
04/27/1990	08:18	58	309	16
05/03/1990	07:19	58	297	16

Table -8(b): Output of Step 1 when insulin dose is 16 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
04/06/1989	12:00	60	250	16
08/02/1990	12:13	60	236	16
05/07/1990	11:53	60	242	16
03/04/1989	12:00	60	244	16
07/27/1990	14:03	60	223	16

If code is 58 and blood glucose measurement is in the range 288-309 then insulin dose will be 16.
If code is 60 and blood glucose measurement is in the range 230-250 then insulin dose will be 16.

Table -9: Output of Step 1 when insulin dose is 17 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
08/08/1990	11:56	60	251	17
08/11/1990	11:59	60	251	17

If code is 60 and blood glucose measurement is in the range 251 then insulin dose will be 17.

Table -10(a): Output of Step 1 when insulin dose is 18 and code is 58

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
08/23/1990	07:01	58	336	18
08/18/1990	09:41	58	436	18

Table -10(b): Output of Step 1 when insulin dose is 18 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
10/13/1989	12:00	60	284	18
02/20/1989	12:00	60	294	18

If code is 58 and blood glucose measurement is >= 336 then insulin dose will be 18.
If code is 60 and blood glucose measurement is in the range 284-294 then insulin dose will be 18.

Table -11: Output of Step 1 when insulin dose is 20 and code is 60

Date	Time	Code	Blood Glucose Measurement	Insulin Dose
08/01/1990	12:57	60	339	20
08/12/1990	11:54	60	431	20
04/23/1990	11:22	60	350	20

Table-12: Reduced Table - output of Step2 for pre breakfast

CODE	BLOOD GLUCOSE MEASUREMENT	INSULIN DOSE
58	<=85	0
58	92-128	10
58	132-158	11
58	161-185	12
58	192-220	13
58	232-241	14
58	253-268	15
58	288-309	16
58	>= 336	18

If code is 60 and blood glucose measurement is >= 339 then insulin dose will be 20.

After combining all the tables which are related with code 58, reduction method is applied, instead of considering all the individual rows the range of the blood glucose measurement value whose insulin dose is same are taken. Conditional attributes date and time are also discarded.

Output of Step3 for pre breakfast:
So, the missing ranges are 86-89, 90-91, 129-131,159-60, 186-191, 221-231, 242-252, 269-287, 310-335.
Using Interpolation the values for all these ranges are as:
86-89…….....0
90, 91-129…….....10
130-139-159…….11
160,186-189……..12
190-191……13
221-231, 242-249…..14
250-252, 269-279…..15
280-287…….....16
>=310…….....18

Using Step 4 the modified table is shown in Table 13:

Table 13: Modified Reduced Table

CODE	BLOOD GLUCOSE MEASUREMENT	INSULIN DOSE
58	<=89	0
58	90-129	10
58	130-159	11
58	160-189	12
58	190-220	13

Table 14: Reduced Table: output of Step2 for pre lunch

CODE	BLOOD GLUCOSE MEASUREMENT	INSULIN DOSE
60	<=85	0
60	94-129	12
60	134-155	13
60	175-182	14
60	192-210	15

58	221-249	14
58	250-279	15
58	280-309	16
58	>=310	18

60	230-250	16
60	251	17
60	284-294	18
60	>= 339	20

After combining all the tables which are related with code 60 as shown above in Table 14, reduction method is applied, instead of considering all the individual rows the range of the blood glucose measurement value whose insulin dose is same are taken. Conditional attributes date and time are also discarded.

Output of Step3 for pre lunch:

So, the missing ranges are 86-89,90-93,130-133,156-159,160-174,183-189,190-191,211-219,221-229,252-279,280-283,295-310,311-338. Using Interpolation the values for all these ranges are as:

86-89…….....0
90-93…….....12
130-133,156-159…….....13
160-174, 183-189……14
190-191,211-219……..15
221-229…….....16
252-279…….....17
280-283,295-310…….....18
>=311-338…….....20

Using Step 4 the modified table is shown below in Table-15:

Table-15: Modified Reduce Table

CODE	BLOOD GLUCOSE MEASUREMENT	INSULIN DOSE
60	<=89	0
60	90-129	12
60	130-159	13
60	160-189	14
60	190-219	15
60	220-250	16
60	251-279	17
60	280-310	18
60	>= 310	20

Classification rules:

With the information deduct shown above, it can generate the necessary rules for aid to the insulin dose for the diabetic patients. The rules are presented below:

Output of Step 5:

Classification Rule for "pre breakfast"

Rule-1: If pre-breakfast Blood Glucose Measurement <=89 then insulin dose = 0
Rule-2: If pre-breakfast Blood Glucose Measurement is 90-129 then insulin dose = 10
Rule-3: If pre-breakfast Blood Glucose Measurement is 130-159 then insulin dose = 11
Rule-4: If pre-breakfast Blood Glucose Measurement is 160-189 then insulin dose = 12
Rule-5: If pre-breakfast Blood Glucose Measurement is 190-220 then insulin dose = 13
Rule-6: If pre-breakfast Blood Glucose Measurement is 221-249 then insulin dose = 14

Rule-7: If pre-breakfast Blood Glucose Measurement is 250-279 then insulin dose = 15
Rule-8: If pre-breakfast Blood Glucose Measurement is 280-309 then insulin dose = 16
Rule-9: If pre-breakfast Blood Glucose Measurement is >=310 then insulin dose = 18

Classification Rule for "pre lunch"

Rule-1: If pre-lunch Blood Glucose Measurement <=8 then insulin dose = 0
Rule-2: If pre-lunch Blood Glucose Measurement = 90-129 then insulin dose = 12
Rule-3: If pre-lunch Blood Glucose Measurement =130-159 then insulin dose = 13
Rule-4: If pre-lunch Blood Glucose Measurement =160-189 then insulin dose = 14
Rule-5: If pre-lunch Blood Glucose Measurement =190-219 then insulin dose = 15
Rule-6: If pre-lunch Blood Glucose Measurement =220-250 then insulin dose = 16
Rule-7: If pre-lunch Blood Glucose Measurement =251-279 then insulin dose = 17
Rule-8: If pre-lunch Blood Glucose Measurement =280-309 then insulin dose = 18
Rule-9: If pre-lunch Blood Glucose Measurement =>=310 then insulin dose = 20

IV. CONCLUSION

With the variation of the time how the insulin dose will vary for the same blood glucose measurement is not considered here. Another important factor in medical science to deal with diabetes is concerned with obesity and duration of suffering from the disease. Then the insulin dose will also vary which is also not considered in this paper. In future communication we shall try to incorporate these factors.

REFERENCES

[1] Sushmita Mitra and Tinku Acharya, DATA MINING Multimedia, Soft Computing, And Bioinformatics, Willy-Interscience, ISBN 9812-53-063-0. 2003.

[2] Julio Ponce and Adem Karahoca(Editors), "Rough Set Theory – Fundamental Concepts,Principals, Data Extraction, and Applications", Data Mining and Knowledge Discovery in Real Life Applications, I-Tech, Vienna, Austria, ISBN 978-3-902613- 53-0, pp. 438, February 2009.

[3] I. S. Jacobs P. Prabhavathy, Dr. B. K. Tripathy, "An Efficient Rough Set Approach in Querying Covering Based Relational Databases", International Journal of Computer Science and Business Informatics, Vol. 1, No. 1, ISSN: 1694- 2108 , MAY 2013

[4] Pawlak, Z, 1982, „Rough Sets", International Journal of Computer and Information science, vol.11, no.5, pp.341-356, 1982

[5] Pawlak, Z, „Rough sets - Theoretical aspects of reasoning about data", Dordrecht: Kluwer Academic Publishers, pp. 68-162, 1991

[6] Michael Kahn, " DIABETES data sets", AIM-94 data set, Washington University, St. Louis, MO, 1994.

[7] Klingensmith, GJ., American Diabetes Association, Intensive Diabetes Management, Third Edition, p. 107, 2003.

[8] National Library of Medicine, http://www.nlm.nih.gov/tsd/acquisitions/cdm/subjects58.html.

[9] https://archive.ics.uci.edu/ml/datasets/Diabetes