Medical Information System for Diabetes Mellitus by using Rough Set Theory

Hnin Cherry
Computer University, Pathein
hnincherryhc@gmail.com

Abstract

Rough set theory is based on the establishment of equivalence classes within the given training data. All of the data samples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data. The application of the rough set theory to identify the most important attributes and to induce decision rules from the medical data set with diabetes mellitus are discussed in this paper. Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. Hence, a medical information system for diabetes mellitus is developed using Rough Set Theory. Diabetes is a serious and rapidly escalating global health problem and one of the leading causes of death. Diabetes is caused by a defect in insulin secretion, insulin action, or both. Information system is needed to have patients who test themselves at home instead of clinical test. Applying Rough set , this system can be used for prediction or classification problem in the medical domain and shown high performance.

Keywords: Data mining, Rough Set Theory

1. Introduction

Rough set theory has been proved to be very useful in practice as clear from the record of many real life application; e.g. in medicine, pharmacology, engineering, banking, financial and market analysis [1, 6, 8, 9]. The theory of rough sets provides a power foundation for discovery of important regularities in data and for object classification. The rough set is defined as the pair of two crisp sets corresponding to approximations. If both approximations of a given subset of the universe are exactly the same, then one can say that the mentioned above subset is definable with respect to available information.

Rough set theory was introduced by Pawlak [2] and since then a number of applications have been reported in diverse fields such as medicine and process control. Rough sets can either be used for the purpose of generating if…then rules (machine learning) or as a technique for eliminating redundant information (data analysis) prior to the use of, say, artificial neural networks.

In this system, we employ a rough sets based classifier in order to determine which attributes in the input signature are important. It is based on the idea that any inexact concept (for example, a class label) can be approximated from below and from above using an indiscernibility relationship (generated by information about objects). The objective of this system is to analyze Rough sets theory and to make a decision for diabetes mellitus.

This paper is organized as follow: section 2 describes related work. In section 3, Rough set theory is explained with example. Section 4 illustrates system overview and implementation of proposed system. Finally we conclude this system in Section 5.

2. Related Works

Rough sets theory provides efficient algorithms for hidden patterns in data, finds minimal sets of data (data reduction), evaluates significance of data, and generates minimal sets of decision rules from data. It is easy to understand and offer straightforward interpretation of results [Pawlak, Z., 1996]. Those advantages can make the analysis easy that is why the rough sets approach is applied widely in many researches.

Rough sets theory has been applied to the analysis of many issues, including medical diagnosis, engineering reliability, expert systems, empirical study of material data, machine diagnosis, travel demand analysis, data mining, the research proposal of a general approach for a progressive construction of a rule-based assignment model to solve the linear programs.

The rough sets theory is useful method to analyze data and reduce information in a simple way. Rough set theory provides a new different mathematical approach to analyze the uncertainly, and with rough sets we can classify imperfect data or information easily. Pawlak [2] points out that one of the most important and fundamental notions to the rough sets philosophy are the need to discover redundancy and dependencies between features.

3. Rough Set Theory

The rough set theory, originally introduced by Pawlak [2] is chosen as a basic tool to analyse diabetes mellitus data. One of the main advantages of rough set theory is that it does not need any preliminary or additional information about data. Also, that programs implementing it's methods may easily run on parallel computers, rough set theory was proposed as a new approach to vague concept description from incomplete data is based on the lower and upper approximations [3].

This algorithm approaches for induction of rules using rough sets. The step for each attribute do indicates a single attribute, a pair of attributes, and so on. The step Select attribute indicates a single attribute during the first pass, a pair of attributes during the second pass (the previously selected attribute paired with each other attribute) and so on.

```
For each decision class do
Begin
  Initialise universe of objects
  Select decision class
  Find class relation
  Repeat
        For each attribute do
        begin
            Select attribute
            Find equivalence relation
            Find lower subset
            Find upper subset
            Calculate discriminant index
        end
        Select attribute with highest
         discriminant index
        Generate rules
        Reduce universe of objects
        Reduce class relation
  Until no objects with selected decision
      class
End
```

Figure1. Rough Set Algorithm

Rough set algorithm is shown in figure1. In this section, a simple example is employed in order to illustrate some basic definitions and concepts of the rough set theory [4,5].

Table1 shows a universe of objects containing 5 objects (i.e., rows). Each object is described using 4 attributes (A1, B1, C1, D1) and a decision class (Class) with 2 possible values (either + or -).

Table1. A simple decision table

Object	A1	B1	C1	D1	Class
u1	A1	b2	c1	d3	+
u2	A1	b2	c2	d2	-
u3	A2	b1	c3	d1	-
u4	A3	b2	c4	d3	+
u5	A1	b2	c5	d2	+

Each step of the algorithm is now illustrated using the data from this decision table.

The universe of objects, denoted by U, is a set containing all objects (i.e., all row-id's) of the decision table.

$$U = \{u1, u2, u3, u4, u5\}$$

The last column, denoted by D, is a set containing all possible values of the decision class. The user selects one of the two decision classes.

$$D = \{+, -\}$$

A class relation is the union of all objects which have the same class. A class relation is denoted by Y_c where c is the previously selected decision class. Assuming that the '+' class was chosen previously then

$$Y_+ = \{u1, u4, u5\}$$

An equivalence relation is the unions of all objects which for a selected attribute A (or attributes) have the same pre-defined attribute value v (or values). An equivalence relation is denoted by $R_v(A)$. For example with A=A1 and v=a1, the equivalence relation is

$$R_{a1}=\{u1,u2, u5\}$$

Since objects u1, u2, and u5 have a common value a1 for attribute A1. With A = B1 D1 and v=b2d3, the equivalence relation is

$$R_{b2d3}=\{u1, u4\}$$

Since objects u1 and u4 have common values b2d3 for attribute B1 D1. The union of all equivalence relations for a particular attribute is denoted by R (A).

For A = A1,

$$R(A)=\{\{u1, u2, u5\}, \{u3\}, \{u4\}\}$$

For A = B1D1,

$$R(A) = \{\{u1, u4\}, \{u2, u5\}, \{u3\}\}$$

The lower set is the union of all equivalence relations of the current attributes which are a subset of the class relation. Using set notation this is denoted as:

$$\underline{B}(A) = \bigcup_{R_v(A) \in R(A)} \{ R_v(A) \subseteq Y_c \} \_\_\_\_\_ (1)$$

The lower subset is denoted by $\underline{B}(A)$.

Setting $A = A_1$, $\underline{B}(A) = \{ u4 \}$.

Therefore, all the objects that are included in the lower set can be assigned with certainly to the selected class.

The upper subset is the union of all equivalence relations of the current attribute which when intersected with the class relation give a non-empty set.

$$\overline{B}(A) = \bigcup_{R_v(A) \in R(A)} \{ R_v(A) \cap Y_c \neq \emptyset \} \_\_\_\_\_ (2)$$

The upper subset is denoted by $\overline{B}(A)$.

Setting A=A1, $\overline{B}(A) = \{u1, u2, u4, u5\}$

Therefore, all the objects that are included in the upper set can probably be assigned to the selected class.

The discriminant index, denoted by α, is calculated using the following formula:

$$\alpha = 1 - \frac{\text{card } \overline{B}(A) - \text{card } \underline{B}(A)}{\text{card } U} \_\_\_\_\_ (3)$$

The description of the lower subset of the attribute with the largest discriminant index provides the rules. Assuming that attributes A1 had the largest discriminant index then its lower subset generates the following rule: if A1 is a3, the class is +.

The universe of objects is reduced using the following formula:

$$U^{new} = U - [\, U - \overline{B}(A) \cup \underline{B}(A) \,] \_\_\_\_\_ (4)$$

The class relation is reduced using the following formula:

$$Y_{new} = Y_c - \underline{B}(A) \quad \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (5)$$

4. System Overview and Implementation of Proposed System

4.1. System Overview

Our Proposed system design is shown in figure2. In this system, dataset is collected from medical experts and then two third of dataset (200) are used for training data and remaining one third of dataset (100) is used for testing data. Training data are analysed by rough set classification algorithm and then generate classification rules. Number of 17 rules are shown in below.

Rule(1) If Age=G and Gender=F and FH=Y and FBG=N and C=BH then Class =P

Rule(2) If Age=G and Gender=M and FH=Y and FBG=N and C=N then Class =P

Rule(3) If Age=G and Gender=M and FH=Y and FBG=N and C=H then Class =P

Rule(4) If Age=L and Gender=M and FH=N and FBG=N and C=H then Class =P

Rule(5) If Age=G and Gender=F and FH=N and FBG=N and C=H then Class =P

Rule(6) If Age=G and Gender=F and FH=Y and FBG=D and C=BH then Class =P

Rule(7) If Age=L and Gender=F and FH=N and FBG=N and C=BH then Class =P

Rule(8) If Age=L and Gender=M and FH=Y and FBG=N and C=BH then Class =P

Rule(9) If Age=G and Gender=F and FH=Y and FBG=D and C=N then Class =P

Rule(10) If Age=G and Gender=M and FH=Y and FBG=N and C=BH then Class =P

Rule(11) If Age=G and Gender=M and FH=Y and FBG=D and C=H then Class =P

Rule(12) If Age=G and Gender=F and FH=Y and FBG=N and C=H then Class =P

Rule(13) If Age=G and Gender=M and FH=N and FBG=D and C=BH then Class =P

Rule(14) If Age=G and Gender=F and FH=Y and FBG=D and C=H then Class =P

Rule(15) If Age=L and Gender=F and FH=Y and FBG=N and C=BH then Class =P

Rule(16) If Age=G and Gender=F and FH=N and FBG=N and C=BH then Class =P

Rule(17) If Age=G and Gender=F and FH=N and FBG=D and C=H then Class =P

Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples. User can input new data sample using the classification rules to decide whether diabetes mellitus is or not. Moreover, you can analyze percentage of male, female and age with diabetes mellitus.

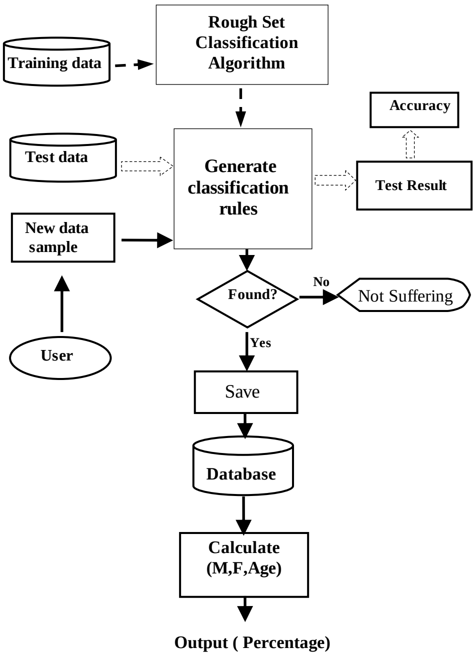


Figure2. System design

4.2. Implementation of the System

This system is an implementation for the diabetes mellitus patients. This system is divided into two phases. Firstly, the system checks whether the diabetes mellitus is or not. Secondly, the system calculates the percentage of male, female and age with diabetes mellitus.

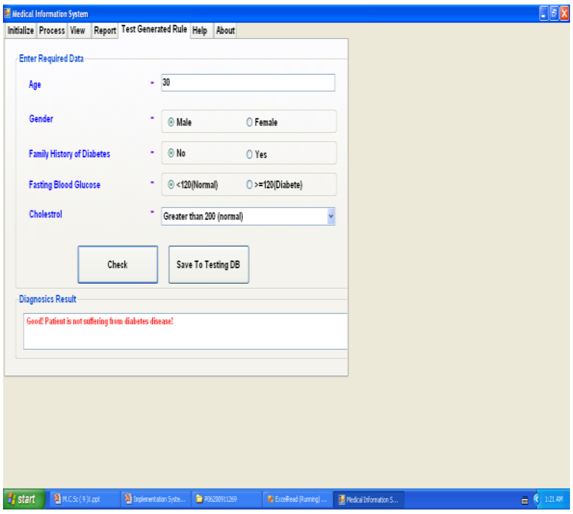
This system contains 5 attributes and classes for positive (P) and negative (N). These attributes and their values are shown in table2.

Table2. Attributes and their values

Attribute Name	Values
Age	<=34 (L), >35(G)
Gender	Female(F), Male(M)
Family History (FH)	Yes(Y), No(N)
Fasting Blood Glucose (FBG)	<120(Normal), >=120(Diabetes)
Cholesterol (C)	<200(Normal), 200-239 (Borderline High), >=240(High)

If user wants to test whether diabetes is or not, user select symptom of their diagnosis. The user will input Age and choose Gender, Family History of Diabetes, Fasting Blood Glucose, and Cholesterol from Test Generated Rule menu.

After filling with the attributes from form completely, user must use the Check button and then, user can see the message with "Good! Patient is not suffering from diabetes disease" or "Patient is suffering from diabetes disease" then data record must be saved as new record in database.



Enter Required Data

Age - 30

Gender - ● Male ○ Female

Family History of Diabetes - ● No ○ Yes

Fasting Blood Glucose - ● <120(Normal) ○ >=120(Diabete)

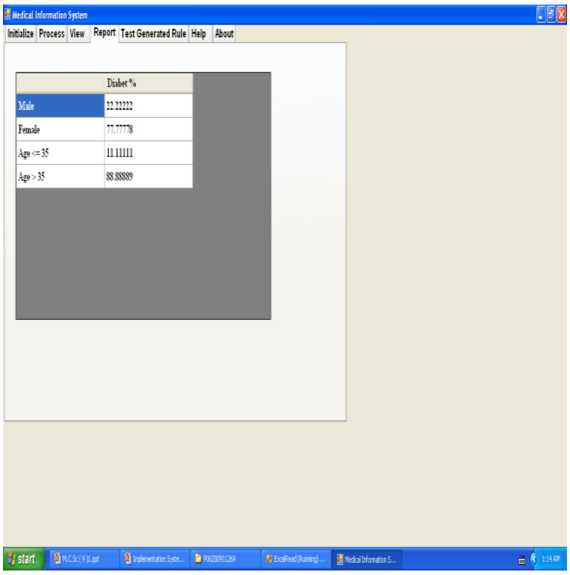
Cholestrol - Greater than 200 (normal)

Diagnosics Result

Good! Patient is not suffering from diabetes disease!

Figure3. Testing whether diabetes is or not

Moreover, user can also view information about diabetes. If user wants to know the percentage of the diabetes for gender and age, this system can be used from Report menu is shown in figure4.



	Diabet %
Male	22.22222
Female	77.77778
Age <= 35	11.11111
Age > 35	88.88889

Figure4. Percentage of diabetes

5. Conclusion

The diabetes mellitus data set has been drawn from a real life medical problem. The rough set based analysis showed that the most important aspects about diabetes mellitus. The result of our thesis and the extracted rules are also consistent with knowledge about diabetes. By using this system, user can know actually whether diabetes mellitus is or not because of 98% accuracy of process from test result. This paper the mining of patient data based on rough set theory to determine diabetes mellitus. We generate a decision rules by including the correct decision class and were able to predict with a high degree of accuracy whether the attempt was legitimate or not based on the decision rules that we generated from rough sets (98% or more classification accuracy). The rough set concept can be of importance for inductive reasoning when classification of objects is required on the basis of their properties. Rough set method goes beyond the individual application in diabetes mellitus.

6. References

[1] S.K.Pal, A.Skowron (Eds.): Rough-Fuzzy Hybridization A New Trend in Decision Making, Springer-Verlag, 1999.

[2] Pawlak, Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.

[3] Pawlak, Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.

[4] Pawlak, (1982) Rough Sets. International Journal of Computer and Information Sciences 11, 5: 341-356.

[5] Pawlak, (1984) Rough classification, International Journal of Man-Machine Studies 20:469-483.

[6] L.Polkowski, A.Skowron (Eds.): Rough Sets in Knowledge Discovery, Studies in Fuzziness and Soft Computing. Physica-Verlag, Heidelberg, 1998.

[7] A.Skowron, J.Stepaniuk .: Approximations of Relations, (ed.) W. Ziarko, Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer Verlag, London Berlin 1994, pp. 161-166.

[8] J. Stepaniuk: Tolerance Rough Sets and Boolean Reasoning in Knowledge Discovery, 6Th Int Conf, On Soft computing MENDEL 200, Brno, Czech Republic p.p 297-302.

[9] J. Stepaniuk: Tolerance Rough Sets and Information Granules, 7th Int Conf, On Soft computing MENDEL 2001, Brno, Czech Republic p.p 297-302.

[10] J.Stefanowski , K.Slowinski.: Rough Set Theory and Rule Induction Techniques for Discovery of Attribute Dependencies in Medical Information Systems, Lecture Notes in Artificial Intelligence 1263, Springer-Verlag, 1997, pp. 36-46.