Volume 5, Issue 7, July 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Novel Approach for Classification on Breast Cancer Data Set

Lekha Bhambu
(Phd Scholar)
Guru Kashi University, Talwandi Saheb
Punjab, India

Dr. Dinesh Kumar
(Associate Professor)
Guru Kashi University, Talwandi Saheb
Punjab, India

Abstract- Extensive amounts of knowledge and data stored in medical databases require the development of specialized tools for storing, accessing, analysis, and effectiveness usage of stored knowledge and data. Intelligent methods such as neural networks, fuzzy sets, decision trees, and expert systems are, slowly but steadily, applied in the medical fields. Rough set theory is an intelligent technique used for the discovery of data reduction, approximate set classification, and rule induction from databases.

In this paper, we present a rough set method for generating classification rules from a set of observed 699 samples of the breast cancer data. The rough set reduction technique based on discernibility matrix with some modification(K-Map) is applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification. Here we are use LEM2 algorithm to generate rules.Experimental results from applying the rough set analysis to the set of data samples are given and evaluated.

Key words- knowledge discovery, data mining, rough sets, breast cancer analysis, rule generation and reduction, reduct, K-map, LEM2.

I. INTRODUCTION

The growth of the size of data and number of existing databases far exceeds the ability of humans to analyze this data, which creates both a need and an opportunity to extract knowledge from databases (Cios *et al.*, 1998).Medical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Analysis of medical data is often concerned with treatment of incomplete knowledge, with management of inconsistent pieces of information and with manipulation of various levels of representation of data. Existing intelligent techniques (Lavrajc *et al.*, 1997; Wolf *et al.*, 2000) of data analysis are mainly based on quite strong assumptions (some knowledge about dependencies, probability distributions, large number of experiments), are unable to derive conclusions from incomplete knowledge, or can not manage inconsistent pieces of information. The most commonly intelligent techniques used in medical data analysis are neural network (Choi and Rockett, 2002; Setiono, 2000), Bayesian classifier (Cheeseman and Stutz, 1996), genetic algorithms (Grzymala-Busse *et al.*, 1999), decision trees (Hassanien,
2003), fuzzy theory (Parido and Bonelli, 1993). In this contribution the rough set theory is introduced. Rough set concept was introduced by Polish logician, Professor Z. Pawlak in early eighties (Pawlak, 1982; Pawlak, 1991; Pawlak *et al.*, 1995)

The theory of rough sets is a mathematical tool for extracting knowledge from uncertain and incomplete data based information. The theory assumes that we first have necessary information or knowledge of all the objects in the universe with which the objects can be divided into different groups. If we have exactly the same information
of two objects then we say that they are indiscernible (similar), i.e., we cannot distinguish them with known knowledge. The theory of RS can be used to find dependence relationship among data, evaluate the importance of attributes, discover the patterns of data, learn common decision-making rules, reduce all redundant objects and attributes and seek the minimum subset of attributes so as to attain satisfying classification.

This paper discusses how rough set theory can be used to analysis medical data, and for generating classification rules from a set of observed samples of the breast cancer data. The modified rough set reduction technique is applied to find all reducts of the data which contains the minimal subset of attributes that are associated with a class label for classification.

This paper is organized as follows. The characteristics of breast cancer data are discussed in Section 2. Theoretical aspects of rough set data analysis, which are relevant to the work and rule generation algorithm are introduced in Sections 3 and 4, respectively.K-map is introduced in section 5. Experimental results and discussion are reported in Section 6.Finally, conclusion is discussed in Section 7.

II. DATA COLLECTION AND KNOWLEDGE REPRESENTATION

2.1Characteristics of Breast Cancer Data

The data sets used in our experiments consists of 699 samples taken from fine needle aspirates from human breast tissue. They have been collected by Dr. W. Wolberg at the university of Wisconsim. Each sample consists of nine measurement

or features along with a label that denotes its class. Each instance has one of two possible classes: benign or malignant. These features has integer values in the range 1 to 10 as shown in Table 1.

Table 1: Condition and decision attributes of breast cancer dataset

Label	**Attribute**	**Domain**
A_1	Clump thickness	1–10
A_2	Uniformity of cell size	1–10
A_3	Uniformity of cell shape	1–10
A_4	Marginal adhesion	1–10
A_5	Single epithelial cell size	1–10
A_6	Bare nuclei	1–10
A_7	Bland_Chromatin	1–10
A_8	Normal Nucleoli	1–10
A_9	Mitoses	1–10
d= A_{10}	class	Benigm–Malignant

2.2. Information Systems

Knowledge representation in rough sets is done via information systems, which are a tabular form of an OBJECT → ATTRIBUTE VALUE relationship. More precisely, an information system, $\Gamma =< U,\Omega, V_q, f_q >q\varepsilon\Omega$, where

U is a finite set of objects, $U = \{x_1, x_2, x_3, \ldots, x_n\}$;

Ω is a finite set of attributes (features), the attributes in Ω are further classified into disjoint condition attributes A and decision attributes D,$\Omega = A \cup D$;

For each $q \in \Omega$,

- Vq is a set of attribute values for q,
- Each fq: $U \rightarrow Vq$ is an information function which assigns particular values from domains of attributes to objects such that $fq(xi) \in Vq$ for all $x_i \in U$ and $q \in \Omega$.

An example of a breast cancer information system is presented in Table 2. Each sample is a patient described in terms of the attributes $\{A_1, A_2, \ldots, A_9\}$.

III. ROUGH SET THEORY

Rough set theory proposed by Z. Pawlak in 1982, is a new mathematical approach to imperfect knowledge having close connections with many other theories. Despite its connections with other theories, the rough set theory may be considered as own independent discipline. Recently rough sets have been exhaustively

Table 2 : Information system for breast cancer dataset

Objects	Condition Attributes									Decision Attribute d
	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	
x_1	5	3	3	3	2	3	4	4	1	M
x_2	1	1	1	1	2	3	3	1	1	B
x_3	8	7	5	10	7	9	5	5	4	M
x_4	7	4	6	4	6	1	4	3	1	M
x_5	4	1	1	1	2	1	2	1	1	B
x_6	4	1	1	1	2	1	3	1	1	B
x_7	10	7	7	6	4	10	4	1	2	M
x_8	6	1	1	1	2	1	3	1	1	B
x_9	7	3	2	10	5	10	5	4	4	M
x_{10}	10	5	5	3	6	7	7	10	1	M
x_{11}	3	1	1	1	2	1	2	1	1	B
x_{12}	8	4	5	1	2	4	7	1	1	M
x_{13}	1	1	1	1	2	1	3	1	1	B
x_{14}	5	2	3	4	2	7	3	6	1	M
x_{15}	3	2	1	1	1	1	9	1	1	B

3.1Basic concepts of Rough Set Theory

An **information system** is a pair S = (U, P) where U is a non-empty finite set of objects called the universe and P is a non-empty finite set of attributes such that a: $U \rightarrow V_a$ for every aϵ P. The set V_a is called the value set of a.

A **decision system** is any information system of the form S = (U, P ∪ {d}), where d (is not element of P) is the decision attribute. The elements of P are called conditional attributes or simply conditions.

Let Q ⊆ P then Q-indiscernibility relation denoted by IND (Q), is defined as:

$$IND(Q) = \{(x, x') \in U^2 \mid \forall a \in Q\ a(x).a(x')\}$$

If (x, x') ϵ IND(Q), then objects x and x' are indiscernible from each other by attributes from Q. The equivalence classes of the Q-indiscernibility relation are denoted $[x]_Q$.

The **discernibility matrix** of S is a symmetric n× n matrix with entries c_{ij} as given as:

$$c_{ij}=\{ a\epsilon Q \mid a(x_i) \neq a(x_j)\} \text{ for } i, j=1,.....,n$$

Each entry thus consists of the set of attributes upon which objects x_i and x_j differ. Since discernibility matrix is symmetric and $c_{ii} = \emptyset$ (the empty set) for i=1,.,n, this matrix can be represented using only elements in its lower triangular part, i.e. for$1 \leq j < i \leq n$.

A **discernibility function**(f_S) for an information system S is a Boolean function of m Boolean variables a_1^* ,.....,a_m^*(corresponding to the attribute a_1,...,a_m) defined as :

$$f_s(a_1,..., a_m)= \forall\{\exists c_{ij}^* \mid 1 \leq j \leq i \leq n, c_{ij} \neq \varphi\}$$

Where $c_{ij}^* = \{a^* \mid a \in c_{ij}\}$. The set of all prime implicates of f_s determines the set of all reducts of Q. The discernibility function is a Boolean function in POS form.

3.2 Approximations of set

Let X is a subset of U, i.e. X ⊆ U.

- **Lower Approximation:** For a given concept, its lower approximation refers to the set of observations that can all be classified into this concept.

$$Q_*(X) = \{x:[x]_Q \subseteq X\}$$

- **Upper Approximation:** For a given concept, its upper approximation refers to the set of observations that can be possibly classified into this concept.

$$Q^*(X) = \{x:[x]_Q \cap X \neq \emptyset\}$$

Once the reducts have been computed, the rules are easily constructed by overlaying the reducts over the originating decision table and reading off the values.

IV. ALGORITHM USED

LEM2(Learning by Example Module, Version- 2) is a machine learning algorithm based on rough set theory. The usual task of LEM2 algorithm is to learn a discriminate rule set, i.e., to learn the smallest set of minimal rules, describing the concept. This algorithm can generate both certain and possible rules from a decision table with attributes being numerical as well categorical. LEM2 needs discretization for numerical attributes.

For inconsistent data, LEM2 induces two sets of rules: certain rule set and possible rule set. The first set is computed from lower approximations of concepts, the second one from upper approximations. It is assumed that the rule set will be used automatically by a classification component. Nevertheless, induced rules are available and comprehensible by the user. Thus, it is possible to use rules manually, like in other systems.

The LEM2 algorithm is a single local covering approach. It yields single minimal discriminate

description, which means, learning the smallest set of minimal rules for every concept. The local coverings are constructed from minimal complex. A formal definition of minimal complex and local covering is reported next.

Definition: Minimal complex and Local covering Let B be a nonempty lower or upper approximation of a concept represented by a decision-value pair (d,w). The set T is a minimal complex of B if and only if B depends on T and no proper subset T ' of T exists such that B

depends on T '. Let ℑ be a collection of non-empty set of attribute-value pairs for equivalence class [T] of T. Then ℑ is the local covering of B iff

- each member T of ℑ is a minimal complex of B
- U[T]=B and
- ℑ is minimal i.e., ℑ has the smallest possible number of members.

LEM2 algorithm is found suitable in rule generation for inconsistent data.

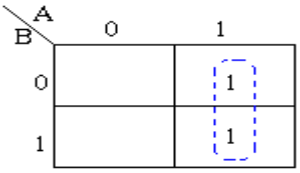
V. KARNAUGH-MAP (K-MAP)

The Karnaugh map, also known as the k-map, a method to simplify Boolean algebra expressions reduces the need for extensive calculations. In Boolean algebra, any Boolean function can be expressed in canonical form using the dual concepts of min-terms and max-terms. Product of Sums (POS) is a conjunction (AND) of max terms. The canonical forms allow for greater analysis into the simplification of Boolean functions. Here K-map is used for the reduction of attributes by simplifying discernibility function which is a Boolean function in POS form. Since discernibility function is a monotone Boolean function (Boolean function without negation), so negative terms in K-map are not considered [9], [10].

Advantages of K-map:

- It takes less space
- It takes less time

Example: Consider the following map. The function plotted is: Z = f (A, B) = A $\overline{B}$ + AB



- The values of the input variables form the rows and columns, which are the logic values of the variables A and B (with one denoting true form and zero denoting false form) form the head of the rows and columns respectively.
- The above map is a one dimensional type which can be used to simplify an expression in two variables.
- There is a two-dimensional map that can be used for up to four variables, and a three-dimensional map for up to six variables.

Using algebraic simplification,

Z = A $\overline{B}$ + AB

Z = A ($\overline{B}$ + B)

Z = A

Variable B becomes redundant due to Boolean Theorem. Referring to the map above, the two adjacent 1's are grouped together. Through inspection it can be seen that variable B has its true and false form within the group. This eliminates variable B leaving only variable A which only has its true form. The minimized answer therefore is Z =A.

VI. EXPERIMENT RESULTS & DISCUSSION

The aim of this experiment is to identify the essential subset of non-redundant attributes, which is relevant to determine the discovery task for a specific decision attribute. We try to extract relevant features from a given set of features with respect to the decision attribute class (*d*) shown in fogure 1

The experiments come with some steps.

- Find out the reduct
- Extract the rules
- Testing of generated rules

Firstly we compute the discernibility matrix for the conditional attributes.Since with every discernibility matrix, a discernibility function associate. So we obtain the discernibility function associated with the discernibility matrix.Then to simplify the discernibility function, we use K-map to find out the reduct. We observed that the use of K-map for the simplification of discernibility function saves time and space. Time and space are the most important factors for any algorithm.Since K-map is applicable upto 4 variables, after that it becomes so complex.So if whave more than 4 conditional attributes then we make the groups of the attributes and apply the procedure mentioned above to each group. Then we combine the result of all groups in order to generate reduct.Then apply the same procedure on new reduct and continues until no attribute is reducted from the table by applying this i.e. we have the attributes as they are. The advantages of using K-map is that it takes less time and less space. New reduct table is shown in table 3.

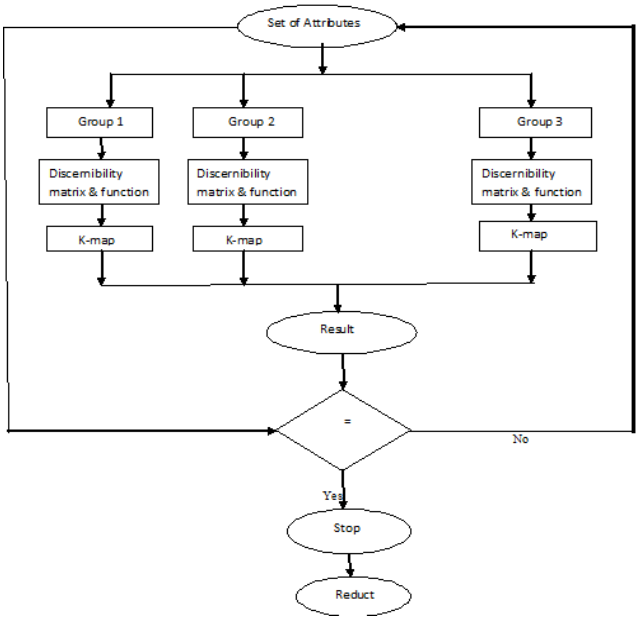


Fig 1: Flow Chart of Finding reduct

Table 3 : New Reduct Table For BCDS

Objects	Condition Attribute				Decision Attribute
	A_1	A_2	A_6	A_7	d
x_1	5	3	3	4	M
x_2	1	1	3	3	B
x_3	8	7	9	5	M
x_4	7	4	1	4	M
x_5	4	1	1	2	B
x_6	4	1	1	3	B
x_7	10	7	10	4	M
x_8	6	1	1	3	B
x_9	7	3	10	5	M
x_{10}	10	5	7	7	M
x_{11}	3	1	1	2	B
a_{12}	8	4	4	7	M
x_{13}	1	1	1	3	B
x_{14}	5	2	7	3	M
x_{15}	3	2	1	9	B

After finding reduct ,we hold a thresold value(t) and compare the value of every attribute for every object with this thresold value and change it as follows:

$$Va=1 \text{ if } Vq \geq t$$
$$Va=0 \text{ if } Vq < t$$

Here t=5. So the new data set is shown in table 4

Table 4 : New BCDS

Objects	Condition Attribute				Decision Attribute
	A_1	A_2	A_6	A_7	d
x_1	1	0	0	0	M
x_2	0	0	0	0	B
x_3	1	1	1	1	M
x_4	1	0	0	0	M
x_5	0	0	0	0	B
x_6	0	0	0	0	B
x_7	1	1	1	0	M
x_8	1	0	0	0	B
x_9	1	0	1	1	M
x_{10}	1	1	1	1	M
x_{11}	0	0	0	0	B
a_{12}	1	0	0	1	M
x_{13}	0	0	0	0	B
x_{14}	1	0	1	0	M
x_{15}	0	0	0	1	B

Then we generate rules using LEM2 algorithm.The total number of rules is equal to 9 for 699 sample data are given below

1 $(A_2=0)\&(A_7=0)\&(A_6=0)\&(A_1=0)=>(d=\{B[334]\})$
2 $(A_6=0)\&(A_1=1)\&(A_7=0)\&(A_2=0)=>(d=\{M[13],2[90]\})$
3 $(A_1=1)\&(A_6=1)\&(A_2=1)\&(A7=1)=>(d=\{M[87]\})$
4 $(A_1=1)\&(A_7=0)\&(A_6=1)\&(A2=1)=>(d=\{M[31]\})$
5 $(A_1=1)\&(A_2=0)\&(A_6=1)\&(A7=1)=>(d=\{M[24]\})$
6 $(A_1=1)\&(A_6=0)\&(A_2=1)\&(A7=1)=>(d=\{M[21]\})$
7 $(A_7=0)\&(A_1=1)\&(A_2=0)\&(A6=1)=>(d=\{M[18]\})$
8 $(A_2=1)\&(A_1=0)=>(d=\{M[22]\})$
9 $(A_6=0)\&(A_1=1)\&(A_2=1)\&(A_7=0)=>(d=\{M[12]\})$

VII. CONCLUSION

The experimental result for 699 samples is shown in Table 5 for a comparison with other existing classification approaches.

Table 5: Comparison With Other Classification Schemes

Approaches	Testing Accuracy(%)
FEBFC	95.14
IRSS	95.89
ARFIS	96.63
Proposed	96.4

As shown in Table 5, the proposed technique achieved quite compatible or competitive classification performance on the given data set.

REFERENCES

[1] Aboul Ella Hassanien and Jafar M.H. ALI(2004) 'Rough Set Approach for Generation of Classification Rules of Breast Cancer Data' INFORMATICA Vol. 15, No. 1, 23–38

[2] A. Chatterjee and A. Rakshit(2004)'Influential Rule Search Scheme (IRSS)– A New Fuzzy Pattern Classifier' IEEE Transaction On Knowledge and Data Engineering, vol. 16, no. 8, pp. 881-893

[3] C. Lee, A. Zaknich, T. Bräunl(2008) 'A Framework of Adaptive T-S type Rough-Fuzzy Inference Systems (ARFIS)'IEEE 2008 International Conference on Fuzzy Systems (FUZZ-IEEE 2008), pp. 567-574

[4] Junbo Zhang , Tianrui Li , Da Ruan b, Zizhe Gao and Chengbing Zhao (2012) 'A parallel method for computing rough set approximations' Information Sciences 194 (2012) 209–223 Elsevier

[5] Prerna Mahajan,Rekha Kandwal and Ritu Vijay(2012) ' Rough Set Approach in Machine Learning: A Review' International Journal of Computer Applications (0975 – 8887) Volume 56– No.10

[6] Shampa Sengupta and Asit Kr. Das(2012) 'Single Reduct Generation Based On RelativeIndiscernibility Of Rough Set Theory' International Journal on Soft Computing (IJSC) Vol.3, No.1

[7] Prasanta Gogoi, Ranjan Das, B Borah and D K Bhattacharyya(2011) 'Efficient Rule Set Generation using Rough Set Theory for Classification of High Dimensional Data' International Journal of Smart Sensors and Ad Hoc Networks (IJSSAN) ISSN No. 2248-9738 (Print) Vol-1, Issue-2

[8] Sabu M.K. and Raju G.(2011) 'Dimensionality Reduction – 'A Rough Set Approach' International Journal of Machine Intelligence ISSN: 0975–2927 & E-ISSN: 0975–9166, Vol 3, Issue 4, pp-349-353

[9] Saeki T.,Nishiura, T. And Kato,T.(2011) 'Studies on an effective algorithm to reduce the decision matrix-A technique on a rule extraction by rough set theory'Systems,Man and Cybernetics(SMC), IEEE Internationa Conference

[10] Puniethaa Prabhu and Duraiswamy(2010) 'Feature selection for HIV database using Rough system'. Second International conference on Computing, Communication and Networking Technologies,978-1-4244-6589-7/10/ IEEE

[11] Julie M. David and Kannan Balakrishnan(2010) 'Machine Learning Approach for Prediction of Learning Disabilities in School-Age Children' International Journal of Computer Applications (0975 – 8887) Vol 9– No.11

[12] Z. Pawlak(1982) 'Rough sets' International Journal of Computer and Information Sciences, 11,341-356, 1982.

[13] Shan N.' Rule discovery from data using decision matrices ' M.Sc. Thesis, University of Regina.

[14] "Simplifying Logic Circuits with Karnaugh Maps". The University of Texas at Dallas. Retrieved 7 October 2012.

