Available online at www.pelagiaresearchlibrary.com

Pelagia Research Library

Pelagia Research Library

Advances in Applied Science Research, 2011, 2 (3): 334-345

Advances in Applied Science Research

Pelagia Research Library

ISSN: 0976-8610
CODEN (USA): AASRFC

Classification and Rule Extraction using Rough Set for Diagnosis of Liver Disease and its Types

S. Karthik*, A. Priyadarishini, J. Anuradha and B. K. Tripathy

School of Computer Science and Engineering, VIT University, Vellore

ABSTRACT

The liver supports almost every organ in the body and is vital for our survival. Liver disease may not cause any symptoms at earlier stage or the symptoms may be vague, like weakness and loss of energy. Symptoms partly depend on the type and the extent of liver disease. Liver diseases are diagnosed based on the liver functional test. Though this disease cannot be predicted at earlier stage due to lack of symptoms and signs, in this paper we attempt to apply soft computing technique for intelligent diagnosis of liver disease. The classification and its type detection are implemented in three phases. In first phase, ANN classification is applied for classifying the liver disease. In second phase rough set rule induction using LEM algorithm is applied to generate classification rules. This rule induction overcomes the drawback of MLP and hence improves the accuracy. in third phase fuzzy rules are applied to identify the types of the liver disease. Using LEM algorithm 6 rules are generated with accuracy of 96% in correct classification. On applying rules generated by LEM, improves the classification accuracy by 6% compared to MLP. The 4 fuzzy rules are framed to indentify the types of liver disease.

Keywords: Artificial neural networks, Classification, Learn by Example(LEM), Rough sets, Rule extraction.

INTRODUCTION

Liver is the largest internal organ in the human body, playing a major role in metabolism and serving several vital functions e.g decomposition of red blood cells etc.,. Liver disease are usually caused by inflammation or damaged hepatocytes, registers a tenacious presence on the list of top ten fatal diseases in World. On the worldwide scale, live cancer, or more specifically hepatocellular carcinoma (HCC),remains the third most common cause of cancer-related deaths and the fifth-most frequent cancer with an estimated 560,000 new cases every year[1].LFTs (liver function tests) are a group of blood tests that can help to show how well a person's liver is working. LFTs include measurements of albumin, various liver enzymes (ALT, AST, GGT and ALP), bilirubin, prothrombin time, cholesterol and total protein. All of these tests can be performed at the same time. Mean corpuscular volume-A normal value in humans is 27 to 31 picograms/cell. AST or SGOT (aspartate aminotransferase or serum glutamic oxaloacetic transaminase) AST (SGOT) is not only found in the liver. It is also normally found in heart, muscle, brain, and kidney tissue. Injury to any of these tissues can cause an elevated blood level. AST (SGOT) normal range is 10-34 IU/L. Alkaline Phosphates (ALP) is an enzyme in the cells which line the biliary ducts of the liver. ALP is also found in other organs including bone, placenta, and intestine. When ALP is elevated, another test known as GGT (gamma-glutamyltransferase) can be ordered by the doctor to confirm that the elevated ALP is being derived from the liver or biliary tract.ALP normal range is 20-140 IU/L (international units per liter). Gamma-glutamyl transpeptidase (GGTP) or transferase (GGT) Gamma-glutamyl transferase (GGT) is an enzyme which is useful when compared to ALP. By comparing this two, it can be determined if the patient has bone or liver disease. GGT or GGTP normal range is 0-51 IU/L[2].

1.2 Types of liver disease

The types of liver disease includes Alcoholic Liver Diseases, Hepatitis, Acute Liver disease, Liver cancer etc., The following table1.1 shows the types of liver disease and its cause and conditions.

Table 1.1 Type of Liver Disease

Type of liver disease	Description	Causes/Conditions
Acute liver failure	Rapid decrease in liver function	Drugs, toxins, a variety of liver diseases
Hepatitis	Acute or chronic liver inflammation	Viruses, alcohol abuse, drugs, toxins, autoimmune, nonalcoholic fatty liver disease.
Liver cancer	A cancer that originates in the liver	Increased risk with cirrhosis and chronic hepatitis; hepatocellular carcinoma (HCC) is most common primary liver tumor
Cirrhosis	Scarring of liver tissue leads to decreased liver function	Can be caused by a variety of conditions but usually a result of chronic hepatitis, alcoholism, or chronic bile duct obstruction

Out of these types acute liver diseases, hepatitis are more predomenant in population. The former is occurred due to alcohol consumption and can be further classified into two viz. Alcoholic liver diseases and induced liver disease.

The structure of this paper is organized as follows Section 2 presents a comprehensive literature review on ANN Classification, Rule Extraction and Fuzzy rules. Section 3 gives the outline architecture of our plan of work in two phases. Section 4 gives the is experimental results of phase I and its results analysis. 5. Conclusion and future work.

Literature Survey

Over the years softcomputing plays the major role in machine learning and data analysis. In recent trend rough set has been widely used in decision making, medical diagnosis, data analysis. It can be applied to hetrogenous, numerical or categorical data sets [8].

Data warehouse and data mining

A data warehouse is a subject-oriented, integrated, time-varying, non-volatile collection of data in support of the management's decision making process. Data mining [14] is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes and detecting anomalies. The structures that are the outcome of data mining process must meet certain conditions: validity, understandability, utility, novelty and interestingness.

2.1.2 Knowledge Discovery in Databases

Knowledge Discovery [14], [15] in Databases is the process of identifying a valid, potentially useful and ultimately understandable structure in data. It involves selecting or sampling data from a data warehouse, cleaning or preprocessing it, transforming or reducing it, applying a data mining component to produce and evaluate a structure.

2.1.3 Basic Rough Sets

Let U be a universe of discourse, which cannot be empty and R be an equivalence relation or indiscernibility relation [13], [16], [19] over U. By U/R we denote the family of all equivalence class of R, referred to as categories or concepts of R and the equivalence class of an element x ∈ U is denoted by $[x]_{R.}$ By a knowledge base, we understand a relational system k = (U, R), when U is as above and R is a family of equivalence relation or indiscernibility relation over U and k is called an approximation space. Elementary sets in k are the equivalence classes of R and any definable set in k is a finite union of elementary sets in k.

Therefore for any given approximation space defined on some universe U and having a n equivalence relation R imposed on it, U is partitioned into equivalence classes called elementary sets which may be used to define other sets in k; Given that X ∈ U, X can be defined in terms of definable sets in k by the following

Lower approximation of X in A is the set

$\mathbf{R_*X = \square\{Y \in U \mid R: Y \ \square\ X\}}$

Upper approximation of X in A is the set

$\mathbf{R^*X = \square\{Y \in U \mid R: Y \ \square X \neq \emptyset\}}$

Another way to describe the set approximations is as follows. Given the lower and upper approximations R_*X and R^*X, of X a subset of U, the R-positive region of X is $POS_R(X)$ and is given by $POS_R(X) = R_*X$, the R-negative region of X is $NEG_R(X)$ and is given by $NEG_R(X) = U\text{-}R^*X$, and the boundary or the R-borderline region of X is BN_R (X) and is given by

BN_R (X) = $R^*X - R_*X$. The elements of R_*X are those of U which can certainly be classified as elements of X and the elements of R^*X are those elements of U, which can possibly be classified as elements of X, employing knowledge of R. We say that X is rough with respect to R if and only if $R_*X \neq R^*X$, equivalently BN_R (X) $\neq \emptyset$. X is said to be R-definable if and only if, R_*X = R^*X or BN_R (X) =$\emptyset$.

The tuple $\{R_*X, R^*X\}$ composed of the lower and upper approximation is called a rough set.

2.1.4 Rule Induction

Rule induction [12], [18] is one of the most important techniques of machine learning. Regularities hidden in data are frequently expressed in terms of rules; rule induction is one of the fundamental tools of data mining. Rules are generally in the following form

If (attribute$_1$, value$_1$) and (attribute$_2$, value$_2$) and (attribute$_n$, value$_n$) then (decision, value)

Data from which rules are induced are usually presented in a form similar to a table in which cases (or examples) are labels (or names) for rows and variables are labeled as attributes and a decision. Attributes are independent variables and the decision is a dependent variable. The set of all cases labeled by the same decision value is called a concept.

2.1.5 Classification

Extracting comprehensible classification rules is the most emphasized concept in data mining researches. In order to obtain accurate and comprehensible classification rules from data bases, a new approach was attempted on combining advantages of artificial neural networks (ANN) and swarm intelligence[1]. Artificial neural networks (ANNs) are a group of very powerful tools applied to prediction, classification and clustering in different domains. The main disadvantage of this general purpose tool is the difficulties in its interpretability and comprehensibility. In order to eliminate these disadvantages, an approach was developed to uncoverandde code the information hidden in the blackbox structure of ANNs. Therefore, knowledge extraction from trained ANNs for classification problems is carried out. This approach makes use of particle swarm optimization (PSO) algorithm to trans form the behaviors of trained ANNs into accurate and comprehensible classification rules. The weights hidden in trained ANNs A novel improvement in neural network training for pattern classification is presented in this paper. The proposed training algorithm is inspired by the biological plasticity property of neurons and Shannon’s information theory. This algorithm applicable to artificial neural networks (ANNs) in general, although here it is applied to multilayer perceptions (MLP). During the training phase, the artificial met plasticity multilayer perceptron (AMMLP) algorithm assigns higher values for updating the weights in the less frequent activations than in the more frequent ones. AMMLP achieves a more efficient training and improvesMLP performance [3].

2.1.6 MLP

We studied the efficiency of multilayer perceptron networks to classify eight different medical data sets with typical problems connected to their strongly non-uniform distributions between output classes and relatively small sizes of training sets. We studied especially the possibility mentioned in the literature of balancing a class distribution by artificially extending small classes of a data set.The results obtained supported our hypothesis that principally this does somewhat improve the classification accuracy of small classes, but is also inclined to impair the classification accuracy of majority classes[6].

2.1.7 Rule Extraction

In recent years, support vector machines (SVMs) were successfully applied to a wide range of applications. However, since the classifier is described as a complex mathematical function, it is rather incomprehensible for humans. This opacity property prevents them from being used in many real-life applications where both accuracy and comprehensibility are required, such as medical diagnosis and credit risk evaluation. To overcome this limitation, rules can be extracted from the trained SVM that are interpretable by humans and keep as much of the accuracy of the SVM as possible. It provides an overview of the recently proposed rule extraction techniques for SVMs and introduce two others taken from the artificial neural networks domain, being Trepan and G-REX. The described techniques are compared using publicly available datasets, such as Ripley's synthetic dataset and the multi-class iris dataset. The experiments show that the SVM rule extraction techniques lose only a small percentage in Performances compared to SVMs and therefore rank at the top of comprehensible classification techniques[7].

III. ARCITURE AND DESIGN

In this section we are going to discuss on present work and the work in progress of diagnosis liver disease and its categories. The following diagram gives the various steps involved in our proposed work. Phase 1 of this work implemented and explain in this section 4 and 5. Phase 2 we are currently working under phase 2. The architecture of the both phases are given bellow.

3.1 Phase 1: Classification of liver diseases

Step 1: From the data set extract 70% of data for training using Ann algorithm.
Step 2: MLP algorithm is applied for classification.Training, testing and validation is preformed.
Step 3: classification analysis is performed. The correct classified records and wrong classified records are identified. The accuracy of the classification is obtained.
Step 4: Calculate the classification error.

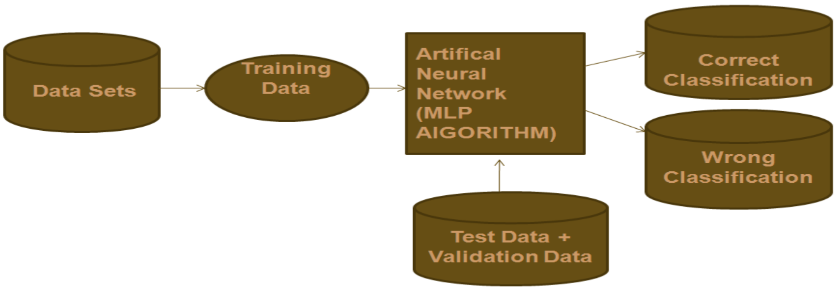
Data Sets
Training Data
Artifical Neural Network (MLP AIGORITHM)
Correct Classification
Wrong Classification
Test Data + Validation Data

Figure :3.1 Phase: 1 Liver diagnosis model

PHASE 2: Improving accuracy and identify types of liver diseases.
Step 1: Rule extraction is performed with correct classified data to improve the accuracy..
Step 2: Fuzzy rules is used to identify the type of liver diseases which is currently not done and will be implemented in future

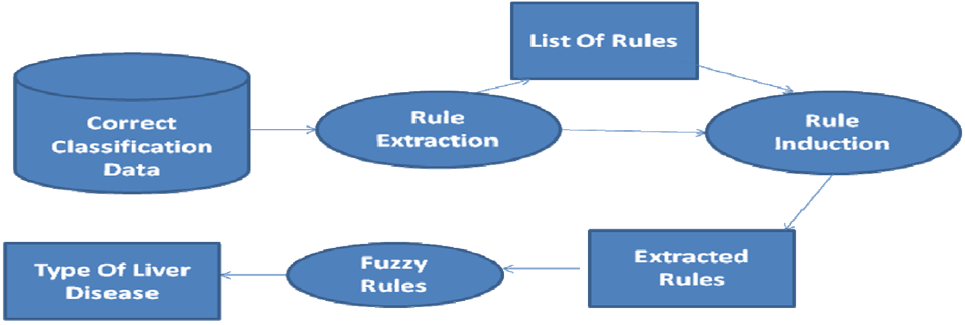

Figure:3.2 Phase: 2 Liver diagnosis proposed future model

3.2 LEM1

In general, rule induction algorithms may be categorized as global and local. In global rule induction algorithms the search space is the set of all attribute values, while in local rule induction algorithms the search space is the set of attribute-value pairs. We discuss lem1 rule induction algorithms.

Notations

A- set of all attributes
{d} - decision attribute
{d}* - partition of {d}
{G} - global cover where $\{G\} = \{g_1,g_2, \ldots, g_n\}$, $g_1,g_2, \ldots g_n \in A$
g_a - attribute name, a= 1, 2, … p
v_{ab} - value of the attribute g_a, b= 1, 2, …q
(g_a, v_{ab}) – denotes a attribute-value pair
R – set of rules generated

Algorithm:

```
for each record in decision table
R′: =∅ , G′: =G
while ( k > 1 )
G′: =G′- g_k
if ( ∩ (g_a, v_ab ) ) ≤ {d}* ) ∀ g_a ∈ G' of the record then
G′: =G′
else
G′:=G′ + {g_k}
k:=k-1
END if
if( k = 1 )
R=G′
```

```
else
R′=R′+{$g_k$}
END if
END while
R=R+{R′} END for
```

IMPLEMENTATION AND RESULTS

PHASE 1: Classification of Liver Disease

Algorithm: Back Propagation Algorithm: Levenberg-Marquardt
Levenberg-Marquardt (trainlm)

Like the quasi-Newton methods, the Levenberg-Marquardt algorithm was designed to approach second-order training speed without having to compute the Hessian matrix. When the performance function has the form of a sum of squares (as is typical in training feed forward networks), then the Hessian matrix can be approximated as

$$\mathbf{H} = \mathbf{J}^T\mathbf{J} \quad ....(5.1)$$

and the gradient can be computed as

$$\mathbf{g} = \mathbf{J}^T\mathbf{e} \quad ....(5.2)$$

where **J** is the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, and **e** is a vector of network errors. The Jacobian matrix can be computed through a standard back propagation technique (see [HaMe94]) that is much less complex than computing the Hessian matrix.

The Levenberg-Marquardt algorithm uses this approximation to the Hessian matrix in the following Newton-like update:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [\mathbf{J}^T\mathbf{J} + \mu\mathbf{I}]^{-1}\mathbf{J}^T\mathbf{e} \quad ------- (5.3)$$

When the scalar µ is zero, this is just Newton's method, using the approximate Hessian matrix. When µ is large, this becomes gradient descent with a small step size. Newton's method is faster and more accurate near an error minimum, so the aim is to shift toward Newton's method as quickly as possible. Thus, µ is decreased after each successful step (reduction in performance function) and is increased only when a tentative step would increase the performance function. In this way, the performance function is always reduced at each iteration of the algorithm.

4.1. Implementation of classification algorithm

For the given training data MLP(Levenberg-Marquardt) algorithm is applied. This algorithm takes 5 attributes(mcv,ast,alt,alp,ggpt) as input nodes with 1 hidden layer containing 20 hidden neurons. The output layer has one node.Total number of 200 records are considered for classification. Out of which 70% of the data is taken for training and 15% for testing and 15% for validation.

We obtained a result of 65.44% of accuracy after classification and the accuracy are computed as follows. The sensitivity and specificity measures can be used, respectively, for this purpose. Sensitivity is also referred to as the true positive (recognition) rate (that is, the proportion of positive tuples that are correctly identified), while specificity is the true negative rate (that is, the proportion of negative tuples that are correctly identified). In addition, we may use precision to access the percentage of tuples labeled as "Liver" that actually are "liver" tuples. These measures are defined as

Sensitivity=t_pos/pos.
Specificity=t_neg/neg.

where t pos is the number of true positives ("liver disease" tuples that were correctly classified as such), pos is the number of positive ("liver disease") tuples, t neg is the number of true negatives ("not liver disease" tuples that were correctly classified as such), neg is the number of negative ("not liver") tuples, and f pos is the number of false positives ("not liver disease" tuples that were incorrectly labeled as "liver disease"). It can be shown that accuracy is a function of sensitivity and specificity:

Accuracy=sensitivity(pos/(pos+neg))+specificity(neg/(pos+neg)).

Table 4.1:Attribute details

ATTRIBUTE	VALUE
Mcv	Nominal
Alkphos	Nominal
Sgpt	Nominal
Sgot	Nominal
Gammagt	Nominal
Selector	Nominal(healthy,disease)

Various classification algorithm is applied on complete dataset and corresponding result is obtained.Table 2 contain classification results.Table 3 contains accuracy of various able 4.2 clasification algorithm applied on complete dataset.

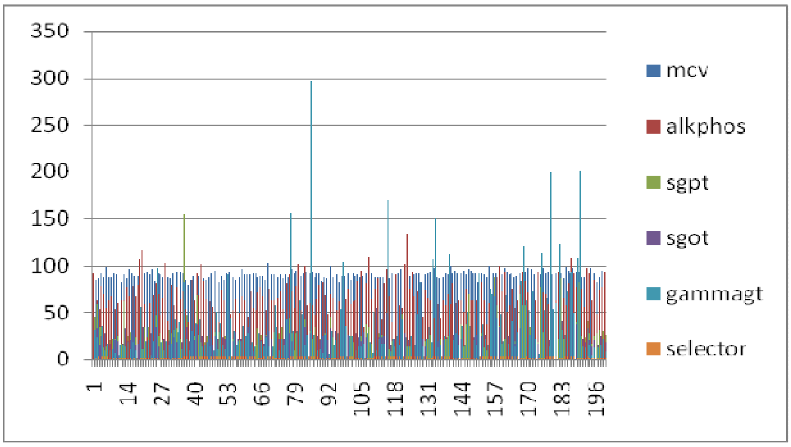

Figure:4.1. Attribute range

Table 2:classification result

ALGORITHM	CLASSIFICATION CORRECT	WRONG
NAIVEBAYES	55%	45 %
MLP	76.5 %	23.5%
RBFNETWORK	61.5%	38.5 %

Table 3:Accuracy

ALGO	CIR	SENSITIVITY	ACCURACY
NaviesBayes	0.55	0.397	55%
MLP	0.765	0.263	76.5%
RBFNetwork	0.615	0.466	61.5%

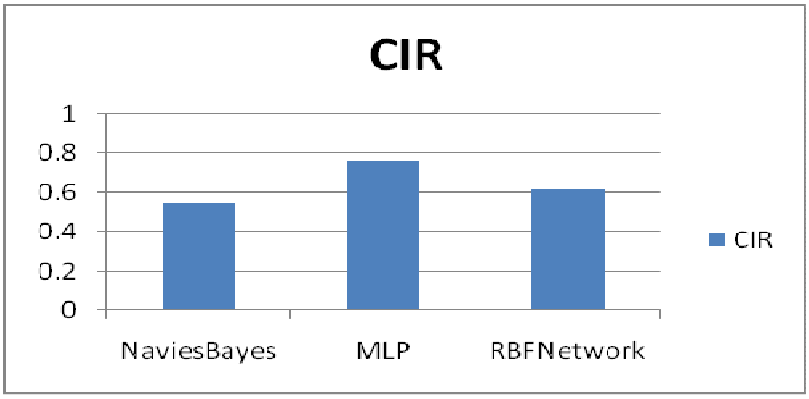

Figure 4: Comparision of various algorithm on basis of CIR

Table 4: Comparison

ALGO	ACCESS TIME	ERROR RATE
Navies Bayes	0.02sec	44.89%
MLP	0.33sec	32.52%
RBF	0.02sec	46.57%

Phase 2 Rules Generated by the LEM! above algorithm

Algorithm for LEM1

The following gives the algorithm for the new LEM1 approach, which is ELEM (Effective LEM1)
Rule1: [alp,norm]&[sgot,high]&[ggpt,norm]->[result,disea].
Rule2: [alp,norm]&[sgot,norm]&[ggpt,high]->[result,disea].
Rule3: [alp,high]&[sgot,norm]&[ggpt,norm]->[result,disea].
Rule4: [alp,low]&[sgot,norm]&[ggpt,norm]->[result,disea].
Rule5:[alph,norm]&[sgot,norm]&[ggpt,norm]-> [result,healthy liver]

FUZZY RULE FOR IDENTIFYING THE TYPES OF LIVER DISEASE

Rule 1: if(mcv=high)&(alp=norm)&(sgot=norm)&(ggmt=high)&(sgpt=norm)=Type1
Rule 2: if(mcv=high)&(alp=norm)&(sgot=norm)&(ggmt=norm)&(sgpt=high)=Type2
Rule 3: if(mcv=high)&(alp=norm)&(sgot=high)&(ggmt=norm)&(sgpt=norm)=Type3
Rule 4: if(mcv=high)&(alp=high)&(sgot=norm)&(ggmt=norm)&(sgpt=norm)=Type4

Type 1: Chronic Hepatitis
Type 2: Biliary Diseases
Type 3: Fatty liver Diseases
Type 4: Alcoholic liver Diseases

CONCLUSION

In spite of the constant advancement in the field of medical sciences, diagnosis of disease remains a challenging task. Liver disease in particular is not easily discovered at its initial stage; early diagnosis of this leading cause of morality is therefore highly important. As a part of the ongoing efforts to make diagnosis more effective, this study accordingly developes a two-phase intelligent diagnosis model aiming to provide a comprehensive analytic framework to raise the accuracy of liver diagnosis. In classification phase, MLP is employed to distinguish between healthy liver and diseased liver. In the concluding phase, Rule extraction is improved the accuracy.

Future Work

The context of this study, the model works effectively in helping physicians diagnosis of liver disease. The model, how-ever, may gave a complete exposure on soft computing techniques on classification, rule extraction, and fuzzy rules, it can be further applied to any machine learning techniques to have better performance in learnring with less computation. This work can be further extended to identify other types of liver diseases apart from the one focused here.

Acknowledgments
We would like to thank Dr.Deepa, VIT health centre for her valuable suggestions in fixing of the parameters and finding the type of liver disease.

Data source:
http://archive.ics.uci.edu/ml/datasets/Liver+Disorders
http://archive.ics.uci.edu/ml/machine-learningdatabases/Liver-disorders

REFERENCES

[1] Rong-Ho Lin a, Chun-LingChuang, *Computers in Biology and Medicine* 40, (**2010**),Page No: 665–670.
[2] Mehmet Aci , Cigdem _Inan, Mutlu Avci, *Expert Systems with Applications* 37, (**2010**) , Page No:5061–5067.
[3] A. Marcano-Ceden, J.Quintanilla-Domı´nguez,D.Andina, *Neurocomputing*, (**2010**).
[4] Lale Ozbaklr, Yilmaz Delice, *Engineering Applications of Artificial Intelligence*, (**2010**).
[5] J.H. Ang, K.C. Tan, A. Al-Mamun, *Neurocomputing* 71, (**2008**), page No: 3493–3508.
[6] Lassi Autio, Martti Juhola, Jorma Laurikkala, *Computers in Biology and Medicine* 37, (**2007**), page No:388 – 397.
[7] S.K. Majumder, A. Gupta, S. Gupta, N. Ghosh,P.K. Gupta, *Journal of Photochemistry and Photobiology B: Biology* 85,(**2006**), Page No: 109–117.
[8] B. Baesens, R. Setiono, C. Mues, J. Vanthienen, *Management Science* 49 (3), (**2003**), Page No: 312–329.
[9] J.M. Keller, H. Tahani, *Inform. Sci.* 62, (**1992**),Page No: 205-221.
[10] B. Baesens, R. Setiono, C. Mues, J. Vanthienen, *Management Science* 49 (3), (**2003**), Page No: 312–329.
[11] J.M. Keller, H. Tahani, *Inform. Sci.* 62, (**1992**),Page No: 205-221.
[12] Guo, S Wang, Z Y Wu, Z C & Yan, “A novel dynamic incremental rules extraction algorithm based on Rough set theory”. In proceedings of the fourth International Conference on machine learning and cybernetics pp 18 – 21.
[13] James F. Peters, Andrzej Skowron, Zdzislaw Pawlak, “Transactions on Rough Sets- I”, Lecture Notes in Computer Science 3100 Springer **2004**, ISBN 3-540-22374-6.
[14] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, 2ND edition, Morgan Kaufmann Publishers, An imprint of Elsevier (**2006**).
[15] Tseng, T. L.(Bill), Quantitative approaches for Information modeling, Ph.D. Dissertation, University of Iowa.
[16] Yu-Neng Fan, Tzu-Liang(Bill),Ching-Chin Chern,Chun-Che Huang, “Rule induction based on an incremental Rough Sets”, Expert Systems with Applications pp 11439 – 11450.

[17] Zdzislaw Pawlak and Andrzej Skowron,"Rudiments of Rough Sets" Institute of Mathematics Warsaw University Banacha 2, 02-097 Warsaw, Poland.
[18] Jerzy W. Grzymala-Busse, "Rule Induction", University of Kansas (**1998**).
[19] Z.Pawlak, "Rough Sets Theoretical Aspests Of Reasoning about Data", Kluwer acadmeic Publishers.