Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol 2, No.2, 2012

Using Rule Based Classifiers for the Predictive Analysis of Breast Cancer Recurrence

Srinivas Murti* , Mahantappa

Department of Computer Science and Engineering,PG Extension Center, Vishweshwarayya Technological University , Bagalkot, India

* E-mail of the corresponding author: s.sri108@gmail.com

Abstract

The Aim of this work is to assess the Effectiveness of Rule Based Classifiers to help an Oncology Doctor for prediction of Breast Cancer Recurrence ,286 Cancer patient data ,obtained from UCI Machine learning Repository ,are used to determine Recurrence Events for New patients .This dataset is processed with WEKA Data Mining Tool ,by applying Rule Based Classifiers(RIPPER,DT,DTNB) and Rule Set is generated .Further from Experimental Results, it has been found that DTNB is providing improved Accuracy compared to other two Classifiers .Based on the patients' characteristics and the Rule set generated by DTNB ,New patients may be labeled as developing or not 'Recurrence Events ',thus supporting an Oncology Doctor in making Decisions about disease in a shorter time.

Keywords: Medical Data Mining, KDD, Breast Cancer Recurrence, Classification, Association, WEKA, Rule Based Classifiers

1. Introduction

Breast Cancer is the leading cause of deaths in Women .One in Nine women is expected to develop Breast Cancer .Breast cancer can recur at any time, But most of the Breast Cancer recurrences occur in first three to five years of initial treatment .Breast cancer can come back as a local recurrence (in the treated Breast or mastectomy scar) or as a distant recurrence somewhere else in the body .

1.1 Problem Definition and our Proposed Technique

The nature of the Medical Data is Noisy, In complete and Uncertain. Too much of the Medical Data are now collected due to the computerization. Too many Disease markers (attributes) are also available in decision making. The Relationships and patterns within this Medical Data could provide New Medical Knowledge and enhance our Knowledge of disease progression and management .Evaluation of stored medical data using the tools like WEKA (Waikato Environment for Knowledge Analysis) may lead to discovery of trends and patterns. Techniques are needed to search large quantities of this Medical Data for these Patterns and Relationships. The major challenge facing Health Industry is the Quality of Service at affordable costs. A Quality of Service means diagnosing the patients correctly and treating them correctly. Poor Clinical Decisions can lead to disastrous results which is unacceptable. In this Context Medical Data Mining came into existence.

Data Mining also referred to as -Knowledge Discovery in Databases or KDD, is the search for Relationships and Global Patterns that exist in the Large Data Bases but are *"hidden"* among vast amounts of data . Classification and Association are two mechanisms to represent Extracted Information. Even most technologically advanced Hospitals in India have no such software that predicts a disease through KDD process or through a Data Mining Technique. Our Work attempts to predict efficiently the Breast Cancer Recurrence (as a particular disease) using Classification Data Mining Technique, Thus helping an Oncologist to diagnose and treat the disease in a short time and over come the problems stated above that Persists in the Medical Data and Health Industry.

The rest of the paper is organized as follows. The concept of Rule Based Classifiers and some of the advanced Rule Based Classifiers - discussed in Section 2; and in Section 3- Methodology for our proposed work has been detailed; Section 4 outlines the Results and Section 5 illustrates Conclusions and future work.

1.2 Literature Survey

Up to Now, Several studies have been reported that have focused on Medical Diagnosis .These studies have applied different approaches to the given problem and have achieved higher prediction accuracies ranging from 62% or higher, Using the dataset from UCI Machine Learning Repository .

Belciug,S. ; Gorunescu ,F. ; Salem ,A,-B.; Gorunsecu ,M.; have accessed the effectiveness of three different Clustering algorithms used to detect Breast Cancer recurrent events and Experimental Results showed that Best Performance was obtained by Cluster Network followed by SOM and K –means. Based on the Patients segmentation regarding the occurrence of Recurrent Events, New Patients were labeled according to their Medical Characteristics as developing or not recurrent events.

Machine learning approaches have focused on models (e.g., neural nets, Bayesian nets, hyperplanes) that are unfamiliar to most non-analyst users. Although data mining models in the form of if- then rules [Apte and Hong, 1996; Holte, 1993; Michal ski et al.,1986], decision trees [Quinlan, 1993] and association rules [Agrawal et al., 1996; Han and Fu, 1995; Liu et al.,1998] are considered to be easy to understand, problems arise when the size of trees or the number of rules become very large. Many successful applications that employ association rule mining algorithms tend to produce very large numbers of rules. Recent work on identifying “interesting” rules and on visualization techniques has been done to improve the comprehensibility of the models [Bayardo and Agrawal, 1999; Liu et al.,1999]. Decision trees are often viewed as one of the most comprehensible models. To their surprise, Kohavi and Sommerfield [Kohavi and Sommerfield, 1998] found that it took longer than expected to explain the meaning of decision tree models to their clients. Rule Based Classifiers like Decision Table, on the other hand ,are easy to interpret and explain because of widespread familiarity with the tabular representations in spreadsheets and relational databases.However, relatively little work has been done on algorithms that use decision tables as representations of hypotheses induced from data sets,perhaps because complex models tend to be more accurate [Kohavi and Sommerfield, 1998]. Users however, would often be willing to sacrifice some accuracy for a more easily interpretable model, and it would be useful to know how well Rule Based Classifiers like decision tables, DTNB (Decision Table with Naïve Bayes’) can perform in abstracting accurate yet comprehensible models from the Medical data.

Hence, We Propose –A Novice Predictive Modeling approach known as Rule Based Classification as an alternative technique to decision trees, Association rules, Neural networks, Bayesian Networks in context of Predictive analysis of Breast Cancer Recurrence .Further Owing the working principle of Rule Based Classifiers and their advantages, we have generated a Rule Set which can predict Breast Cancer Recurrence using the Patients’ Characteristics (attributes); thus Helping an Oncologist in Predicting Breast Cancer Recurrence

2. Rule Based Classifiers

A Rule Based Classifier is a technique for classifying records using a collection of “If …. Then... ” Rules .Figure1 shows an example of a Model generated by a Rule Based Classifier for a vertebrate Classification problem .The Rules for the model are represented in Disjunctive Normal Form ,R=(r1 v r2 v ...rk),Where R is called as a *Rule Set* and ri’s are called *Classification Rule* or Disjuncts

Table 1. Example of Rule Set Vertebrate Classification Problem

r1:(gives birth=no)^ (Aerial creature=yes) → Birds
r2:(gives birth=no) ^(Aquatic Creature =yes) → Fishes
r3(Gives Birth=yes)^(BodyTemperature=Warm) → Mammals
r4(Gives Birth =no) ^(Aerial Creature=no) → Reptiles
r5:(Aquatic Creature=semi) → Amphibians

2.1 Advantages of Rule Based Classifiers

- Produce Descriptive Models ,which are easier to interpret, but gives comparable performance to decision tree Classifiers

- The Rule Based Classifiers create rectilinear partitions of attribute space and assign class to each partition, nevertheless if Rule Based Classifier allows multiple rules to be triggered for a given record ,then more complex decision boundary can be constructed

2.2 Working Principle

The Rule Based Classifier classifies a test record based on rule triggered by a record .To illustrate how a Rule Based Classifier works ,Consider a rule set shown in the Table 2 and the following Vertebrates

Table 2 .Working Principle of Rule Based Classifiers

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature
Lemur	Warm Blooded	Fur	Yes	No
Turtle	Cold Blooded	Scales	No	Semi
Dog Fish	Cold Blooded	Scales	Yes	Semi
Shark	Cold Blooded	Scales	Yes	Yes

The Working Principle of Rule Based Classifiers is shown in Table 2.A Rule Based Classifier classifies a test record based on the rule triggered by the record .To illustrate how a Rule Based Classifier works ,Consider the Rule Set shown in Table2 and the following vertebrates

- The First Vertebrate, which is a Lemur, is aWarm Blooded and gives birth to its young. It triggers the Rule r3 ,and thus ,is Classified as a *"Mammal"*.
- The second vertebrate, which is a Turtle, triggers the rule r4 and r5, since the Classes predicted by the rules are contradictory (reptiles versus amphibians),their conflicting classes must be resolved
- None of the Rules are applicable to Dogfish Shark. In this case ,we need to ensure that the classifier can make a reliable prediction even though a teat record is not covered by any Rule

2.3 Advanced Rule Based Classifiers

Recently Number of Rule Based Classifiers has been proposed to get good Rule set; some of them are as follows

2.3.1 Ripper

In our Experiment Java class implementing a Propositional Rule learner, RIPPER (Repeated Incremental Pruning to produce Error reduction) which was proposed by William .W.Cohen is run against the Breast Cancer Dataset taken from UCI Machine Learning Repository in WEKA Environment. The Algorithm is briefly described as follows

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one,

DO:

1. Building stage:

Repeat 1.1 and 1.2 until the description length (DL) of the rule set and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

1.1. Grow phase:

Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

1.2. Prune phase:

Incrementally prune each rule and allow the pruning of any final sequences of the antecedents; The pruning metric is (p-n)/(p+n) -- but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

2. Optimization stage:

After generating the initial rule set {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The variant with the minimal DL is selected as the final representative of Ri in the ruleset.After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.

3. Delete the rules from the rule set that would increase the DL of the whole rule set if it were in it. and add resultant rule set to RS

END DO

2.3.2 Decision Table

In our Experiment, A Java Class for building and using a simple decision table majority Classifier is run against the Breast Cancer Dataset derived from UCI Machine Learning Repository. Given a training set of labeled instances, an induction algorithm builds a classifier.There are two variants of decision table classifiers based conceptually on a simple lookup table. The classifier, called DTMaj (DecisionTable Majority) returns the majority of the training set if the decision table cell matching the new instance is empty, i.e., it does not contain any training instances.

A decision table has two components- A schema, which is a list of attributes and a A body, which is a multiset of labeled instances. Each instance consists of a value for each of the at- tributes in the schema and a value for the label. The set of instances with the same values for the schema attributes is called a cell.

Given an unlabeled instance, ~x, the label assigned to the instance by a decision table classifier is computed as follows. Let I be the set of labeled instances in the cell that exactly matches the given instance ~x, where only the attributes in the schema are required to match and all other attributes are ignored. If I ≠0;; return the majority class in I , breaking ties arbitrarily. Otherwise (I = 0;), A DTMaj returns the majority class in the decision table ,can be implemented in a universal hash table

2.3.3 Decision Table with Naïve Bayes'*(DTNB)*

In our Experiment, A Java Class for building and using a decision table/naive bayes hybrid classifier is run against Breast Cancer dataset derived from UCI Machine Learning Repository. The algorithm for learning the combined model (DTNB) proceeds in much the same way as the one for stand-alone DTs. At each point in the search it evaluates the merit associated with splitting the attributes into two disjoint subsets: one for the DT, the other for NB. We use a forward selection, where, at each step, selected attributes are modeled by NB and the remainder by the DT and all attributes are modeled by the DT initially. At each step, the algorithm also considers dropping an attribute entirely from the model.

3. Proposed Methodology

The Breast Cancer Dataset obtained from UC- Irvine archive of Machine Repository has 201 instances of One Class and 85 instances of other Class .The instances are described by 9 attributes some of which are Linear and some are Nominal.

Attribute Information:

1.Class:no-recurrence-events,recurrence-events
2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89,90-99.
3.menopause:lt40,ge40,premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39,40-44,45-49,50-54,55-59.
5.inv-nodes:0-2,3-5,6-8,9-11,12-14,15-17,18-20,21-23,24-26,27-29,30-32,33-35,36-39.
6.node-caps:yes,no.

7.deg-malig:1,2,3.
8.breast:left,right.
9. breast-quad: left-up, left-low, right-up, right-low, central.
10. Irradiat: yes, no.

The Proposed Methodology for the automation of Breast Cancer Health Records is depicted in fig 1. The Experiments are carried out as depicted in the fig1. Initially Breast Cancer Data Set Collected from UCI Machine Learning Repository is Pre Processed to Remove Missing Values in order to make it suitable for Mining Process. Once the Pre Processing gets over ,Rule Based Classifiers like DT,DTNB,RIPPER are applied one by one, and accuracy is measured .DTNB gives better accuracy compared to other two rule based classifiers .Hence the Rule Set generated by DTNB is stored in MS-Access Database

A New patient is labeled by validating the Rule set shown in Table 7 generated by DTNB and by using the working principle of Rule Based Classifier and Patients' Characteristics.

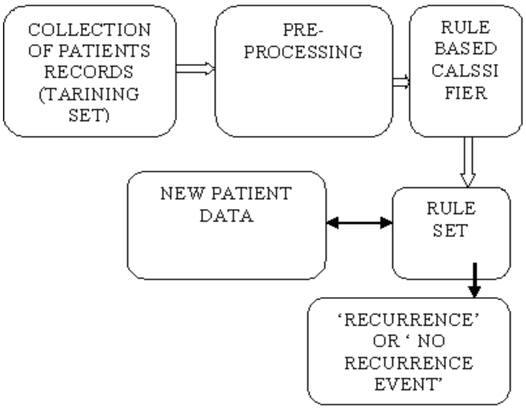
COLLECTION OF PATIENTS RECORDS (TARINING SET)
PRE-PROCESSING
RULE BASED CALSSI FIER
NEW PATIENT DATA
RULE SET
'RECURRENCE' OR ' NO RECURRENCE EVENT'

Figure 1. .Proposed Methodology for Breast Cancer Recurrence Detection

4. Experimental Results

The Classification Accuracy of the three classifiers run against Breast Cancer dataset is shown in Table 3

Table 3 -Classification Accuracy of RIPPER, DT, and DTNB

Sl no	Dataset	RIIPER	DT	DTNB
1	Breast Cancer	72.27%	72.72%	75.17%

The Confusion Matrices of *RIPPER, DT, and DTNB,* when run against Breast Cancer Dataset are shown in Table 4, 5, 6 respectively. Table 7 gives the overview of Rule Set Generated by DTNB.

We Employed four Performance Measures –Precision, Recall-Measure and ROC Space .A distinguished Confusion Matrix (Contingency Table) is obtained to calculate the four measures .It contains ,The cell which denotes the number of Samples Classified as true while they were true(TP) and the Cell which denotes the number of Samples Classified as False while they were False(TN) . The other two cells denote the Number of Samples Misclassified –Specifically the Cell denoting Number of Samples Classified as False while they were actually true (FN) and the Cell denoting Number of Samples Classified as True while they were actually False (FP) .Once the Confusion Matrix is constructed ,The Precision Recall and F-Measure are calculated as follows.

Recall=TP/ (TP+FN) (1)

Precision=TP/ (TP+FP) (2)

F-Measure= (2*TP)/ (2*TP+FP+FN) (3)

Precision measures percentage of actual Patients (TP) among patients that get declared disease .Recall measures percentage of actual patients that were discovered .F-Measure balances between Precision and Recall. A ROC (Receiver Operator Characteristic) space is defined by FP Rate and TPR Rate as X and Y Axes respectively

TPR=TP/ (TP+FN) (4)

FPR=FP/ (FP+TN) (5)

Table 4: Confusion Matrix of RIPPER

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC - Area	Class
0.846	0.635	0.759	0.846	0.8	0.589	No-Recurrence-Events
0.365	0.154	0.5	0.365	0.422	0.589	Recurrence-Events

Table 5: Confusion Matrix of DT

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC-Area	Class
0.91	0.765	0.738	0.91	0.815	0.638	No-Recurrence-Events
0.235	0.09	0.526	0.235	0.325	0.638	Recurrence-Events

Table 6: Confusion Matrix of DTNB

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.896	0.35	0.769	0.896	0.828	0.676	No-Recurrence-Events
0.365	0.104	0.596	0.365	0.453	0.676	Recurrence-Events

Table 7: Rule Set of DTNB

age	menopause	tumor-size	inv-nodes	breast	breast-quad	Class
50-59	premeno	35-39	6-8	right	right up	no-recurrence-events
60-69	ge40	15-17	15-19	left	right-low	no-recurrence events
40-49	premeno	45-49	12-14	right	right up	no-recurrence events

5. Conclusion and Future Work

In this paper, we have discussed need for Data mining in the Medical field. In this context we have discussed a New Predictive Modeling approach known as Rule Based Classification as an alternative technique to Neural Networks, Bayesian Networks, Decision Trees and Association Rules along with its advantages and working principle. We have assessed the performance of three rule based classifiers namely DT, RIIPER and DTNB and found that classification accuracy of DTNB is better compared to other two classifiers, when run against Breast Cancer Dataset obtained from UCI Machine learning Repository. Further we have found Rule set containing some interesting rules which were easy to interpret and familiar to represent them in spreadsheet were obtained from DTNB.The Experimental results also reveal that DTNB is an efficient approach for extraction of patterns from Breast cancer dataset. In the Future Work, Decision Table can be combined with Genetic Algorithm to reduce the attributes for the prediction of Breast Cancer Recurrence through a Graphical User Interface or -Fuzzy Weighted Association Rule Mining Technique may be used along with a GUI and may be used to label new patients as developing or not Breast Cancer Recurrence Events

References

Cleveland Clinical Database http:// www.clevelandclinic.org/health

Feelders, A., Daniels, H. and Holsheimer, M. (2000) ‘Methodological and Practical Aspects of Data Mining’, Information and Management, 271-281

UCI Machine Laening Repository http://www.ics.uci.edu/`mlax/MLrepository.html

Belciug,S. Gorunescu,F. Salem ,A,-B.Gorunsecu ,M.(2010).,”Clustering Based Approach for detecting Breast Cancer Recurrence” .In International Conference on Intelligent system Design and Application (ISDA) ,533-538

Jiawei Han, Micheline Kamber, and Jian Pei (2001), “Data Mining: Concepts and Techniques “3rd edition, Morgan Kaufmann publication (Chapter 1)

William W. Cohen(1995): “Fast Effective Rule Induction.” In: Twelfth International Conference on Machine Learning, 115-123

Ron Kohavi(1995): The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189

Mark Hall, Eibe Frank: Combining Naive Bayes and Decision Tables(2008), In: Proceedings of the 21st Florida Artificial Intelligence Society Conference (FLAIRS),318-319

D.SenthilKumar ,G.SathyaDevi and S.Sivanesh(May 2011) “Decision Support System for Medical Diagnosis using Data Mining ”In International Journal of Computer Science Issues(IJCSI) , Vol 3, Issue 3, No.1,147-153

H.W. Ian ,E.F(2005), “Data Mining Practical Machine Learning Tools and Techniques ”,Morgan Kaufman publication

First Author Srinivas Murti received the B.E degree in Information Science and Engineering from Vishweswarayya Technological University (VTU), Belgaum (INDIA) in 2008.He is currently an M.Tech (QIP) Candidate in Department of Computer Science and Engineering, PG Extension Centre, VTU, Bagalkot.He has completed 2.5 Years of teaching in various courses of Undergraduate Engineering Programs. He has Industrial Experience of 1 year in Software Development and Maintainance.His Research Interests include Medical Data Mining, Machine learning and related Real world applications and published various research papers in leading journals and conferences

Second Author Mahantappa is currently an M.Tech (QIP) Candidate in Department of Computer Science and Engineering, PG Extension Centre, VTU, Bagalkot.He has completed 2.5 Years of teaching in various courses of Undergraduate Engineering Programs. His Research interests include Image processing, Radio Frequency Identification Devices (RFID) and Technology, Networking and Data Mining