Discussing the Role of Classification Algorithms in Clinical Predictions with help of Case Studies

Mohammad Taha Khan
Research Scholar, Suresh Gyan Vihar University
Mahal Jagatpura ,Jaipur,Rajasthan
e-mail: mdtahakhan@gmail.com

Professor Dr. Shamimul Qamar
Department of Computer Networks and Communication Engg.
College of Computer Science
King Khalid University,Abha, Ksa
e-mail: drsqamar@rediffmail.com.com

Laurent F. Massin
Specialist Developer, C4 Advanced Solution, Abu Dhabi, UAE
e-mail: massinfossard@aim.com

***Abstract*—**This paper discuss about the important role of classification algorithms in clinical predictions , two case studies one for breast cancer and other for heart disease prediction with help of classification data mining techniques is presented in this paper. Online freely accessible data is used for the said case studies. Used data is publicly available data on internet consisting of 909 records for heart disease and 699 for breast cancer. C4.5 and the C5.0 Two well-known decision tree algorithms used to get the rules for predictions, and these rules used for improving the quality of an open source Pathology Management System based on Care2x.Performances of these algorithms are also compared.

This Paper will further discuss about the importance of open source software in healthcare as well as how a pathology management system can adopt Evidence Based Medicine (EBM).

EBM is a new and important approach which can greatly improve decision making in health care. EBM's task is to prevent, diagnose and medicate diseases using medical evidence [5].Clinical decisions must be based on scientific evidence that demonstrates effectiveness.
This paper is basically extension of our previous work ‘A Prototype of Cancer/Heart Disease Prediction Model Using Data Mining’.

Keywords- *Health care Prediction, data mining, EBM*

*****

I. INTRODUCTION

Healthcare is the world’s second largest and also the fastest growing service sector. A well-managed and low cost healthcare system is of great importance to a country where large population with diverse social, educational and economical background is to be served. Generally healthcare services are provided by Government. Most of the Government Hospitals provide services with very limited resources. Managing the resources optimally and efficiently is major challenge for healthcare sector, information and communication technology (ICT) can play a major role to overcome this challenge. As ICT based tools can be used for resource management, patient record keeping, sharing the information, faster processing of data, managing the people at hospital etc. The development of an exhaustive healthcare system involves complex issues like finance, performance, security, scalability, and adherence to standards. Further, open source software solutions can help the hospitals to achieve the required services at lower cost.

Accurate and error-free of diagnosis and treatment given to patients has been a major issue highlighted in medical service nowadays. Quality service in health care field implies diagnosing patients correctly and administering treatments that are effective [8].Hospitals can also minimize the cost of clinical tests by employing appropriate computer-based information and/or decision support systems. Most hospitals today use some sort of hospital information systems to manage their healthcare or patient data [7]. As we know computer based systems use to generate a huge amount of data which can be process to find out the hidden useful information, and these information can be used in clinical decision making.

The main goal of this research is: “How open source software can provide low cost systems for healthcare?” and “How we can process the data to get useful information to support decision making by healthcare practitioners?”.

World is looking for better treatment for deadliest diseases like cancer and heart disease. For a better treatment planning it is very important for clinician as well as patients to know the future holds of cancer/heart disease. A good prediction system for heart disease and cancer can be proved as a better tool for improving the efficiency of a hospital and clinicians. Now a days modern healthcare system rapidly accepting the data mining approaches mainly because the effectiveness of these approaches to classification and prediction systems has improved, particularly in relation to helping medical practitioners in their decision making. This type of research can play an important role in improving patient outcomes, cost reduction of medicine, and further advance clinical studies.

Case studies discussed in this paper use the publicly available dataset of breast cancer and heart disease, and use C4.5, C5.0 classification algorithms to predict about heart/cancer disease, analyse the results for generating some rules related cancer and heart diseases and the rules generated by these algorithms in an open source pathology system for further predictions.

The paper has been divided into following section;-section 2 discus the importance of data mining in healthcare context, section 3 discuss a brief description of classification algorithms c4.5 and c5.0,section 4 describes two case studies about cancer and heart disease prediction using data mining, section 5 is relevance of open source software in healthcare and section 6 is about open source pathology system and use of rules generated through case studies and section 7 is conclusion and future work.

II. WHY DATA MINING IN HEALTHCARE?

In the modern age of information and communication technology (ICT) healthcare industry generates large amounts of complex data about patients, hospitals resources, and disease diagnosis etc. This large amount of data is main motivation for researchers to mine useful information and knowledge which enables support for cost-saving and decision making for healthcare systems. Decision is always responsibility of clinicians, these increased volume of stored data provide additional source of knowledge for decision making with help of data mining. Extracted hidden information and knowledge provide better patient care and effective diagnostic capabilities. Data mining techniques can help answer several critical questions, such as [6]:

- Given the records of dialysis patients, what can be done to improve the treatment of these patients?
- Given the historical patient records on cancer, should the treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?

Data mining is the most important step in knowledge discovery process; data mining brings different tools together to find out hidden and unknown facts and information from a large amount of data that are difficult to find out manually or by traditional methods. These techniques and methods are based on statistical techniques, visualization, machine learning, etc.

Data mining algorithms try to fit a model closest to the characteristics of data under consideration. These Models can be descriptive or predictive [6].

Descriptive models are used to identify patterns in data, clustering, association rules, and visualization are some are the tasks of descriptive modeling.

Predictive models are used to make predictions, for example, to make a diagnosis of a particular disease. A patient may be subjected to particular treatment not because of his own history but because of results of treatment of other patients with similar symptoms. Predictive modeling consists some of the tasks like Classification, regression, and time series analysis. Classification is the one of the most important task of prediction modeling. A brief description of classification and C4.5,C5.0 algorithms is given under section 3.

Fig.1 depicts how data mining play an important role in modern clinical practice.

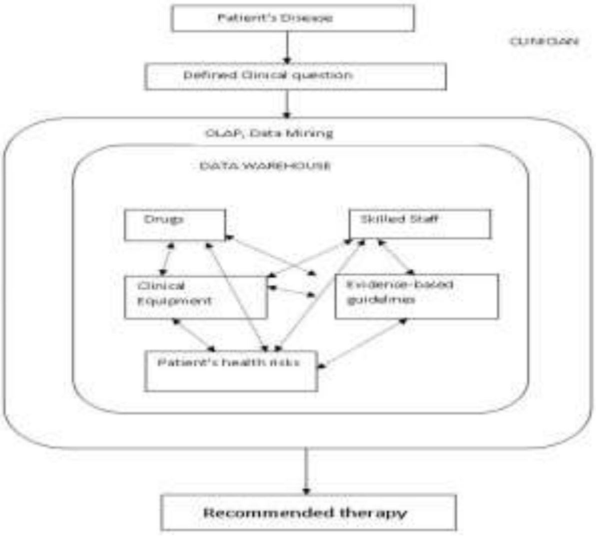


Figure1 use of data mining in better health delivery [4]

III. A BRIEF DESCRIPTION OF CLASSIFICATION TECHNIQUE

As per Fayyad et.al. (1996) Classification is finding models that analyse and classify a data item into several predefined classes Classification is a concept or process of finding a model which finds the class of unknown objects. Classification is the one of the most important data mining technique. Medical diagnosis is an important application of classification for example; diagnosis of new patients based on their symptoms by using the classification rules about diseases from known cases.

The classification problem is basically to define a
Function; f = D→ C where each ti € D is mapped to f (ti) belonging to some Cj [3].
Where:

1. D is a database of patients with tuples (x1, x2 … xn)
2. x1, x2 … xn are values of attributes A1,A2 … .An relevant to a particular disease.
3. C= {C1, C2 … Cn} is set of classes of disease depending on its severity.

Decision tree is a way of implementing the classification. Decision trees have become one of the most powerful and popular approaches in knowledge discovery and data mining. Decision tree is used as a predictive model. More descriptive names for such tree models are classification trees or regression trees. Decision trees need two kinds of data: Training and Testing data.

Training data, which are usually the bigger part of data, are used for constructing trees. The more training data collected, the higher the accuracy of the results. The other group of data, testing, is used to get the accuracy rate and misclassification rate of the decision tree.

A. Classification Algorithms C4.5 and C5.0

C4.5 algorithm: [14] c 4.5 algorithm is used for classification. C4.5 builds decision trees from a set of training data using the concept of information entropy in the same way as ID3, at each node of the tree; C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the smaller sublists. This algorithm has a few base cases.

1. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
2. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
3. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

i). Tree Generation:

Entropy and Gain is used in creating the tree.

$$I(P) = -\sum_{i=1}^{k} p_i * \log(p_i)$$

Where p_i is the proportion of instances in the dataset that take the i_{th} value of the target attribute.

Gain is:

$$Info(X,T) = \sum_{i}^{n} \frac{|T_i|}{|T|} Info(T_i)$$

Where i is a value of X, |Ti| is the subset of instances of T where X takes the value i, and |T| is the number of instances

ii). Pruning Trees.

Pruning algorithm is use to reduce error and avoiding the overfitting. Pruning a tree is the action to replace a whole subtree by a leaf. The replacement takes place if the expected error rate in the subtree is greater than in the single leaf. In our case study we will start by generating the whole (generally overfitted) classification tree and simplify it using pruning just after.

B. C5.0 Algorithm [Wikipedia]

Pseudo code for C5.0 and C4.5 algorithm is same. But there is some basic difference between these two algorithms. C4.5 made a number of improvements to ID3 and C5.0 offers a number of improvements on C4.5. Some of these improvements are:

1. Speed - C5.0 is significantly faster than C4.5 (several orders of magnitude)

2. Memory usage - C5.0 is more memory efficient than C4.5

3. Smaller decision trees - C5.0 gets similar results to C4.5 with considerably smaller decision trees.

4. Support for boosting - Boosting improves the trees and gives them more accuracy.

5. Weighting - C5.0 allows you to weight different attributes and misclassification types.

6. Winnowing - C5.0 automatically winnows the data to help reduce noise.

IV. Case Studies of Cancer and Heart Disease Prediction

This section will discuss about two case studies of heart disease and cancer prediction.

A. Breast Cancer Prediction Case Study

Breast cancer is one of the deadliest diseases in women. It is said to be second leading cause of cancer deaths in women today [19]. Breast cancer cases are rising at very high speed in India; it took the first spot leaving the cervical cancer at second. Recent data from Indian Council of Medical Research (ICMR) Show the seriousness of the situation. The documentation is mainly about rise of breast cancer in metros but it can be safely said that many cases in rural areas are still unnoticed. Reports say that one in 22 women in India is likely to suffer from breast cancer during her lifetime[12], while in America with one in eight being a victim of this deadly cancer. University of Wisconsin Hospitals, Madison (Dr. William H. Walberg) [16] is having dataset for breast cancer online .This online available is used for the breast cancer prediction case study.

i). Attributes of Breast Cancer data:

Total Cases:	**599**
Attribute	**Domain**
1. Sample code number	id number
2. Clump Thickness	1 – 10
3. Uniformity of Cell Size	1 – 10
4. Uniformity of Cell Shape	1 – 10
5. Marginal Adhesion	1 – 10
6. Single Epithelial Cell Size	1 – 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class:	(2 for benign, 4 for malignant)

ii). Specification of Attributes:

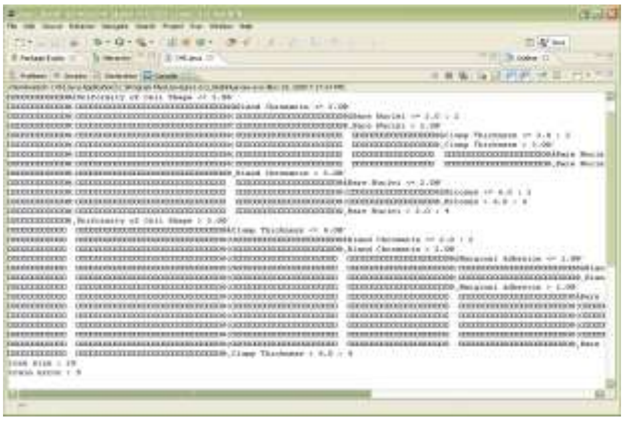
The target attribute:
Class

Sample code number:
ignore
Clump Thickness:
continuous
Uniformity of Cell Size:
continuous
Uniformity of Cell Shape:
continuous
Marginal Adhesion:
continuous
Single Epithelial Cell Size:
continuous
Bare Nuclei:
continuous
Bland Chromatin:
continuous
Normal Nucleoli:
continuous
Mitoses:
continuous

The target attribute is class which can have two values either 2(Benign) or 4(Malignant).Malignant is cancerous.
Malignant tumors can invade and destroy nearby tissue and spread to other parts of the body Benign is not cancerous. Benign tumors may grow larger but do not spread to other parts of the body. Value to class attribute is given 2 and 4 to avoid the conflict with the values of other attributes. There are several attributes mentioned above which can have value from1 to 10.C 4.5 and C5.0 Programs supports three type of files: Names files Provides names for classes, attributes, and attribute values, Data file describe the training cases for generating the decision tree and/or and test file used to evaluate the produced classifier.

iii). Decision Tree and Rules Generated:

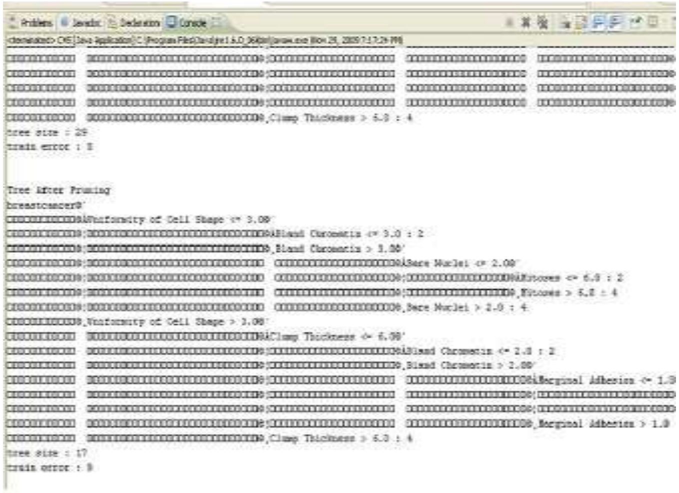
Following Fig.2 depicts the tree generated using c4.5 algorithm. Tree size is 29 with 5 train error.5 train errors means after running the 400 records on C4.5 there are five cases where error was noted down.



```
Uniformity of Cell Shape <= 3.0
Bland Chromatin <= 3.0
[illegible]
Clump Thickness > 6.0 : 4
tree size : 29
train error : 5
```

Fig.2.Tree Generated before pruning using c4.5

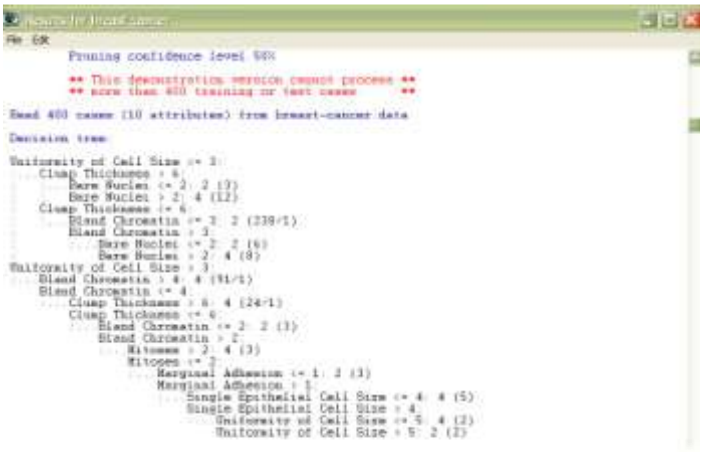
As pruning a tree is the action to replace a whole subtree by a leaf which reduces the size of tree. Following Fig.3 depicts tree after pruning. Tree size is 17.



```
Clump Thickness > 6.0 : 4
tree size : 29
train error : 5


Tree After Pruning
breastcancer
Uniformity of Cell Shape <= 3.0
Bland Chromatin <= 3.0 : 2
Bland Chromatin > 3.0
Bare Nuclei <= 2.0
Mitoses <= 6.0 : 2
Mitoses > 6.0 : 4
Bare Nuclei > 2.0 : 4
Uniformity of Cell Shape > 3.0
Clump Thickness <= 6.0
Bland Chromatin <= 2.0 : 2
Bland Chromatin > 2.0
Marginal Adhesion <= 1.0
Marginal Adhesion > 1.0
Clump Thickness > 6.0 : 4
tree size : 17
train error : 9
```

Fig.3.Tree Generated after pruning using c4.5

Fig.4 shows the tree generated after running C5.0, which reads 400 cases with 10 attributes.



```
        Pruning confidence level 99%

        ** This demonstration version cannot process **
        ** more than 400 training or test cases      **

Read 400 cases (10 attributes) from breast-cancer data

Decision tree:

Uniformity of Cell Size <= 3:
:...Clump Thickness > 6:
:   :...Bare Nuclei <= 2: 2 (3)
:   :   Bare Nuclei > 2: 4 (12)
:   Clump Thickness <= 6:
:   :...Bland Chromatin <= 3: 2 (238/1)
:       Bland Chromatin > 3:
:       :...Bare Nuclei <= 2: 2 (6)
:           Bare Nuclei > 2: 4 (8)
Uniformity of Cell Size > 3:
:...Bland Chromatin > 4: 4 (91/1)
    Bland Chromatin <= 4:
    :...Clump Thickness > 6: 4 (24/1)
        Clump Thickness <= 6:
        :...Bland Chromatin <= 2: 2 (3)
            Bland Chromatin > 2:
            :...Mitoses > 2: 4 (3)
                Mitoses <= 2:
                :...Marginal Adhesion <= 1: 2 (3)
                    Marginal Adhesion > 1:
                    :...Single Epithelial Cell Size <= 4: 4 (5)
                        Single Epithelial Cell Size > 4:
                        :...Uniformity of Cell Size <= 5: 4 (2)
                            Uniformity of Cell Size > 5: 2 (2)
```

Fig.5.Rules Generated using c5.0

B. *Heart Disease Prediction*

Heart diseases are also one of the most deadliest diseases. Because of the life style now a days heart disease are becoming the very common. Prior knowledge of chances of getting a heart disease is very helpful for patient as well as clinicians for planning a better and effective treatment. This case is all about prediction of heart disease using the heart disease data set. The algorithms which are used again are C5.0 and C4.5. The purpose is to predict the presence or absence of heart disease given the results of various medical tests carried out on a patient.

We have used a total of 909 records with 75 medical attributes. This dataset is taken from Cleveland Heart Disease database [14].We have split this record into two categories: one is training dataset (455 records) and second is testing dataset (454 records). The records for each category are selected randomly. “Diagnosis” attribute is the target predictable attribute. Value “1” of this attribute for patients with heart disease and value “0” for patients with no heart disease. “PatientID” is used as the key; the rest are input attributes. It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

```
 i). Attribute Information:
```

----

1. Age (age in years)
2. Sex (1 = male; 0 = female)
3. Chest pain type (4 values)
   - -- Value 1: typical angina
   - -- Value 2: atypical angina
   - -- Value 3: non-anginal pain
   - -- Value 4: asymptomatic
4. Resting blood pressure
5. Serum cholesterol in mg/dl
6. Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. Resting electrocardiography results (values 0, 1, 2)
   - -- Value 0: normal
   - -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
   - -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. Maximum heart rate achieved
9. Exercise induced angina (1 = yes; 0 = no)
10. Old peak = ST depression induced by exercise relative to rest
11. The slope of the peak exercise ST segment
    - -- Value 1: upsloping
    - -- Value 2: flat
    - -- Value 3: downsloping
12. Number of major vessels (0-3) colored by flourosopy
13. Thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

ATTRIBUTES TYPES

Real: 1, 4,5,8,10,12
Ordered: 11,
Binary: 2, 6, 9
Nominal: 7,3,13

Variable to be predicted

Absence (1) or presence (2) of heart disease

ii). Decision Tree Rules Generated By C5.0

```
See5 [Release 2.06]      Sat Nov 21
          19:36:52 2013

Read 150 cases (13 attributes) from
          heartdisease.data
```

DECISION TREE:

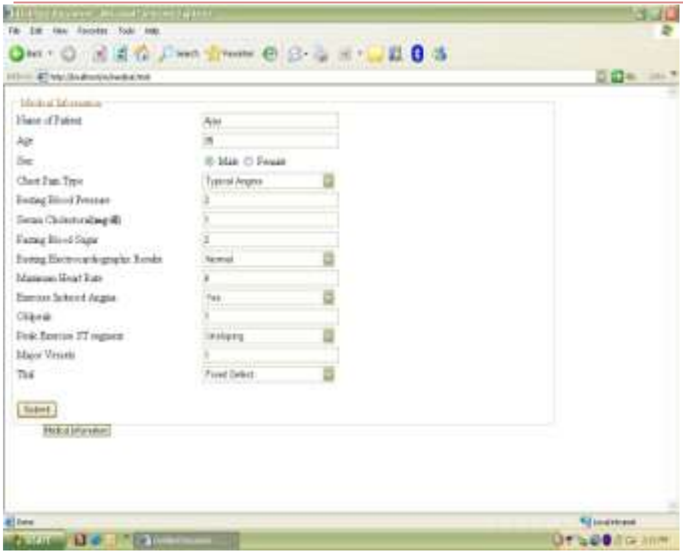
```
                        Thal > 6:
                        :...ChestPain  >
 3: 2 (32/2)
            :   ChestPain <= 3:
     :    :...STSlope <= 1: 1 (8/2)
     :        STSlope > 1: 2 (12/3)
                Thal <= 6:
          :...OldPeak > 2.8: 2 (6)
                OldPeak <= 2.8:
          :...ChestPain <= 3: 1 (60/6)
                   ChestPain > 3:
             :...Vessels <= 0: 1 (23/6)
                 Vessels > 0: 2 (9/1)
```

RULES:

```
Rule 1: (60/6, lift 1.6)
         ChestPain <= 3
         OldPeak <= 2.8
           Thal <= 6
      ->  class 1  [0.887]

Rule 2: (51/5, lift 1.6)
         ChestPain <= 3
          STSlope <= 1
      ->  class 1  [0.887]

Rule 3: (65/9, lift 1.5)
         OldPeak <= 2.8
          Vessels <= 0
           Thal <= 6
      ->  class 1  [0.851]

Rule 4: (27/1, lift 2.1)
          ChestPain > 3
           Vessels > 0
      ->  class 2  [0.931]

Rule 5: (32/2, lift 2.0)
          ChestPain > 3
            Thal > 6
      ->  class 2  [0.912]

Rule 6: (31/3, lift 2.0)
           STSlope > 1
            Thal > 6
      ->  class 2  [0.879]


 Rule 7: (6, lift 2.0)
          OldPeak > 2.8
           Thal <= 6
      ->  class 2  [0.875]


      Default class: 1

Evaluation on training data (150 cases):

                   Rules
             ----------------
               No      Errors
              7   20(13.3%)   <<


    (a)    (b)    <-classified as
           ----   ----
      77     6     (a): class 1

      14    53     (b): class 2
```

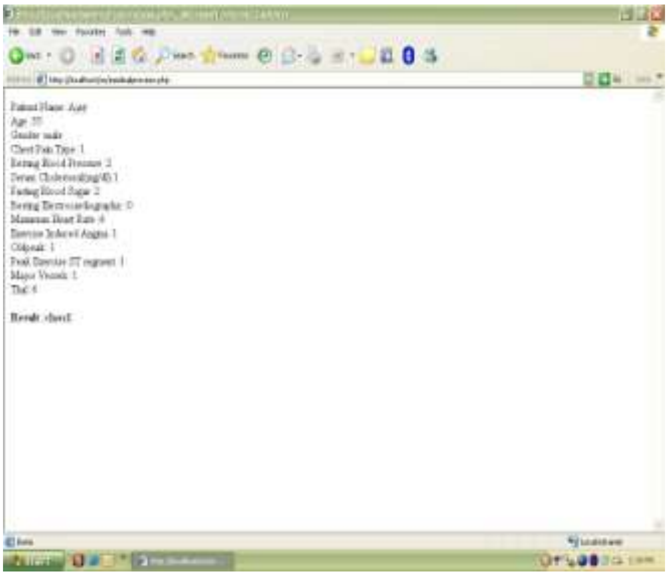
C. *Working Prediction Model for Cancer*

As part of our project we have designed a working model for cancer/heart disease prediction. This model will predict the breast cancer's or heart disease class based on the rules created by C4.5 and C5.0 algorithms.Fig.6 shows the interface for input, which take Medical profiles of a patient such as age, sex, blood pressure and blood sugar etc as input and it can predict about presence or absence of cancer/heart disease.



Name of Patient	Ajay
Age	35
Sex	● Male ○ Female
Chest Pain Type	Typical Angina
Resting Blood Pressure	2
Serum Cholestoral(mg/dl)	1
Fasting Blood Sugar	2
Resting Electrocardiographic Results	Normal
Maximum Heart Rate	4
Exercise Induced Angina	Yes
Oldpeak	1
Peak Exercise ST segment	Unsloping
Major Vessels	1
Thal	Fixed Defect

Fig7. Bellow shows a particular case belonging to class1 for heart disease.



Patient Name: Ajay
Age: 37
Gender: male
Chest Pain Type: 1
Resting Blood Pressure: 2
Serum Cholestoral(mg/dl): 1
Fasting Blood Sugar: 2
Resting Electrocardiographic: 0
Maximum Heart Rate: 4
Exercise Induced Angina: 1
Oldpeak: 1
Peak Exercise ST segment: 1
Major Vessels: 1
Thal: 6

Result: class1

Fig.7.Interface for output

V. Relevance of Of Open Source Software in Healthcare

Providing quality services at affordable prices is major challenge healthcare organizations (hospitals and medical centres).In term of ICT infrastructure hardware and software are capital goods for an organization. Less price of a software means ICT is available at lower cost. This helps an organization to add to its resources and improves its process.

Open Source software is an important and growing class of software. Open Source software is distinguished not by programming language, operating environment, nor application domain, but rather by the license(s) that governs the use, distribution, and, most importantly, the rights to access and modify the software's source code [21]. The philosophy of open source permits users to use, change, and improve the software, and to redistribute it in modified or unmodified forms. Together, software source code, licensing, and community have dramatically changed many conventional assumptions about software and the software industry itself. Acceptability of open source software is increasing day by day. Some of the reasons for using open source software include low total cost of Ownership, lack of software piracy issues, and availability of source code leading to high degree of customizability and scalability and extensive support freely available on Internet. When the source code of a program is available anyone can contribute by improving the code, adding new features, correcting errors, etc.

Healthcare is one of the important sectors for the economy of any developing country; if we get low cost ICT solutions for healthcare it is very beneficial for economical growth. Open source software have potential to be a key player for low cost quality healthcare delivery. Care2x, OpenVista, OpenEMR are some of free and open source healthcare software worldwide used.

VI. Pathology Management System based on Care2x

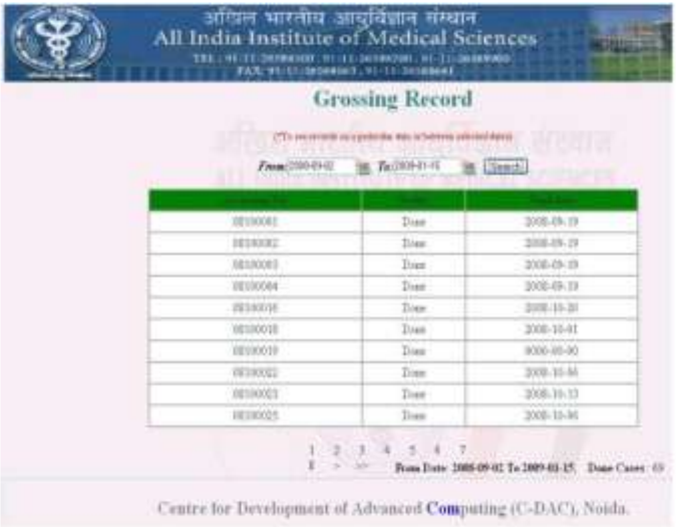
We have developed one Advance Pathology Management System based on Care2x for AIIMS Pathology.AIIMS is one of the premier Govt. hospitals in India with 1600 beds. It handles near about 1500000-1600000 in-patient and outpatient yearly. AIIMS Pathology reports 400-500 cases per day. To improve the work process of AIIMS Pathology there is a requirement of an advance pathology management system.

And maintaining the low cost was our primary goal. For this purpose we opted one existing open source software Care2x to customise it as per our requirement . CARE2X is an open source Web based Integrated Healthcare Environment (IHE)[22] under GNU/GPL. The project was started in May 2002 Until today the development team has grown to over 100 members from over 20 countries. Its source code is freely distributed and available to the general public.

CARE2X [22] HIS is built upon other open-source projects: the Apache web server from the Apache Foundation the script language PHP [23] and the relational database management system mySQL [24]. CARE2X is modular and highly scalable so it is very easy to scale this application as per requirements. CARE2X is currently composed of four major components. Each of these components can also function individually. These components are HIS - Hospital/Health service Information System, PM - Practice (GP) management, CDS - Central Data Server, HXP - Health Xchange Protocol [22].

This advanced pathology management system is providing all features like Grossing, Sectioning, Reporting and Sample tracking with decision support.

Rules generated are used in this system to help the clinician in decision making.



अखिल भारतीय आयुर्विज्ञान संस्थान
All India Institute of Medical Sciences

Grossing Record

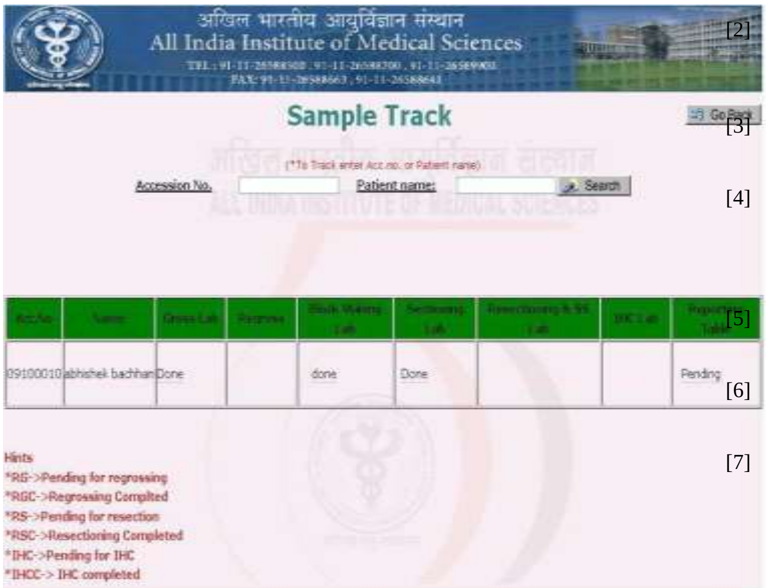
From: 2008-09-02 To: 2009-01-15 Search

[illegible]	[illegible]	[illegible]
08100001	Done	2008-09-19
08100002	Done	2008-09-19
08100003	Done	2008-09-19
08100004	Done	2008-09-19
08100016	Done	2008-10-20
08100018	Done	2008-10-01
08100019	Done	0000-00-00
08100022	Done	2008-10-06
08100023	Done	2008-10-13
08100025	Done	2008-10-06

1 2 3 4 5 6 7
1 > >> From Date: 2008-09-02 To 2009-01-15 Done Cases: 69

Centre for Development of Advanced Computing (C-DAC), Noida.

Fig.8.Grossing option in APMS



अखिल भारतीय आयुर्विज्ञान संस्थान
All India Institute of Medical Sciences
TEL.: 91-11-26588500, 91-11-26588700, 91-11-26589900
FAX: 91-11-26588663, 91-11-26588641

Sample Track

(*To Track enter Acc.no. or Patient name)

Accession No. Patient name: Search

Acc.No	Name	Gross Lab	Regross	Block Making Lab	Sectioning Lab	Resectioning & SS Lab	IHC Lab	Reporting Table
09100010	abhishek bachhan	Done		done	Done			Pending

Hints
*RG->Pending for regrossing
*RGC->Regrossing Complted
*RS->Pending for resection
*RSC->Resectioning Completed
*IHC->Pending for IHC
*IHCC-> IHC completed

Fig.9.Sample tracking in APMS

At the time of reporting system prompt the suggestions based on the rules. Based on the sample symptoms suggestion populated. This system is a step towards the evidence based medicine.

VII. Conclusions and Future Work

If we talk about performance of these two algorithms, C5.0 handles missing values easily but C4.5 shows some errors due to missing values. Over running the dataset of breast cancer of 400 records C4.5 shows 5 train error whereas C5.0 show only 3 train errors. C5.0 produces rules in a very easy readable form but C4.5 generates the rule set in the form of a decision tree.

Data mining techniques play an important role in finding patterns and extracting knowledge from large volume of data. It is very helpful to provide better patient care and effective diagnostic capabilities. Evidence Based Medicine (EBM) is a new direction in modern healthcare.EBM is as an important approach to make clinical decisions about the care of individual patients. This decision about patient is based on the best available Evidence. Its task is to prevent, diagnose and medicate diseases using medical evidence. It is all about providing best evidence, at right time in right manner to the clinician. External evidence-based knowledge cannot be applied directly to the patient without adjusting it to the patient's health condition. If the rules generated by this system is approved by medical experts that can be used as evidence for further use.

CARE2X is flexible generic multi-language open-source project. CARE2X is a very feature rich HIS, fully configurable for any clinical structure. After customization, it has the potential to become functional software to support workflows of Indian hospital. Efforts were made to explore the possibility of providing a low cost solution to Indian hospitals.

References

[1] Jaree Thongkam, Guandong Xu, Yanchun Zhang and Fuchun Huang 'Breast Cancer Survivability via AdaBoost Algorithms'*HDKM,2008,wollongon,australia.*

[2] Diana Dumitru 'Prediction of recurrent events in breast cancer using the Naive Bayesian classification' Annals of University of Craiova, *Math. Comp. Sci. Ser.*Volume *36(2), 2009, Pages 92-96 ISSN: 1223-6934.*

[3] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 194-200, 2006.

[4] Nevena Stolba and A Min Tjoa "The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making" December 24, 2005.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", J Healthcare Information Managment. 16(4), 50-55, 2002.

[6] Siri Krishan Wasan, Vasudha Bhatnagar and Harleen Kaur*The impact of data mining techniques on medical diagnostics" *Data Science Journal, Volume 5, 19 October 2006".*

[7] Herbert Diamond, Michael P. Johnson, Rema Padman, Kai Zheng, "Clinical Reminder System: A Relational Database Application for Evidence-Based Medicine Practice " INFORMSSpring National Conference, Salt Lake City, Utah-April 26, 2004.D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[8] Sellappan Palaniappan , Rafiah Awang "Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques" Proceedings of iiWAS2007.

[9] Infectious Disease Informatics and, outbreak detection,Daniel Zeng1, Hsinchun Chen, Cecil Lynch, Millicent Eidson, and Ivan Gotham.

[10] AMPATH Medical Record System AMRS): Collaborating toward An EMR for Developing Countries Burke W. Mamlin, M.D. and Paul G. Biondich, M.D., M.S.Regenstrief Institute, Inc. and Indiana University School of Medicine, Indianapolis, IN

[11] Global Epidemiological Outbreak Surveillance System Architecture:Ricardo Jorge Santos(1) and Jorge Bernardino CISUC – Centre of Informatics and Systems of the University of Coimbra – University of Coimbra)ISEC – *Engineering Institute of Coimbra – Polytechnic Institute of Coimbra portugal*

[12] http://www.medindia.net/news/view_news_main.asp?x=7279

[13] Managing Diagnostic Process Data Using Semantic Web,Vili Podgorelec, Luka Pavlic Institute of Informatics, FERI, University of Maribor, Slovenia.Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07) 0-7695-2905-4/07

[14] http://en.wikipedia.org/wiki/C4.5_algorithm.

[15] ARIHITO ENDO, TAKEO SHIBATA, HIROSHI TANAKA 'Comparison of Seven Algorithms to Predict Breast Cancer Survival' *Biomedical Soft Computing and Human Sciences,* Vol.13, No.2, pp.11-16 (2008).

[16] http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Prognostic) (Breast cancer dataset).

[17] http://archive.ics.uci.edu/ml/datasets/Heart+Disease (Heart Disease dataset).

[18] DMS Tutorial: http://dms.irb.hr/tutorial/tut_dtrees.php.

[19] ICMR Bulletin: http://www.icmr.nic.in/bufebruary03.pdf.

[20] Tipawan Silwattananusarn and Dr. KulthidaTuamsuk ' Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012' *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012.*

[21] Michael Tiemann President Open Source Initiative Vice President Open Source Affairs, Red Hat November 1, 2009 'How Open Source Software Can Save the ICT Industry One Trillion Dollars per Year'.

[22] CARE2X; an Open Source Project. http://www.CARE2X.org .

[23] PHP An Open Source widely used language for web development, http://www.php.org .

[24] MySql Largest Open Source Database used by many renowne leading organizations http://www.mysql.com.