# Clinical Data Mining – An Approach for Identification of Refractive Errors

D.V. Chandra Shekar and V.Sesha Srinivas

*Abstract*— **Information Technology offer health care industry significant potential to improve productivity and quality of patient care. The area of Data mining in health care is growing rapidly because of strong need for analyzing the vast amount of clinical data bases stored in hospitals. Hospitals are adopting data mining technique to identify hidden information from clinical databases. This information could provide new medical knowledge. In this paper we provide an over view of Data mining classification technique . Our study on Refractive errors in schools going children identifies the need of mining the clinical data. The database includes nearly 1,87,513 student records. This huge data was analyzed by decision tree algorithm and finally we found that Myopia was most common refractive error in both male and female students.**

*Index Terms*— **Reflex errors, classification, pediatric ophthalmology, Regression.**

## I. Introduction

The Internet and web based technologies have a significant impact on the health care industry. More physicians are now adopting e-health care tools to provide better patient care. Major improvements have been witnessed in key health care metrics as reduction of medical errors rates, lowering laboratory costs and administrative expenses.

A few popular IT application tools fore physicians health care industry include computer based patient record, e-prescribing online communication through e-mail and chats between physicians and patients , web based order entry tools , e-medicine , telemedicine, cyber surgery, micro-robotic surgery and 3D image modeling.

Manuscript received January 15, 2008. Reviewed and accepted on January 17,2007

F. Author is Assistant Professor, Department of Computer Science , TJPS College (P.G Courses ), Guntur, Andhra Pradesh–522006,INDIA,
Phone : 9440454033;e-mail : chand.info@gmail.com
S. Author is Lecturer , Department of MCA, RVR & JC Engg College , Guntur AndhraPradesh-522006, INDIA
e-mail: vangipuramseshu@yahoo.com

Not with standing the applications of IT in the health care industry a few hospitals worldwide are in the advanced stage of developing clinical data repositories over and above their basic documentation systems. Very few hospitals are fully capitalizing on the benefits offered by physicians order entry, clinical decision support system and rule engines.

It has been estimated that the amount of information in the world doubles every 20 months and the size of the number of databases increasing even.

Faster (Dilly 1995). The use of traditional verification based approaches analysis is difficult when the data is massive, highly dimensional, distributed and uncertain.

While technological advancements in the form of computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable, few tools exist to evaluate and analyze this clinical data after it has been captured and stored. Evaluation of stored clinical data may lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management. Techniques are needed to search large quantities of clinical data for these patterns and relationships. Past efforts in this area have been limited primarily to epidemiological studies on administrative and claims databases. These data source lack the richness of information that is available in database comprised of actual clinical data. Data needs to be classified in order to make analysis easier & more meaningful. Classification of data can help in revealing patterns of the disease hitherto hidden or unknown.

Health care facilities have at their disposal of vast amount of data. Though analysis of available data on a given problem can lead to more efficient decision-making. The challenge is to extract relevant knowledge from this data and act upon it in a timely manner. Data mining is one of method to classify and analyze clinical data [2]. In this study we propose the

introduction of a recently developed this methodology to clinical databases of Eye hospital.

## 2. STATEMENT OF THE PROBLEM

Refractive errors are the most common cause of visual impairment & the second most common cause for treatable blindness in the world. Uncorrected visual errors have severe social & economic effects on individuals & communities, restricting educational & employment opportunities of otherwise healthy people. Studies have shown that refractive error in children causes up to 62.5% of blindness (<6/60 in the better eye) in Chile, 22% in Nepal,77% in urban India, and 75% in China. For visual impairment in children (<6/12 in the better eye), refractive error is responsible for 55% in Chile, 86% in Nepal, 93% in China, 70% in rural India, and 83% in urban India. What is also disturbing is the amount of this refractive error that is uncorrected on presentation – 46% in Chile , 92% in Nepal , 58% in china , 86% in rural India [6]. Data from the Sankara –ORBIS pediatric opthalmology school screening programmed were entered in a computer database. this involved a total of 1,87,513 students being screened of which2,837 were found to have defective vision . this huge database needs to be analyzed for doctors to get an understanding of the problem at hand . proper analysis of this data reveal patterns of refractive errors particular to age , sex or location . this would in turn help in understanding the nature of refractive errors in the school children in a particular area or age group enabling proper enabling proper allocation of resources .

The main objective of the study was to classify the sample data according to age , gender & type of refractive error using the data mining technique decision tree that was constructed using ID3alogorithm.

## 3. DATA MINING AND CLASSIFICATION TECHNIQUES

Data mining technologies were developed in order to extract meaningful information from these large stores of data, taking a place in the overall field of knowledge Discovery methods whereby new information is derived from a combination of previous knowledge (Geobel & Grunewald 1999). The subset of knowledge Discovery that involved an expert, in order to make use of their domain expertise in determining how to discover new knowledge, has been labeled as Knowledge Acquisition (Gaines, B.R. 1993). Automated Machine Learning techniques have therefore become the main focus of research into Knowledge Acquisition (Grefenstette, Ramsey&Schultz 1990; Hong et al.2000; Sester 2000). Classification Maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examination of data. Classification algorithms require that the classes be defined based on data attribute values.

### 3.1 DECISION TREE

Decision tree are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that in contrast to neural networks, decision trees represent rules. Rules can readily be expressed so that humans can understand them. There are variety of algorithms for building decision trees that share the desirable quality of interpretability. Decision tree induction is typical inductive approach to learn knowledge on classification. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance.

Decision tree is a classifier in the form of a tree structure ( Figure 1), where each node is either:

- a leaf node - indicates the
- value of the target attribute (class) of examples, or
- a decision node - specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test.

The key requirements to do mining with decision tree are:

1: Attribute –Value description – Object properties

2: Predefined classes – category to which examples are assigned

3. Discrete classes- A case does not belongs to a particular class

4. Sufficient Data – Usually hundred and thousands of training cases

## 4. CONSTRUCTING DECISION TREE USING ID3 ALGORITHM

Decision Trees are an example of a global model. By contrast, to k-Nearest Neighbor algorithm and Support Vector Machines, decision trees such as ID3
(interactive dichotomizer) and C4.5 typically select only a subset of the attributes to form the
hypothesis – a decision tree. ID3 was restricted to attributes with discrete values for not only the output variable, but also the input variables.

*Function ID3*

*Input : (R : A set of non target attributes*

*C: The target attributes*

*S: A training set –returns a decision tree*

*Begin*

*If S is empty , return a single node with value Failure;*

*If S consists of records all with the same value for the target attributes, return a single leaf node with that value.*

*If R is empty then return a single node with the value of the most frequent of the values of the target attribute that are found in records of S.*

*Let A be the attribute with largest Gain(A,S) among attributes in R;*

*Let { aj|j=1,2…m} be the values of attributes A;*

*Let {Sj|j=1,2,….m} be the subsets of S consisting respectively of records with value aj for A;*

*Return a tree with root labeled A and arcs labeled a1,a2,….am going respectively to the trees(ID3{R-A},C,S1,ID3(R-{A},C,S2),……..*

*ID3(R-{A}, C,Sm};*

*Recusrively apply ID3 to subsets {Sj|j=1,2,…m}*

*Until they are empty*

*End*

ID3 searches through the attributes of training instances and extracts the attributes that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 stops; Other wise it recursively operates on the m(where m = no. of possible values of an attribute) portioned subsets to get their best attribute. For each of the possible discrete values of this attribute a descendant node is created and the relevant training examples assigned to the correct node. The algorithm uses a greedy search that is it picks the best attribute and never looks back to consider earlier choices.

The central focus of ID3 decision tree growing algorithm is selecting which attribute to test at each node in the tree. For the selection of the attribute the concept of Entropy is used.

## 5. RELATED WORK IN CLINICAL DATABASES

This section deals with data mining & related researches, and focuses on current research work on data mining. Margaret R. Kraft et al [4] have explored the process of knowledge generation from the Veteran's Administration healthcare information system. This inquiry is concerned with predicting the length of stay of a subset of the total patient population, specifically those with spinal cord injuries (SCI).

Alapontetal [5] have presented a research project, which is directed to the partial automation of Data Mining (DM) in hospital information systems (HIS).

Bharat Rao [12] note that key clinical data in hospital records is typically recorded in unstructured form as free text and images, and most structured clinical information is poorly organized.

Tristan Ronald Ling [7] concludes that the Exposed MCRDR (Multiple Classification Ripple-Down Rules) method is a valuable knowledge Discovery approach, even for domains that require extensive background knowledge, have large volumes of difficult to interpret cases, and have complex and specific target knowledge.

Susan Jensen [10] has analyzed data relating to patient information and medical exams connected with thrombosis attacks using SPSS Clementine data mining workbench.

Madigan EA et al [9] employed the data mining approach CART (Classification and Regression Trees) to determine the drivers of home healthcare service outcomes (discharge destination and length of stay)and examine the applicability of induction through data mining to home healthcare data.

The Andhra Pradesh Eye Disease Study (APEDS) was a population based cross sectional study of all ages representative of the urban-rural and socioeconomic distribution of the population of the Indian state of Andhra Pradesh. The ethics committee of the LV PRASAD EYE INSTITUTE, Hyderabad, India, approved this study. APEDS was conducted from October 1996 to February 2000.

## 6. OUR APPROACH TO CLASSIFY CLINICAL DATABASE

The present data analysis was carried out on data collected under the Sankara-ORBIS pediatric ophthalmology program. This program is an EYE screening programme for school children in GUNTUR, District of Andhra Pradesh, India. An estimated 1.5. million children are blind ,world wide. According to a study done in South India 5.1% school Children are visually impaired. Early detection of visual impairment will prevent a child from going into a stage of AMBLYOPIA or SQUINT. Uncorrected refractive errors are an important cause of vision impairment. Total 1,87,513 students studying in 975 schools of Guntur Distict in Andhra Pradesh are Screened.

### 6.1 CLASSIFICATION PROCESS

Training Data Set:

Children in the age group of 5-16 years in the schools in the jurisdiction of EYE hospital.

First step of the the algorithms is to determine the Root node. This is done by calculating the ENTROPY of the attributes. The main attributes are age, gender & refractive error.

The Entropy is calculated using the formula

$$ENTROPy(P)=[p_1(\log(p_1) + p_2(\log(p_2) +\ldots\ldots p_n \log(p_n)\ ]$$

Each step in ID3 chooses a state that orders splitting the most. A database state is completely ordered if all tuples in it are in the same class. ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as the difference between how much information is needed to make a correct classification before the split versus how much information is needed after the split. The entropies of split data are weighted by the fraction of the data set being placed in that division.

Table 1: Information Gain for the attributes

| Attribute | Information Gain |
|---|---|
| Age | 0.72 |
| Gender | 0.43 |
| R-errors | 0.41 |

Hence the largest Information gain is provided by attribute age. By choosing age as first attribute to split according to the highest information gain we find that 7% in the age group of 5-8 years, 28% in the age group of 9-12 years and 65% in the age group of 13-16 yrs. After choosing and gender as the second splitting attribute according to information gain we find that we got the below classified data.

Table 2: Age wise classification of data

| AGE (years) | FEMALE % | MALE % |
|---|---|---|
| 5-8 | 7.84 | 6.12 |
| 9-12 | 19.6 | 36.74 |
| 13-16 | 72.5 | 57.14 |

Repeating the above steps to choosing Refractive errors as splitting attribute the entropy is calculated and information gain now is 0.1087.

Table 3: Age wise Astigmatism

| AGE (years) | Myopia | Hyperopia | Astigmatism (%) |
|---|---|---|---|
| 5-8 | 0 | 0 | 100 |
| 9-12 | 61.11 | 16.67 | 22.22 |
| 13-16 | 64.29 | 7.14 | 28.57 |

The complete decision tree is shown in the below figure.

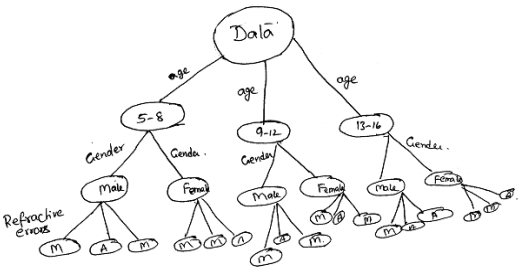


Figure 1: Complete Decision Tree

## 7. CONCLUSION

Classifying the given data, it was found that there was no significant difference in incidence of refractive errors between males (49%)& females (51%) . Further it was found that myopia was the most common refractive error in both males& females. In the 5-8 years age group , myopia was the most common refractive error among females whereas in males astigmatism was more common . In the 9-12 &13-16 years age group , myopia was again the most common refractive error .

## 8. FUTURE SCOPE

In this study, 1,87,513students data was analyzed . To get a better picture of this classification, visualization techniques may be used Further, more attributes could be used for analysis like for example, geographic area, school of the students, urban or rural areas etc.

## 9. REFERENCES

[1] Bary Robson and Richand Mushlin – Clinical and pharmacogenomic data mining.

[2] Chae Ym – Data mining approach topolicy –analysis in health insurance domain 2001 Jul 62(2-3):103-11

[3] Bary Robson and Richand Mushlin- A simple method for combination of information from Associations and Multivarience to Facilitate Analysis Decision and Design in Clinical Research.

[4] Margaret R.Kroft et al Data mining in Health care Information Systems – Proceddings of 36 the Hawai International Conference on System Services(HICSS'03)

[5] J Alapont , A.Bella Sanjan et al –Specialized Tools for Automating Data mining for Hospital Management

[6] Brien A Holden Serge REsmikoff –The role of optometry Invision 2020 –Community Eye Health vol15 Issue 43, 2003

[7] Tristan Ronal Ling , Nov -2006 – AnIncremental Learning Method for Data mining from Large Databases .

[8] H.Hamilton. E. Gurak, L. Findlater W. Olive -Overview of Decision Trees

[9] Madigan EA Curet OL –A data mining approach in home healthcare – outcomes and services use , 2006 Feb24,6:18

[10] Susan Jensen – Mining Medical Data for predictive and sequential patterns PKDD2001, SPSS Uk ltd.

[11] Kamber M, Winstone L and Hanj – Generation and decision Tree Induction – Efficient classification in Data mining

[12] R. Bharat Rao , Sathyakama Sandilya et al – Clinical & Financial outcomes Analysis with existing hospital patient record.