"A SURVEY ON DATA MINING TECHNIQUES FOR CLINICAL PREDICTION"

Rekha Bhor, Dhanashree Chobhe,Shital Naik.

Guide By,

Prof. J. U. Lagad

Computer Department

SCSMCOE, A. Nagar

Prof. D. B. Zine (H. O. D.)

Computer Department

SCSMCOE, A. Nagar

Abstract:

This project is consisting of a deep overview of the applications in various techniques such as data mining techniques, medical, research, and educational aspects of Clinical Predictions. Due to the regulations and due to the availability of computer system in medical and health care system areas there is a large amount of data is becoming available. Such a large amount of data cannot be handled by humans in a short period of time to make diagnosis, prognosis and schedule of all the treatments as well as practitioners are expected to use all this data in their work.

A main motive or main goal of this paper is to elaborate the data mining techniques in clinical and health care system to evaluate the accurate decisions about particular disease.

The paper also provides the brief overview of clinical data mining techniques can improve various aspects of clinical predictions. Due to the modern world ,a wrong diagnosis by the hospital lead to earn a bad name and loosing reputation. The use of this paper is to implement a cost effective treatment using data mining technologies for providing database decision support system.

In this paper by using various data mining techniques the support is made to assist in the diagnosis of the disease .

Keywords: Naive Bayes, medical research, data mining, decision making.

Introduction:

Among all most information intensive industries is the Healthcare industry .

In medical and health care system information and data keeps growing on day-to-day life.

Unfortunately, hospital system may generate approximately four terabytes of data in a year.

It becomes critical for quality healthcare system to use such a large amount of data to extract the useful information. Data Mining is the important step,which results in the discovery of hidden but it is very useful knowledge from the database.

A formal definition of Knowledge discovery in databases is given as follows:

Data mining is the extraction of hidden predictive and also extraction of implicit previously unknown and potentially useful information from large databases.

Data mining technology also provides a user-oriented approach to novel and hidden patterns in the data.

"Clinical Prediction is the computerization of medical information to support and optimize:

(1) administration of health services;

ISSN: 2278 – 7798

(2) clinical care;
(3) medical research; and
(4) training.

Data mining techniques predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven and accurate decisions.Data mining techniques offers prospective analysis to provide the past events by the retrospective tools of decision support system.Data mining tools can answer business questions that traditionally were too time consuming to resolve.They inding predictive information that expert may miss because it lies outside their expectation.It is the application to be used for computing and communication technologies to optimize health information processing by collection,effective retrieval (includes time and place),analysis and decision support for administration,clinicians,researchers and education of medicines.Computerized information retrieval may help in quality decision making and to avoid human error also help in accurate decision making.Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable.Hospitals must also minimize the cost of clinical tests. They can gain these results by appropriate computer-based information and/or decision support systems.Health care data is bulky.It includes patient centric data, resource management data and transformed data.Clinical care system must have ability to analyze data.Patients records like treatment record of millions of patients can be stored and computerized.Data mining techniques may help in answering several important and critical questions related to health care.

Research Elaboration:

Although human decision-making is often optimal, but it is poor when there are large amounts of data to be classified.Due to the humans stress and immense work efficiency and accuracy of decision will decreases. Typical problems is that data mining addresses are how to classify data, cluster data, find associations between data items, and perform time series analysis. Various data mining techniques have been invented for each type of problem.Each problem requires data mining techniques to analyze large quantities of data. Today diagnosing patients correctly and administering effective treatments have be challenge.Poor clinical decision may be tends to patients death and which cannot be afforded by the hospital as it looses its reputation.The cost to treat a patient with a heart problem is quite high and not affordable by every patient to achieve a correct and cost effective treatment computer-based information and/or decision support Systems can be developed to do the task. Most hospitals today use some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making.

This raises an important question: “How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?”

So there is need of developing a master’s project which will help practitioners predict the diseases before it occurs.The diagnosis of diseases is a vital and intricate job in medicine. The recognition of diseases from diverse features or signs is a multi-layered problem that is not free from false assumptions and is frequently accompanied by impulsive effects.

Project Scope:

We suggest that the age, gender, chest pain, blood pressure, personnel history, previous history, cholesterol, fasting blood sugar, resting ECG, Maximum heart rate,

ISSN: 2278 – 7798

slope, etc. that may be used as reliable indicators to predict presence of heart disease.That data should be explored and must be verified from the team of heart disease specialist doctors. In future, we will try to increase the accuracy for the heart disease patient by increasing the various parameters suggested from the doctors by using different data mining techniques.

As a future work, the researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

Mathematical Model:

Naive Bayes or Bayes' algormith is the basis for many machine-learning and data mining methods. The algorithm is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It learns from the "evidence" by calculating the correlation between the dependent and independent Variables.

Probabilistic model

Abstractly, the probability model for a classifier is a conditional model

$p(C|F_1,\ldots,F_n)$

over a dependent class variable $C$with a small number of outcomes or *classes*, conditional on several feature variables F_1 through F_n. The problem is that if the nsumber of features n is large or if a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate the model to make it more tractable.

Using Bayes' theorem, this can be written

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F}{p(F_1,\ldots,F_n}$$

In plain English, using Bayesian Probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on $C$and the values of the features $F_i$are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$p(C, F_1, \ldots, F_n)$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned} p(C, F_1, \ldots, F_n) &= p(C)\ p(F_1, \ldots, F_n|C) \\ &= p(C)\ p(F_1|C)\ p(F_2, \ldots, F \\ &= p(C)\ p(F_1|C)\ p(F_2|C, F_1 \\ &= p(C)\ p(F_1|C)\ p(F_2|C, F_1 \end{aligned}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature $F_i$is conditionally independent of every other feature $F_j$for $j \neq i$, given the category C. This means that

$$p(F_i|C, F_j) = p(F_i|C),$$
$$p(F_i|C, F_j, F_k) = p(F_i|C),$$
$$p(F_i|C, F_j, F_k, F_l) = p(F_i|C),$$

and so on, for $i \neq j, k, l$. Thus, the joint model can be expressed as

$$\begin{aligned} p(C|F_1, \ldots, F_n) &\propto p(C, F_1, \ldots, F_n) \\ &\propto p(C)\ p(F_1|C)\ p(F_2|C)\ p(\\ &\propto p(C) \prod_{i=1}^{n} p(F_i|C)\,. \end{aligned}$$

This means that under the above independence assumptions, the conditional distribution over the class variable $C$is:

ISSN: 2278 – 7798

$$p(C|F_1,\ldots,F_n) = \frac{1}{Z}p(C)\prod_{i=1}^{n} p(F_i|C$$

where the evidence $Z = p(F_1,\ldots,F_n)$ is a scaling factor dependent only on $F_1,\ldots,F_n$, that is, a constant if the values of the feature variables are known.
s

Constructing a classifier from the probability model

The discussion so far has derived the independent feature model, that is, the naive Bayes probability model. The naive Bayes classifier combines this model with a decision rule. One common rule is to pick the hypothesis that is most probable; this is known as the *maximum a posteriori* or *MAP* decision rule. The corresponding classifier, a Bayes classifier, is the function classify defined as follows:

$$\text{classify}(f_1,\ldots,f_n) = \operatorname*{argmax}_c p(C =$$

Advantages & disadvantages:

Advantages:

1.The naive bayes algorithm supports simplicity, computational efficiency and good classification performance.

2.The naive bayes algorithm used to obtain the good results of large amount of data.

3. If predictor category is not present in the training data set , naive Bayes algorithm assumes that a new record with that category of the predictor has zero probability.

4. Efficient screening tools reduce demand on costly health care resources

5.Optimize allocation of hospital resources

6.Better insight into medical survey data

7.Computer-based training and evaluation

Disadvantages:

1. Applying data mining in the medical field is a very challenging task in medical profession.
2. According to the doctor intuition the clinical decision are often made.
3. The quality of service provided to patients is affected due to unwanted bias, errors and excessive medical cost.
4. Naive-Bayes states that if you have no occurrences of a class label and a certain attribute value together (e.g. class="good", shape="circle") then the frequency-based probability is set to zero.
5. It affect the posterior probability estimate when all the probabilities are multiplied you will get zero,this is Naive-Bayes' conditional independence assumption.

Conclusion

Data mining techniques is used to extracts the hidden knowledge from the system. This project provides an overview of applications of data mining techniques in administrative, clinical, research, and educational aspects of Clinical Predictions. This project state that there exists a great potential for data mining techniques to improve various aspects of Clinical Predictions, even though their is current practical use of data mining in health related problems is limited, Furthermore, to improve the quality of data rising of clinical data will increase the potential for data mining techniques an help to decrease the cost of heathcare. Data mining has become an essential part of the healthcare advancement because there is the rapid expansion in healthcare industry in reference to provide services and

ISSN: 2278 – 7798

information technology. There are no of data mining classification algorithms are available that helps in uncovering the valuable knowledge hidden behind them and in aiding the decision makers to improve the health care services.The presented study of data mining gives medical practitioners and health care planners a tool to help them in quickly comprehending vast clinical databases timely and precisely.This established that while the current practical use of data mining in health related problems is limited, there exists a great potential for data mining techniques to improve various aspects of Clinical Predictions.

Reference

(1) V. Krishnaiah, G. Narsimha & N. Subhash Chandra," A Study On Clinical Prediction Using Data Mining Techniques", IJICCT, Mar 2013.

(2) Dr. Meenu Dave, Priyanka Dadhich,"Applications of Data Mining Techniques:Empowering Quality Healthcare Services", IJICCT–JUL 2013.

(3) Mehmed, K.: "Data mining: Concepts, Models, Methods and Algorithms", New Jersey: John Wiley, 2003

(4) Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

(5)Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", IT Professional, 28-31, 2000.

(6)Dr. Meenu Dave, Priyanka Dadhich,"Applications of Data Mining Techniques:Empowering Quality Healthcare Services", IJICCT–JUL 2013.

(7) Berner E., "Clinical Decicion Support Systems". Springer Science+Business Media, 2007 .

ISSN: 2278 – 7798