Analysis Of Classifiers Diseases Prediction Using Weka Tool

[1]Rikendra, [2]Deepika
[1]Student, [2]Assistant professor
OM Institute of Technology, Hisar

Abstract - Binary classification is the task of classifying the members of a given set of objects into two groups on the basis of whether they have some property or not. A typical binary classification task in health care management could be diagnosis of medical testing to determine if a patient will die or live. We have used HEPITITIES database from UCI Machine Repository. The database is containing 153 instances and 20 attributes on which various binary classifiers have been applied, we have used mainly J48,NB TREE AND AD TREE classifiers. We have compared these algorithms on various parameters of performance evaluation; our focus will be on mainly four parameters namely: precision, sensitivity, accuracy and error rate. For classification task we have used WEKA and TANAGRA data mining tools. The results of experiment show that AD Tree gives a promising classification result on the basis of sensitivity, precision ,error rate and accuracy.

Keywords - Data mining, Weka tools ,Classification

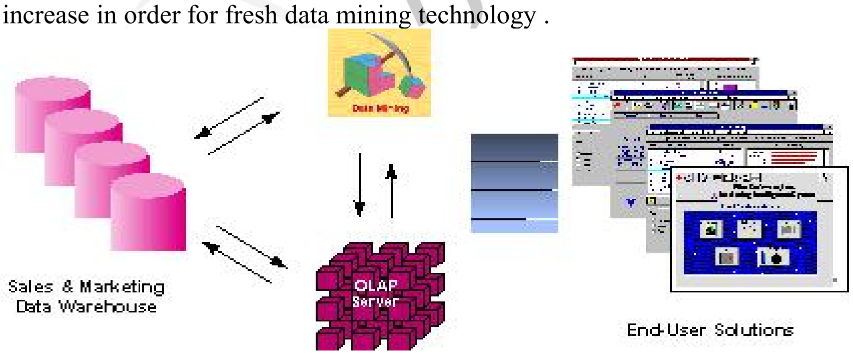
INTRODUCTION

Over the last few years, there has been an extensive growth in the amount of data collection from different resources. It may be from transport, library, financial, hospitals telecom, medical, and shopping records. This data can be taken on a single platform and analyzed digitally. All this is possible only due to the rapid growth in data science, communication media, networking, and computing technologies; despite of this, another technology due to this advancement came onto the market cover the progress of data mining tools that aim to infer valuable fashion from this data. But this simple access to private data causes a risk to the human being confidentiality. In this thesis, we discussed the piecewise quantization approach, which is used for confidentiality reserving clustering.

Data Mining

This technique deals with the pulling out of secret predictive information from big database. Data mining make use of complicated algorithms for the practice of sorting by huge amounts of data sets and expose related information. It can also be defined as "The Analysis step of the Knowledge Discovery in Databases process, or KDD", a relatively new and combination of different domains in computer science. Data mining is used to extract the Patterns from large data sets. For this we are combining different techniques from data analysis, statistics, and artificial intelligence with database organization.

With new technical move forward in many domains it becomes more significant tool. It can transform from current business scenario to extraordinary quantities of digital data into business intelligence. Data mining is presently used in a broad array of profile practice, such as advertising, observation and technical discovery. The emergent agreement that it can hold real cost that has led to a sudden increase in order for fresh data mining technology .



Integrated Data Mining Architecture

Scope of Data Mining

Data mining, it get its name by the likeness involving findings for main business information in a vast database – i.e. , receiving related item in gigabytes of store scanner data and mining a pile for a layer of valuable ore. Processes for data mining require changing all over huge amount of material, or smartly penetrating it to locate accurately where the value exists. This technology may provide new business prospects by providing a variety of features in databases of adequate size and quality. *Automated prediction of trends and behaviors.* The way of find out extrapolative information in huge databases is computerized by data mining. Questions that need wide investigation usually be able to answer directly and quickly with data mining method.

Data mining usually have four classes of responsibilities

- **Association Learning Rule** – this learning rule seek out for the interaction among variables. Let's take an example of a market might collect data for purchasing habits of a customer. Using this learning rule, the market can decide that products are regularly purchased together and can apply this information for marketing purposes.
- **Clustering** – it is the responsibility of discovering groups and structures in the data that are in some way or a different "similar", with no using identified organization of the data.
- **Classification** – in this technique we generalize the known organization to pertain to new data. i.e., an email program capacity endeavors to categorize an email as genuine or spam. Ordinary algorithms contain different techniques as SVM, NN, Adjoining neighbor, decision tree learning, and naive Bayesian classification.
- **Regression** – attempt to find out a job that models the data with the minimum error.

Data Mining Technologies:

- The main data mining techniques are as follows:
- *ANN (Artificial Neural Networks):* it is non-linear analytical models which learn by the training and it mimics the behavior of biological neural networks.
- *Decision Trees:* in this technique the sets of decisions are in tree-shaped structures. For the classification of dataset rules are generated with the help of these decisions.
- *Genetic Algorithms: it is an o*ptimization technique which has the concepts of evaluation design with the help of genetic combination, mutation, and natural selection.
- *Nearest Neighbor Method*: this technique categorize each record in a dataset on the behalf of the combination of the classes of the k record(s) most similar to it in a historical dataset.
- *Rule Induction*: in this technique the mining of useful if-then rules from data by statistical significance.

LITERATURE REVIEW

CLASSIFICATION

Data mining Classification technique is applied for the predicting group membership for data instances [1]. For example, classification techniques may be used to determine whether it will be rainy or sunny weather outside. Some well known classification techniques are decision trees and neural networks.

Clustering and classification analyses are the two very common data mining techniques of finding hidden patterns in data. Though these two techniques are also considered as the two sides of the same coin; but in fact these are two entirely dissimilar approaches in data analysis. Just similarity is that clustering and classification segments customer records into different data segments called classes. But contrasting clustering, classification is known in advance by the user or analyst that how the classes are defined.

TYPES OF CLASSIFICATION TECHNIQUES [1]:

There are following types of classification –

1. Classification based on Decision tree induction
   - i. Decision tree induction
   - ii. Tree pruning
   - iii. Decision trees based on extraction from classification rules
   - iv. Induction based on Scalability and decision tree
2. Classification based on Bayesian
   - i. Bayes theorm
   - ii. Naïve Bayesian classification
   - iii. Bayesian belief classification
3. Classification by backpropogation
4. Association based classification
5. Other classification schemes
   - i. "KNN"
   - ii. "Case based reasoning"
   - iii. "Genetic algorithms"
   - iv. "Rough set theory"
   - v. "Fuzzy set approaches"

CHALLENGES

Geo-spatial data warehouse be likely to be very large. besides, existing GIS data-sets are often split into traits and feature parts, that are unsurprisingly archived in hybrid data management system. Algorithmic supplies vary significantly for relational (attribute) data management and for topological (feature) data management. Associated to this is the array and variety of geographic data format which also show exclusive challenges. The digital geographic data disorder is creating new types of data formats beyond the conventional "vector" and "raster" format. Geographic data warehouse more and more consist of ill-structured data such as images and geo-referenced multi-media.

CONCLUSION

According to above classification techniques result, we can find the best technique for our hepatitis dataset by comparing output of confusion matrix and summary statistic. So the following results are achieved,

Name of the algorithm	Summary	Confusion Matrix
SMO	Correctly Classified Instances 130 90.2778 % Incorrectly Classified Instances 14 9.7222 % Ignored class unknown instance 2	a b 112 5 | a = DIE 9 18 | b = LIVE
NB TREE	Correctly Classified Instances 140 90.2222% Incorrectly Classified Instances 4 2.7772 % Ignored class unknown instances 2	a b 115 2 | a = DIE 2 25 | b = LIVE
NAÏVE BAYES	Correctly Classified Instances 125 86.8056 % Incorrectly Classified Instances 19 13.1944 % Ignored class unknown instances 2	a, b 106 11| a = DIE 8 19 | b = LIVE

From above conclusion we can say that NB Tree gives the more efficient result than others classifiers .But we may get also more promising result by applying other classifiers ,

REFERENCES

[1] Data Mining concept and Techniques jiawei Han and Micheline Kamber :2000,Simon Fraser University

[2] Dr. Varun Kumar, 2Luxmi Verma Department of Computer Science and Engineering, ITM University, Gurgaon, India." Binary Classifiers for Health Care Databases: A Comparative.

[3] Study of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer" IJCST Vol. 1, Iss ue 2, December 2010, ISSN : 2229-4333 (Print) | ISSN : 0976-8491 (Online).

[4] Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.

[5] John C. Platt Microsoft Research jplatt@microsoft.com Technical Report MSR-TR-98-14 April 21, 1998 © 1998 John Platt.

[6] Wikipedia NAÏVE BAYES CLASSIFIER :www.wikipedia.org/en/classification.htm

[7] An Introduction to the WEKA Data Mining System Zdravko Markov Central Connecticut State University markovz@ccsu.edu Ingrid Russell University of Hartford irussell@hartford.edu.

[8] uci machine repository for dataset: http://www.ics.uci.edu/~mlearn/databases/hepatitis/hepatitis.names Web Documents: About Hepatitis domain database.

[9] Benchmark results of Naive Bayes implementations (http:/ / tunedit. org/ results?d=UCI/ & a=bayes).

[10] http://wekadocs.com/node/6 Web Documents: WEKA Software.

[11] BAHÇEŞEHİR UNIVERSITY: APPLYING CLASSIFICATION METHODS ON HEPATITIS – DOMAIN DATASET(pdf) BY: Ergin DEMİREL (0569841)