Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 9, September 2014, pg.263 – 269

RESEARCH ARTICLE

STUDY OF CLASSIFIERS IN DATA MINING

Gaurav Taneja[1], Ashwini Sethi[2]

[1,2]Guru Kashi University, Talwandi Sabo (Bathinda), India

taneja132@gmail.com, mtech.gku@gmail.com

ABSTRACT - Hepatitis virus infection substantially increases the risk of chronic liver disease and hepatocellular carcinoma in humans and also affects majority of population in all age groups. It is the major challenge for many hospitals and public health care services for diagnosing hepatitis. Accurate diagnosis and exact prediction of the disease on time can save many patients life and there health. Hepatitis viruses are the most common cause of hepatitis in the world but other infections, toxic substances (e.g. alcohol, certain drugs), and autoimmune diseases can also cause hepatitis. Data mining is an effective tool to diagnose hepatitis and to predict result. This paper review the many data mining techniques which diagnosis hepatitis virus.

Keywords- Hepatitis, Data Mining, NB TREE, NAÏVE BAYES, SMO

I. INTRODUCTION

A. HEPATITIS

Hepatitis is an inflammation of the liver. The condition can be self-limiting or can progress to fibrosis (scarring), cirrhosis or liver cancer. Hepatitis viruses are the most common cause of hepatitis in the world but other infections, toxic substances (e.g. alcohol, certain drugs), and autoimmune diseases can also cause hepatitis. There are 5 main hepatitis viruses, referred to as types A, B, C, D and E. These 5 types are of greatest concern because of the burden of illness and death they cause and the potential for outbreaks and epidemic spread. In particular, types B and C lead to chronic disease in hundreds of millions of people and, together, are the most common cause of liver cirrhosis and cancer. Hepatitis A and E are typically caused by ingestion of contaminated food or water. Hepatitis B, C and D usually occur as a result of parenteral contact with infected body fluids. Common modes of transmission for these viruses include receipt of contaminated blood or blood products, invasive medical procedures using contaminated equipment and for hepatitis B transmission from mother to baby at birth, from family member to child, and also by sexual contact. Acute infection may occur with limited or no symptoms, or may include symptoms such as jaundice (yellowing of the skin and eyes), dark urine, extreme fatigue, nausea, vomiting and abdominal pain.

II. DATA MINING

Data mining is a technique that deals with the extraction of hidden predictive information from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. Data mining (the Analysis step of the Knowledge Discovery in Databases process, or KDD), a relatively young and interdisciplinary field of computer science, is the process of extracting Patterns from large data sets by combining methods from statistics and artificial intelligence with database management.

A. CLASSIFICATION

Classification is a data mining (machine learning) technique used to predict group membership for data instances[1]. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy". Popular classification techniques include decision trees and neural networks. Two common data mining techniques for finding hidden patterns in data are clustering and classification analyses. Although classification and clustering are often mentioned in the same breath, they are different analytical approaches. Classification is a different technique than clustering. Classification is similar to clustering in that it also segments customer records into distinct segments called classes. But unlike clustering, a classification analysis requires that the end-user/analyst know ahead of time how classes are defined. For example, classes can be defined to represent the likelihood that a customer defaults on a loan (Yes/No). It is necessary that each record in the dataset used to build the classifier already have a value for the attribute used to define classes. Because each record has a value for the attribute used to define the classes, and because the end-user decides on the attribute to use, classification is much less exploratory than clustering. The objective of a classifier is not to explore the data to discover interesting segments, but rather to decide how new records should be classified -- i.e. is this new customer likely to default on the loan?

Example 1- Animal Classification: For centuries, the naming and **classification** of living organisms into groups has been an integral part of the study of nature.

Example 2- Definition: With trademarks, the class or **classification** means what kind (class) of goods or services offered are being represented by a certain trademark.

1) *CLASSIFICATION BY DECISION TREE INDUCTION:* The topmost node is called root node.in order to classify unknown sample, the attribute values of sample are tested against the decision tree. Decision trees can easily be converted into classification rules. The basic algorithm for decision tree induction is a greedy algorithm which constructs decision trees in a top-down recursive divide-and-conquer manner. The, is a version of ID3, a well-known decision tree induction algorithm.
   The basic strategy is as follows:

The knowledge represented in decision trees can be extracted and represented in the form of classification IFTHEN rules. One rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent (\IF" part). The leaf node holds the class prediction, forming the rule consequent (\THEN" part). The IF-THEN rules may be easier for humans to understand, particularly if the given tree is very large.

Example-

IF age = \<30" AND student = no THEN buys computer = no

IF age = \<30" AND student = yes THEN buys computer = yes

IF age = \30-40" THEN buys computer = yes

IF age = \>40" AND credit rating = excellent THEN buys computer = yes

IF age = \>40" AND credit rating = fair THEN buys computer = no.

The efficiency of existing decision tree algorithms, such as ID3 and C4.5, has concern when these algorithms are applied to the mining of very large, real-world databases. Most decision tree algorithms have the restriction that the training samples should reside in main memory. In data mining applications, very large training sets of millions of samples are common. Hence, this restriction limits the scalability of such algorithms, where the decision tree construction can become inefficient due to swapping of the training samples in and out of main and cache memories.

2) *Bayesian classification:* Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

Bayes theorm: Let X be a data sample whose class label is unknown. Let H be some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P(H/X), the probability that the hypothesis H holds given the observed data sample X.P(H/X) is the posterior probability, or a posteriori probability, of H conditioned on X. For example, suppose the world of data samples consists of fruits, described by their color and shape. Suppose that X is red and round, and that H is the hypothesis that X is an apple. Then P(H/X) reflects our confidence that X is an apple given that we have seen that X is red and round. In contrast, P(H) is the prior probability, or a priori probability of H. For our example, this is the probability that any given data sample is an apple, regardless of how the data sample looks. The posterior probability, P(HjX) is based on more information (such as background knowledge) than the prior probability, P(H), which is independent of X. Similarly, P(XjH) is the posterior probability of X conditioned on H. That is, it is the probability that X is red and round given that we know that it is true that X is an apple. P(X) is the prior probability of X. Using our example, it is the probability that a data sample from our set of fruits is red and round. How are these probabilities estimated?" P(X), P(H), and P(X/H) may be estimated from the given data, Bayes theorem is useful in that it provides a way of calculating the posterior probability, P(H/X)from P(H), P(X), and P(X/H). Bayes theorem is: The naive Bayesian classifier makes the assumption of class conditional independence, i.e., that given the class label of a sample, the values of the attributes are conditionally independent of one another. This assumption simplifies computation. When the assumption holds true, then the naive Bayesian classifier is the most accurate in comparison with all other classifiers. In practice, however, dependencies can exist between variables. Bayesian belief networks specify joint conditional probability distributions. They allow class conditional independencies to be defined between subsets of variables. They provide a graphical model of causal relationships, on which learning can be performed. These networks are also known as belief networks, Bayesian networks, and probabilistic networks.

3) *Classification by back propagation:* Back propagation is a neural network learning algorithm. The field of neural networks was originally kindled by psychologists and neurobiologists who sought to develop and test computational analogues of neurons.
4) *Association based classification:* It is a highly active area of research in data mining. One method of association-based classification, called associative classification, consists of two steps. In the first step, association rules are generated using a modified version of the standard association rule mining algorithm known as Apriori. The second step constructs a classifier based on the association rules

discovered. Let D be the training data, and Y be the set of all classes in D. The algorithm maps categorical attributes to consecutive positive integers. Continuous attributes are discretized and mapped accordingly.

5) *KNN CLASSIFICATION:* Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n-dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. These k training samples are the k "nearest neighbors" of the unknown sample."Closeness" is defined in terms of Euclidean distance, The unknown sample is assigned the most common class among its k nearest neighbors. When k = 1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space.

6) *Case-based reasoning:* Case-based reasoning (CBR) classifiers are instanced-based. Unlike nearest neighbor classifiers, which store training samples as points in Euclidean space, the samples or "cases" stored by CBR are complex symbolic descriptions. Business applications of CBR include problem resolution for customer service help desks, for example, where cases describe product-related diagnostic problems. CBR has also been applied to areas such as engineering and law, where cases are either technical designs or legal rulings, respectively.

7) *Genetic algorithms:* Genetic algorithms attempt to incorporate ideas of natural evolution. In general, genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits. As a simple example, suppose that samples in a given training set are described by two Boolean attributes, A1 and A2, and that there are two classes, C1 and C2. The rule \IF A1 and not A2 THEN C2" can be encoded as the bit string "100" where the two leftmost bits represent attributes A1 and A2, respectively, and the rightmost bit represents the class. Similarly, the rule \if not A1 and not A2 then C1" can be encoded as \001". If an attribute has k values where k > 2, then k bits may be used to encode the attribute's values. Classes can be encoded in a similar fashion. Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules in the current population, as well as offspring of these rules. Typically, the fitness of a rule is assessed by its classification accuracy on a set of training samples. Offspring are created by applying genetic operators such as crossover and mutation. In crossover, substrings from pairs of rules are swapped to form new pairs of rules. In mutation, randomly selected bits in a rule's string are inverted. The process of generating new populations based on prior populations of rules continues until a population P "evolves" where each rule in P satisfies a prespecified fitness threshold. Genetic

algorithms are easily parallelizable and have been used for classification as well as other optimization problems. In data mining, they may be used to evaluate the fitness of other algorithms.

8) *Rough set theory***:** Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It applies to discrete-valued attributes. Continuous-valued attributes must therefore be discretized prior to its use.

Rough set theory is based on the establishment of equivalence classes within the given training data. All of the data samples forming an equivalence class are indiscernible, that is, the samples are identical with respect to the attributes describing the data. Given real-world data, it is common that some classes cannot be distinguished in terms of the available attributes. Rough sets can be used to approximately or "roughly" define such classes.

A rough set definition for a given class C is approximated by two sets - a lower approximation of C and an upper approximation of C. The lower approximation of C consists of all of the data samples which, based on the knowledge of the attributes, are certain to belong to C without ambiguity. The upper approximation of C consists of all of the samples which, based on the knowledge of the attributes, cannot be described as not belonging to C. Decision rules can be generated for each class. Typically, a decision table is used to represent the rules.

9) Fuzzy set approaches: Rule-based systems for classification have the disadvantage that they involve sharp cut-offs for continuous attributes. For example, consider applications for customers who have had a job for two or more years, and who have a high income (i.e., of more than $50K) , a customer who has had a job for at least 2 years will receive credit if her income is, say, $51K, but not if it is $50K. Such harsh thresholding may seem unfair. Instead, fuzzy logic can be introduced into the system to allow "fuzzy" thresholds or boundaries to be defined. Rather than having a precise cutoff between categories or sets, fuzzy logic uses truth values between 0:0 and 1:0 to represent the degree of membership that a certain value has in a given category.

REFERENCES

[1] http://hepatitis.about.com/od/overview/a/numbers.htm.
[2] Houzifa M. Hintaya, F. M.-A. (August-2013). *Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach*. International Journal of Scientific & Engineering Research, Volume 4, Issue 8, 680-685.
[3] Data Mining concept and Techniques jiawei Han and Micheline Kamber :2000, Simon Fraser University.
[4] Dr. Varun Kumar, 2Luxmi Verma Department of Computer Science and Engineering, ITM University, Gurgaon, India. *Binary Classifiers for Health Care Databases: A Comparative Study of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer*. IJCST Vol. 1, Issue 2, December 2010, ISSN: 2229-4333(Print) | ISSN: 0976 - 8491(On l i n e) .

