DATA MINING TECHNIQUES FOR HEALTHCARE DATA

N. Thangarasu, G. Manivasagam*

Abstract

Data mining is the method of extracting secret data from large database sets. There are many tools and techniques used nowadays to predict this useful information[1].

Researcher in Data mining field has come across many tools and equipment to develop an application for data analyzing in high dimensional databases. But implementing data mining techniques in the medical field to diagnose, predict and steps to prevention measures of the particular disease are increased. These applications include analysis regarding treatments attempt to improve those early diagnosing capabilities and prevention to reduce the death rate gradually. This paper discusses about the detailed study on general data mining techniques and techniques where are used in health care data.

Key Words: Data Mining, Audio, Video, Image, Sequence Data, Hypertext Data, Health Care Date.

I. INTRODUCTION

The most commonly used techniques to predict useful knowledge from large databases and warehouses are:

a) Artificial Neural Network

These are models that predict knowledge by learning through training.

b) Decision Trees

Hierarchical tree structures that provide a set of decisions based on a given set of conditions. Rules are generated based on these decision trees which impart useful information on the given data set.

c) Genetic Algorithms

Concepts like genetic combination, mutation and natural selection are the process used by genetic algorithms based on the concept of evolution

D) Nearest Neighborhood Method

The data in the dataset are classified based on the records similar to its neighbor.

I.1. Data Types that can be mined

The data that can be mined are of different types they are:

- Data streams
- Time series data
- Graphs
- Spatial data
- Multimedia data
- Text data
- World wide web

I.2. Multimedia Database

A Database is a collection of different forms of information. The Multimedia database holds a large collection of multimedia objects. The data types may be audio data, video data, image data, sequence data and hypertext data. The main focus given in this research is image data mining. The main methods used for Multimedia data mining is

- Similarity search
- Multidimensional analysis
- Classification and prediction analysis
- Mining associations in Multimedia data

Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
*Corresponding Author

In content-based retrieval systems, image content acts as the source for information retrieval. The information used may be color histogram, grey levels of the pixels, texture, pattern, image topology, object shape, layout and locations of certain patterns within the image. The queries are provided in the form of image features. Initially, image features are sketched that are color, texture, or shape information. These functions are then converted into vector form to suit the database of images. There are many applications of this image mining in various fields. Some of them are medical image diagnosis, weather prediction and web search engines for images.

In similarity, based on retrieval process many approaches are used from which the following are important. They are

- Color Histogram Based Signature
- Multi Feature Composed Signature
- Wavelet Based Signature
- Wavelet Based Signature with Region Based Granularity.

I.2.1. Color Histogram Based Signature

This method does not consider the s1hape, texture and location information. The signature involves the color histogram based on the color composition of the image. It matches color information of two images and analyses the produced result.

I.2.2. Multi Feature Composed Signature

The image signature in this method includes multiple features which are considered for retrieval process. The features may be the combination of color, shape, texture, orientation and location. The similarity of each feature is calculated with a different distance function. Among these entire features, few features are selected for the optimal retrieval process.

I.2.3. Wavelet Based Signature

In this method, the representation of the image is in the form of wavelet coefficient. All the features such as shape, texture and location information are represented in a single framework of wavelet coefficient. Since all the features are in an integrated format, the efficiency of prediction is improved and the multiple search conditions are not needed. A single signature co-efficient is calculated for the entire image.

I.2.4. Wavelet Built Signature with Region Established Granularity

Signature in this module is calculated only for a specific granularity of a region. Regions may get matched for similar images. So, the region of one image performs scaling or translation of a matching region with others. The similarity is measured with the query image and target image.

II. APPLICATIONS OF DATA MINING TOWARDS HEALTHCARE DATA

Heart disease related database includes the details about the patient records in huge number; in such cases to analyze those datasets, prior knowledge is needed. Statistical approaches and data mining techniques may help the specialists to diagnose and determine life threatening aspects of the patient. The statistical analyses might help in analyzing the criteria like age, blood pressure, smoking, high cholesterol levels, diabetes, stress, medical history, obesity and lack of exercise. The cognizance helps to identify the patient with major risk factors easily through data mining concepts.

Researchers have implemented numerous data mining techniques to support medical professionals, cardiovascular specialists. Few common methods that are used wider are Naive Bayes, decision tree and K-nearest neighbor[2]. But there are other methods such as the kernel density, neural network, bagging algorithm, self-organizing map (SOM) Support Vector Machine (SVM), etc are explained below.

II.1. Decision Tree

Decision trees vary primarily in the mathematical model they cause in choosing feature class over rule extraction. Gain ratio decision tree is widely used and gains high interest among the researchers in this field. It shows the relationship between entropy and classified information.

Through entropy strategy, the attribute helps to reduce entropy and improves the gathering of knowledge to form the tree root. To achieve this, the information gain of each element must be calculated. Based on this better attribute maximizing entropy might be chosen. Entropy is represented through the below formula[3].

$$E = -\sum_{i=1}^{k} p_i \log_2^{p_i} \qquad (2\text{-}1)$$

In which k is the maximum number of variable classes, pi is the ratio of the number of ith class events dependent on the occurrence of i for the number of samples.

II.2. Bayesian Network

It is a measurable strategy used to anticipate the participation of the class utilizing probability theory. Bayesian network makes use of classification process to assume the effect of the value of a theorem and independent nature from additional elements[4]. This distinction is called conditional independence. This calculation bridges the gap between engaged calculations named as Naive simply.

This very well estimates the previous probability of variable response and the conditional likelihood of another. Initial preparation requires both actual and conditional probabilities[5]. The existence of response variable shall be determined for each test sample of the database. Following this, the response with the highest value is chose through the following relation [6].

$$P(v = c_i) = P(c_i) \times \sum_{j=1}^{n} P(a_j = v_j | class = c_i) \qquad (2\text{-}2)$$

Whereas V, c_i, a_j and v_j are said to be the test sample, variable response value, data attribute, and sample test value respectively.

II. 3. K-nearest Neighbor

A primarily storage-centered technique that starts from preparation samples and stores the memory during runtime[7]. Suppose if a first trial is defined by a1, a2,... to an, and b is the second trial stated as b1, b2,..., then bn calculates the distance between the two by the below formula 2-3.

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_n - b_n)^2} \qquad (2\text{-}3)$$

II.4. Support Vector Machine

For a particular support vector, the Support Vector Machine boundary has to be determined for achieving the better data detection and distinction. In SVM, the records that present within support vectors are considered to be improper record for machine and modeling. In such situation, if the algorithm is less sensitive then it fails to identify the best one compared to the outermost possible space in entire modules. SVM must transfer data to a new space with programmed modules so that data can be categorized and divided accordingly. Each support vector defines a boundary line of the classes in equation form.

II.5. Classification

Classification is one among the machine learning algorithm which predicts the group membership for the given data instances. For instance, it can be used in the departmental store to predict and to know the customer interest over the specific items over a period of time like

sunny, rainy and cloudy. Methods of classification are divided into supervised methods and unsupervised ones.

• Supervised learning

Supervised learning indicates an element of a named information that comprises of a lot of prepared tests. Through supervised learning, each example consists of a couple of data objects called vectors and yield value called supervisory signals. A supervised learning algorithm looks at the information prepared and produces a predetermined capacity, known as classifier. The construed work helps in distinguishing the ideal yield for any substantial information.

• Unsupervised learning

In machine learning, this type of learning algorithm's data seems to be unlabeled. So, it is hard to find the hidden structure from the unlabeled data. As a result, both training and testing those datasets remains in error. Thus, it is the major variance between unsupervised learning and supervised learning.

II.6 Neural Network

For particular applications, neural networks generate high accurate results eventually. It makes use of feed forward neural network model[8], variable learning and back-propagation methods. Heart Diseases database can also train neural network. The model starts with the input of medical data to produce this, and implements an ANN algorithm[9]. It starts predicting the results once the training model is developed. The method to compute begins with the classification of medical data into testing and training data. First, an underlying weight must be allotted to each element arbitrarily. The determined mistakes are balanced by the heaviness of the accessible highlights. Each element's last weight is inspected to keep away from the forthcoming blunders. This procedure is rehashed for number of times till developing the preparation models. At last, this model assists with assessing the testing information.

III. CONCLUSION

A variety of strategies have been explored, such as the use of deep learning to create new learning techniques which are capable of mining essential facts and composite structures from various input materials. Using Fuzzy analytical data mining techniques, we can classify disease forecasting in terms health care, some of the mining techniques such as an automated thresh keeping technique are used to pre-process.

REFERENCES

[1] D. Hand, H. Mannila and P. Smyth, "Principles of data mining", MIT, (2001).

[2] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).

[3] Data Mining. Concepts and Techniques, 3rd Edition Han J, Kamber M, 2006

[4] Shujun Huang, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narandes, Yang Wang, And Wayne Xucancer Genomics Proteomics "Applications of Support Vector Machine (SVM) Learning " Cancer Genomics v.15(1); Jan-Feb 2018

[5] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD process of extracting useful knowledge form volumes of data. commun. ", ACM, vol. 39, no. 11, (1996), pp. 27-34.

[6] J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd ed. The Morgan Kaufmann Series, (2006).

[7] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases",

Commun. ACM, vol. 39, no. 11, (1996), pp. 24-26.

[8] P. R. Harper, "A review and comparison of classification algorithms for medical decision making", Health Policy, vol. 71, (2005), pp. 315-331

[9] R. D. Canlas Jr. , "Data Mining in Healthcare:Current Applications and Issues", (2009).