International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET

A Basic Study of Classification and Prediction Method of Data Mining

Pooja Singh, Dr. G. N. Singh, Dr. Arvind Singh
Department of Computer Science, Sudarshan College, Lalgaon, Awadhesh Pratap Singh University Rewa, Madhya Pradesh, India

Article Info
Volume 8, Issue 2
Page Number : 475-479

Publication Issue :
March-April-2021

Article History
Accepted : 10 April 2021
Published: 30 April 2021

ABSTRACT

Although databases are very rich with hidden information that can be used intelligent business decisions. Prediction and classification there are two main forms of data analysis technique/method which can be used to fetch and create models that can describe important classes of data to predict future trends. While the classification predicts certain labels and the prediction model predicts continuous-valued tasks. A classification model can be classify a particular things are appropriate or not, safe or risky, where as a predictive model can be built to predict about spending capacity of peoples based on their given income. Prediction and classification methods have been proposed by various researchers in decision support system, expert systems, machine learning, statistics etc. Majority of the algorithms and techniques are souvenir denizen, you can assuming as a small data size. Recent database mining research has been done on classification and prediction techniques which are capable of handling large amount, souvenir denizen data. All these techniques considered as a parallel and distributed processing.

Keywords : Souvenir Denizen, Parallel and Distributed Processing

I. INTRODUCTION

In this paper, we have done basic study of basic techniques for classification like induction, decision tree, Bayesian belief networks, neural networks and Bayesian classification. Association-based classification is also discussed along with the integration of data warehousing techniques with classification.

Prediction methods are briefly discussed, including linear, nonlinear, and generalized linear regression models.

II. LITERATURE SURVEY

In this paper, we have done review of literatures on the classification and prediction methods of data mining. Here are some most important research papers given below-

S.Vijiyarani and S.Sudha (2013) this is a survey paper. This paper discusses the problem of summarizing various algorithms of data mining used in the field of medical prediction. The major focus is on using different algorithms and combinations of multiple

target features for the prediction of different types of disease using data mining. [1]
In this paper both researcher discussed about the heart related disease prediction, and machine learning algorithms known as naïve bayes, K-NN, Decision List etc.[1]

Kaixi Zhang, Yingpeng Hu, Yanghui Wu (2018) in this paper, researcher report a research efforts where they developed three prediction models for the demands of the farmers loan. From three of that models, two of the models are from the machine learning and one is belongs from statistics [2].

Aarti Sharma at el. (2014) in her paper she tells that data mining as a "decision support" process, where they search information using certain patterns for patterns of information in data. Such types of patterns are clustering, prediction, classification, association and sequential patterns etc. They also tell that the educational, scientific and commercial, applications are increasingly dependent on above methodologies. Her view is that Decision trees are a reliable and effective decision making technique which can provide high accuracy of classification. They also explain that the data mining can help to play an important role in the field of medicine or health care and prediction of various diseases [3].

Nikita Jain, Vishal Srivastava (2013) in this paper, the concept of data mining was briefly introduced and its importance to the methodologies was shown. Data mining based on neural network and genetic algorithm is extensively researched and the major techniques and methods to achieve data mining on neural network and genetic algorithm are surveyed. A formal review of the area of rule extraction from ANNs and GAs is also done in this paper [4].

Parneet Kaur at el. (2015) in this paper, they are all used classification techniques for prediction on the dataset of students. They also predict and analyze student's performance as well slow learners among them. In his/her study, a model was created using some of the selected student and related input variables collected from schools [5].

III. CLASSIFICATION

The Classification of data is takes very important place in the field of data mining so, data classification is a two-step process, in which the first step a model of data classification is created for describing a predetermined set of data classes or their concepts. The model can be created by analyzing database and their tuples described by attributes. Every tuple is concoctive which is belonging to a pre-concerted class, as pre-concerted by one of the attributes which is also called the class label attribute. In the reference of classification, data tuples are also envisaged to as samples, examples, or objects.
Typically there is models called learned model. The learned model can be represented as a decision tree, classification rule as well as mathematical formula. Suppose given a database of bank customers, by using classification techniques can be learned and identify customers buying capacity and credit ratings. The classification techniques can also be used to categorize samples of available data.
In the second step the classification model is used for classification and the predicted accuracy of the model or classifier is estimated.

2.1 Difference between Prediction and Classification

A prediction is a statement that will be happen in future or may be happen in future. We can think prediction as the construction and a model to the class of an un-labeled object, or to assessing the future event, value or value ranges of an attribute. Classification and regression are two main types of future assessment techniques where forecasting problems realized. Classification is used to predict various discrete or nominal type values, where as regression is used to predict continuous/sequential

values. However, we can say that the use of prediction is to predict/forecast class labels.

There is various application of Classification and prediction technique such as medical diagnosis, financial loan approval, credit approval, forecasting and marketing etc.

2.2 Issues of classification and prediction

The following processing steps can be applied to data classification which helps to improve the accuracy, scalability and efficiency in the classification/prediction process.

A. Cleaning of Data

This refers to the removal or reducing noise by applying smoothing techniques and the treatment of missing values. In the view of missing values replacing value with the most commonly is occurring value for attribute. Most of the classification algorithms have techniques to handle noisy and missing data.

B. Relevance Analysis

Many features in the data may be irrelevant to the classification and tasks of prediction. Assuming you are using data recording the day of the week the financial company applied for registration in stock market is unlikely to be relevant to the success of the application.

In addition, other attributes may be redundant. Therefore, relevance analysis can be processed on the data with the purpose of removing any irrelevant or unnecessary features from the learning process. Learning from a set of features on relevance analysis can help improve classification efficiency and scalability.

C. Transformation of Data

Data can be normalized to higher-level concepts and concept hierarchy can be used for the transformation of data. This is especially useful for constant-valued features.

We can think for example numerical values for income attribute can be normalized to discrete categories such as high, medium and low. Similarly, nominal-valued features, such as a road, can be generalized or can think to higher-level concepts, like a city. Therefore generalization compresses the original training data, fewer input/output processing that can be involved during learning process. Data can also be normalized various way, especially when methods involving like neural networks or distance measurements during in the learning phase. Normalization involves like scaling all values so that they fall within a small range, such as -1.0 to 1.0, or 0 to 1.0.

D. Comparison of Classification Techniques

Classification and prediction techniques can be compared with the following criteria:

Decision Tree

This is a flow-chart-like hierarchical tree structure, where classification processes are done in the form of tree and their each internal node denotes a test condition on a feature. Each branch of tree represents the result of the test condition, and the leaf nodes represent classes. The first topmost node in a tree structure is called root node. The topmost mode in a tree is the root node.

Decision Tree Induction

The main and basic algorithm for decision tree induction is a greedy algorithm that creates a decision tree in as a top-down iterative and divide-and-conquer manner.

The basic strategy is as follows:

Step-1. The first node is called tree root and starts as a single node representing the samples.

Step-2. If the samples belonging from the same class and categories, then the node will becomes a leaf.

Step-3. The algorithm uses an entropy-based measure known as information knowledge.

Step-4. This attribute becomes the "test" or "decision" attribute at the node

Step-5. In this step of algorithm, all attributes are categorized as a discrete-valued.

Step-6. Here a branch is constructed for known values of the test attribute, and the samples are partitioned accordingly.

Step-7. The algorithm uses the same process recursively/iteration to form a decision tree for the samples at every partition.

The iterative partitioning stops only when the conditions are satisfy that is given below:

1. If all samples for a given node belong to the same class or categories.
2. If there are no remaining attributes which can be partitioned.
3. If majority of voting is employed.
4. If involved nodes are converting into a leaf nodes.
5. If there are no samples available for the branch test-attribute. In that condition, a leaf is constructed with the majority categories in samples.

IV. DATA MINING

The process of organizing, analyzing and thinking about data is way by which we can find what the data can do. Data Analysis means it is a process of cleaning, inspecting transforming, and modeling various types of data. The main purpose of data analysis is to finding useful information, providing conclusions, and help in decision making. Data analysis consists of approaches, including various techniques under a list of names, in various business, science, and social science area. [8]

The data mining is an automatic and semiautomatic analysis of large amount of data for the extraction of interesting patterns of data records known as cluster analysis, a collection of records for anomaly detection, and to find out dependencies i.e., association rule mining and sequential pattern mining. The database techniques are spatial indices. These patterns are used in analysis i.e., in machine learning and predictive analytics.

Data Mining is the finding and discovery of unknown information from the databases [13] [14]. Data fetching, data fishing and data snooping refer to the use of data mining method. The sample part of a larger population dataset which are too small for reliable statistical inferences to be made to validate the patterns that discovered. All these techniques can be used in the creation of new hypothesis to test data against the larger data.

The Assessment of data mining functions and products are the results of the impact from many of the disciplines, in which we can includes the databases, information retrieval, statistics, algorithms, and machine learning [8].

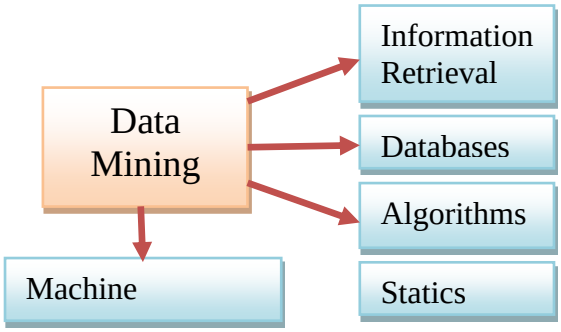

Figure: Historical Data Mining

Task of Data Mining

There are two categories of tasks of Data mining:

- Classification and Predictive
- Descriptive

V. CONCLUSION

In this study of we have studied various prediction and classification techniques that can be used for classifying and categorizing the data and analysis those finding. After all we can make better prediction on the basis of various techniques. In this paper we have also discussed various issues related to

classification and prediction. It is very difficult task for the prediction of various things without using classification techniques, it is quite possible with help of Data Mining applications but this effort decreases the human efforts quality.

VI. REFERENCES

[1]. S. Vijiyarani and S.Sudha, “Disease Prediction in Data Mining Technique – A Survey”, International Journal of Computer Applications & Information Technology, Vol 2, Issue 1, 2013

[2]. Kaixi Zhang, Yingpeng Hu, Yanghui Wu,"Classification and Prediction on Rural Property Mortgage Data with Three Data Mining Methods", Journal of Software Engineering and Applications, 2018, 11, 348-361, 2018

[3]. Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava,"Application of Data Mining – A Survey Paper", International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2023-2025, 2014

[4]. Nikita Jain1, Vishal Srivastava, "DATA MINING TECHNIQUES: A SURVEY PAPER", International Journal of Research in Engineering and Technology eISSN: 2319-1163, Volume: 02 Issue: 11, 2013

[5]. Parneet Kaur at el., "Classification and prediction based data mining algorithms to predict slow learners in education sector", ScienceDirect Procedia Computer Science 57, 500 – 508, 2015

[6]. Boros E., P.L. Hammer, T. Ibaraki, A. Kogan.(1997). Logical Analysis of Numerical Data. Mathematical Programming, 79:163-190.

[7]. Hammer P.L.(1986). The Logic of Cause-effect Relationships, Lecture at the International Conference on Multi- Attribute Decision Making via Operations Research-based Expert Systems, Passau, Germany.

[8]. Proceedings of 4th international conference on statistical sciences Volume (15), University of Gujrat Pakistan, 15: 78