Chapter 8
Data Classification and Prediction

Pudumalar S
Thiagarajar College of Engineering, India

Suriya K S
Thiagarajar College of Engineering, India

Rohini K
Thiagarajar College of Engineering, India

ABSTRACT

This chapter describes how we live in the era of data, where every event in and around us creates a massive amount of data. The greatest challenge in front of every data scientist is making this raw data, a meaningful one to solve a business problem. The process of extracting knowledge from the large database is called as Data mining. Data mining plays a wrestling role in all the application like Health care, education and Agriculture, etc. Data mining is classified predictive and descriptive model. The predictive model consists of classification, regression, prediction, time series analysis and the descriptive model consists of clustering, association rules, summarization and sequence discovery. Predictive modeling associates the important areas in the data mining called classification and prediction.

INTRODUCTION

The greatest challenge in front of every data scientist is making this raw data, a meaningful one to solve a business problem. Data is the beginning point of all data mining process. The raw data or the collected data cannot use directly to build the business models. Hence processing added value to the data called information. The information is the processed data which is stored and managed in the large database. The process of extracting knowledge from the large database is called as Data mining. Data mining software analyses relationships and patterns in stored transaction data based on open-ended user queries. In the data mining Major elements are listed follows 1) Extract, make over and load transaction data onto the data warehouse system. 2) Store and manage the data in a multidimensional database system. 3)

DOI: 10.4018/978-1-5225-4044-1.ch008

Provide data access to information technology professional and business analysts. 4) Analyze the data by application software. 5) Present the data in a useful format, such as a graph or table. Data mining is classified predictive and descriptive model. The predictive model consists of classification, prediction, regression, and time series analysis. The descriptive model is consist of clustering, summarization, association rules and sequence discovery. Predictive modeling associates the important areas in the data mining called classification and prediction. Applications of predictive modeling include customer retention management, cross-selling, direct marketing, and credit approval which are notable by the nature of the variable being predicted. "Why classification is important?" The classification problem attempts to learn the relationship between a set of feature variables and a target variable of interest. For example, the bank manager has massive customer's data, which consists of customer details and who all are applying for the loan. The manager will classify the customer data and easily identify the customers who all are in the risk and safe condition which is called as classification. The classified data are used to create a pattern to forecast the future condition of the customers, which is called as a prediction. Now a day's data classification and prediction holds promise in many fields to enhance efficiency and reduces the time complexity of the application. Classification and Prediction can be performed only when the data comes in the following steps, data pre-processing includes data cleaning, replace missing values, data relevance, data transformation, and data reduction. Most classification algorithms typically have two phases:

1. **Training Phase:** In this phase, a training model is constructed from the training instances. Intuitively, this can be understood as a summary mathematical model of the labeled groups in the training data set.
2. **Testing Phase:** In this phase, the training model is used to determine the class label (or group identifier) of one or more unseen test instances.

Classification predicts a certain outcome based on a given input. In order to predict the result, the algorithm processes a training set containing a set of attributes and the individual outcome, where usually called prediction attribute. The algorithm tries to determine relationships between the attributes that would make it possible to predict the conclusion. The inputs are analyzed by using data mining algorithms and produce a prediction. The prediction accuracy defines how "good" the algorithm is. Decision tree based, Rule-based, Instance-based learning, Bayesian Classification, Neural Networks, Ensemble methods, Support Vector are the most popular and powerful used in classification and prediction. A decision tree is a flow chart; similar to the data structure trees consists of decision node, leaf node, arc or edges, and path. Entropy is used to measure of homogeneity of the dataset. Information gain, Gain ratio, and Gini-index are the methods to select the attributes and generate the tree. Instance-based learning methods consist of K-nearest neighbor's algorithm, weighted regression, and case-based reasoning. The rule-based method is used to generate rules to classify the dataset which is consists of PRIMS and RIPPER. The Bayesian classification includes naïve Bayes and Bayesian belief network. Artificial Neural Network (ANN) of Neural Network is information -processing paradigm that is stimulated as the human brain's information processing mechanism. ANN has three different classes; these are single layer feed forward, multi-layer feeds forward and recurrent. Ensemble methods are used improve the accuracy of the classifiers which is achieved by bagging and boosting. Support vector machine is used for classification and regression methods which satisfying from theoretical points of view. SVM is used in many real-time applications such as text categorization, image classification, bioinformatics, and hand -written character recognitions.