Chapter 44
A Dynamic and Scalable Decision Tree Based Mining of Educational Data

Dineshkumar B. Vaghela
Parul University, India

Priyanka Sharma
Raksha Shakti University, India

Kalpdrum Passi
Laurentian University, Canada

ABSTRACT

The explosive growth in the amount of data in the field of biology, education, environmental research, sensor network, stock market, weather forecasting and many more due to vast use of internet in distributed environment has generated an urgent need for new techniques and tools that can intelligently automatically transform the processed data into useful information and knowledge. Hence data mining has become a research are with increasing importance. Since continuation in collection of more data at this scale, formalizing the process of big data analysis will become paramount. Given the vast amount of data are geographically spread across the globe, this means a very large number of models is generated, which raises problems on how to generalize knowledge in order to have a global view of the phenomena across the organization. This is applicable to web-based educational data. In this chapter, the new dynamic and scalable data mining approach has been discussed with educational data.

INTRODUCTION

Web usage mining refers to non-trivial extraction of potentially useful patterns and trends from large web access logs. In the specific context of web-based learning environments, the increasing proliferation of web-based educational systems and the huge amount of information that has been made available

DOI: 10.4018/978-1-7998-2460-2.ch044

has generated a considerable scientific activity in this field. As an increasingly powerful, interactive, and dynamic medium for delivering information, the World Wide Web in combination with information technology has found many applications. One popular application has been for educational use, as in Web-based, distance or distributed learning. The use of the Web as an educational tool has provided learners and educators with a wider range of new and interesting learning experiences and teaching environments that were not possible in traditional education. These platforms contain a considerable amount of e-learning materials and provide some degree of logging to monitor the progress of learning keeping track of learners' activities including content viewed, time spent at a particular subject and activities done. This monitoring trawl provides appropriate data for many different contexts in universities, like providing assistance for a student at the appropriate level, aiding the student's learning process, allocating relevant resources, identifying exceptional students for scholarships and weak students who are likely to fail. This can be possible by processing and analyzing the data using various classification techniques. Decision trees are simple yet effective classification algorithms. One of their main advantages is that they provide human-readable rules of classification. *Decision Tree Induction* algorithm (Quinlan J. R.-1986) is used for classification by constructing a decision tree. The algorithm constructs decision tree recursively using depth– first divide and conquer approach. At any given node, to further split up the dataset towards identification of a class, the algorithm chooses the most suitable attribute based on the *information gain* value of the attributes. The information gain of an attribute is a measure of the ability of an attribute to minimize the information needed to classify the given entity in the resulting sub-trees.

Decision Tree Construction

Classification is the problem of identifying a category for the given instance whose category is unknown. The classifier tries to classify the given unknown instance based on the data it learns from the training set and features it sees in the instance for which prediction is to be done. This problem arises in various fields ranging from operation research, education, weather forecasting…etc for making decisions. There are wide varieties of classification problems in machine learning domain, all of which cannot be solved using one technique. Therefore, for proving that the results which we are getting from one kind of technique is good enough for us makes it indispensable that we compare the results with other techniques for the given problem. Whilst doing this, we come across the various performance aspects of the algorithm i.e. where it would fail and where it can do remarkably well in classifying the data. One of the most popular techniques for classifying data in Machine Learning domain is Decision Trees. The advantage of using Decision Trees in classifying the data is that they are simple to understand and interpret. Decision tree have been well studied and widely used in the knowledge discovery and decision support system. These trees approximate discrete valued target functions as trees and are widely used practical method for inductive reference. Each line present in the datasets is known as the instance. The instance contains the label and a vector of features present in it. The Decision Trees examines the feature of given instance and comes to a conclusion on what label to assign based on the values present for the various features of that particular instance. Each node in the decision tree is either a decision node or a leaf node. This classifier resembles tree data structure as each decision can have two outcomes, thereby making a binary decision tree that culminates in a label corresponding to each set of given features.

The data classification is a two step process. The first step is training the classifier using training data set. The second step involves predicting the labels for the unknown datasets (or testing datasets). This comes under the training step. Now in order to classify an unknown instance, the attribute values