Chapter 25
An Evaluation of C4.5 and Fuzzy C4.5 with Effect of Pruning Methods

Tayyeba Naseer
PMAS Arid Agriculture University, Pakistan

Sohail Asghar
COMSATs Institute of Information Technology, Pakistan

ABSTRACT

Classification is a supervised learning technique in data mining classify historical data. The decision tree is easy method for inductive inference. The decision tree induction process has three major steps – first complete decision tree is constructed to classify all examples in the training data, the second is pruning this tree to decrease misclassification rate and the third is processing the pruned tree to improve the classification. In this chapter, the empirical comparison of pruning the tree created by C4.5 and the fuzzy C4.5 algorithm. C4.5 and Fuzzy C4.5 decision tree algorithms are implemented using the JAVA language in Eclipse tool. In this chapter, first decision tree is built using C4.5 and Fuzzy C4.5 and five famous pruning techniques is used to evaluate trees and the comparison is achieved between pruning methods for refining the size and accuracy of a decision tree. Cost-complexity pruning produce the smaller tree with minimum increase in error for C4.5 and Fuzzy C4.5 decision trees.

INTRODUCTION

"Data mining is refers to as mining or extracting knowledge from huge data set", according to (Jiawei et *al.,* 2006). Data mining is the solicitation of explicit algorithms for mining patterns from data. The term data mining has habitually been used by statisticians, data analysts and the management information systems (MIS) communitie. Data mining can carry out in numerous terms or to some level diverse meaning from data mining such as data dredging, knowledge mining from data, pattern analysis, data archaeology and knowledge extraction. Machine learning is part of data mining. It rose over the late

DOI: 10.4018/978-1-5225-1759-7.ch025

1980s had made excessiveprogresses all through the 1990s and is expected to endure in the 21st century (John 1989). Clustering, data reduction, prediction, classification, association, or data transformation etc. Are several areas of data mining used to solve diverse types of problems. The development of effectual and dynamic data mining algorithms to process data; and the decreasing cost of computational impact, enabling the use of computationally tough methods for data exploration. Data mining is categorized in two types (Guoxiu 2005).

1. Supervised Learning
2. Unsupervised learning

Supervised learning is the task of machine learning of persuading a function of categorized data. The data used for inferring learning is called training data, it contains set of training examples. However, in supervised learning, examples are the pairsentailing of an input item and aessentialefficiency value which is also known as a managerial indication (Nitesh 2003). A supervised learning algorithm inspects the training data and produces a reliant function, which can be designed for recording inventive examples. Finest situation will allow the algorithm to fittingly fix the class labels for unseen examples. Classification is one of the instances of supervised learning techniques (Guoxiu 2005).

In unsupervised learning, all the observations are estimated to be elicited by hidden variables, that is, the clarifications are hypothesized to be at the end of the fundamental series, (Guoxiu 2005). In unsupervised learning, examples entail of input items only, they do not pair with output values. Clustering is an eminent example of unsupervised learning.

In our daily life and in the working environment we often solve numerous decisions – making problems. Conferring to the real-world, we make a decision based on our past experiences (Semra & Ersoy 2010). Machine learning field frequently lets computers implement or come up with new ideas for the precise result; various decision making methods exist, such as Decision Trees, Bayesian learning and Artificial Neural Networks. The decision tree methods are the emphasis of this investigation. Respectively each technique has its own advantages and drawbacks. The decision tree is one of the foremost use machines learning method for making decisions in pattern recognition (Guoxiu 2005).

The decision tree is a figurativemethodology used in numerous regions because of its benefits, more precise, effective, influential and easy for data preparation and also easy to understand for non – practical peoples (Quinlan 1986), (Wang, Yeung & Tsang 2001) and (Semra & Ersoy 2010). The supplementary benefit is that it can categorize numerical and categorical both types of data.

The first symbolic inductive learning algorithm was CLS [2], it is the prototype of decision tree, and then various additional decision tree algorithms, e.g. ID3, CART, C4.5, SPRINT, SLIQ and BOAT were proposed. C4.5 and CART (a beneficiary of ID3) is the two renowned and extensively used algorithms. CART was developed by Statisticians, while C4.5 (Quinlan 1986) was developed by a computer scientist in the field of machine learning. Decision-tree inductions comprises of three main phases:

Phase 1. Creating the complete decision tree to classify all the training data instances
Phase 2. Pruning the decision tree to show statistical constancy
Phase 3. And tendering out the pruned tree to raise understandability.

Decision Tree is also known as a nested hierarchy of branches like tree because it has different levels for indicating the information, and each branch, division demonstrates the features of the dataset using