DISTRIBUTED FUZZY DECISION TREES FOR BIG DATA

M.Anitha, S.Divya Bharathi, K.Ilakiya, R.Keerthika,

1 Student , Computer Science and Engineering, Saranathan college of Engineering, Tamilnadu, India
2 Student, Computer Science and Engineering, Saranathan college of Engineering, Tamilnadu, India
3 Student, Computer Science and Engineering, Saranathan college of Engineering, Tamilnadu, India
4 Student, Computer Science and Engineering, Saranathan college of Engineering, Tamilnadu, India

ABSTRACT

Fuzzy Decision Trees (FDTs) have shown to be an effective solution in the framework of fuzzy classification. The approaches proposed so far to FDT learning, however, have generally neglected time and space requirements. In this paper, we propose a distributed FDT learning scheme shaped according to the Map Reduce programming model for generating both binary and multi-way FDTs from big data. The scheme relies on a novel distributed fuzzy discretise that generates a strong fuzzy partition for each continuous attribute based on fuzzy information entropy. The fuzzy partitions are therefore used as input to the FDT learning algorithm, which employs fuzzy information gain for selecting the attributes at the decision nodes. We have implemented the FDT learning scheme on the Apache Spark framework. We have used ten real-world publicly available big datasets for evaluating the behaviour of the scheme along three dimensions. Performance in terms of classification accuracy, model complexity and execution time, Scalability varying the number of computing units. Ability to efficiently accommodate an increasing dataset size. We have demonstrated that the proposed scheme turns out to be suitable for managing big datasets even with modest commodity hardware support.

Keyword : - *Fuzzy ,Data Mining ,Preprocessing ,Clustering, Decision tree, Big data, etc*

1. INTRODUCTION

Decision trees are widely used classifiers, successfully employed in many application domains such as security assessment, health system and road traffic congestion. The popularity of decision trees is mainly due to the simplicity of their learning schema. Further, decision trees are considered among the most interpretable classifiers that is, they can explain how an output is inferred from the inputs. Finally, the tree learning process usually requires only a few parameters that must be adjusted.

In a decision tree, each internal (non-leaf) node denotes a test on an attribute, each branch represents outcome of the test, and each leaf (or terminal) node holds a class label. Several works have exploited the possibility of integrating decision trees with the fuzzy set theory to deal with uncertainty leading to the so-called fuzzy decision trees (FDTs). Unlike Boolean decision trees, each node in FDTs is characterized by a fuzzy set rather than a set. Thus, each instance can activate different branches and reach multiple leaves. Thus, a binary split tree is generally deeper and sometimes harder to interpret than a multi-way split tree. Further, in some domain, multi-way splits seem to lead to more accurate trees but, since multi-way splits tend to fragment the training data very quickly they generally need larger data size in order to work effectively.

1.1 EXISTING SYSTEM

This system develops two protocols for privately evaluating decision trees and random forests. We operate in the standard two-party setting where the server holds a model (either a tree or a forest), and the client holds an input (a feature vector). At the conclusion of the protocol, the client learns only the model's output on its input and a few generic parameters concerning the model; the server learns nothing. The first protocol we develop provides security against semi-honest adversaries. We then give an extension of the semi-honest protocol that is

robust against malicious adversaries. We implement both protocols and show that both variants are able to process trees with several hundred decision nodes in just a few seconds and a modest amount of bandwidth.

1.2 PROPOSED SYSTEM

We have proposed a novel protocol for privacy-preserving classification of decision trees, and improved the performance of previously proposed protocols for general hyper plane-based classifiers and for the two specific cases of support vector machines and logistic regression. To derive the parameters for our decision tree classification protocol, we again improve the Support Vector Machines and Logistic Regression concept for classification. The Client and Server can pre-compute this data by themselves with the help of well-known computationally secure schemes with a trusted authority is not available. We present accuracy and runtime results for 2 classification benchmark datasets from the UCI repository.

2. SYSTEM DESIGN

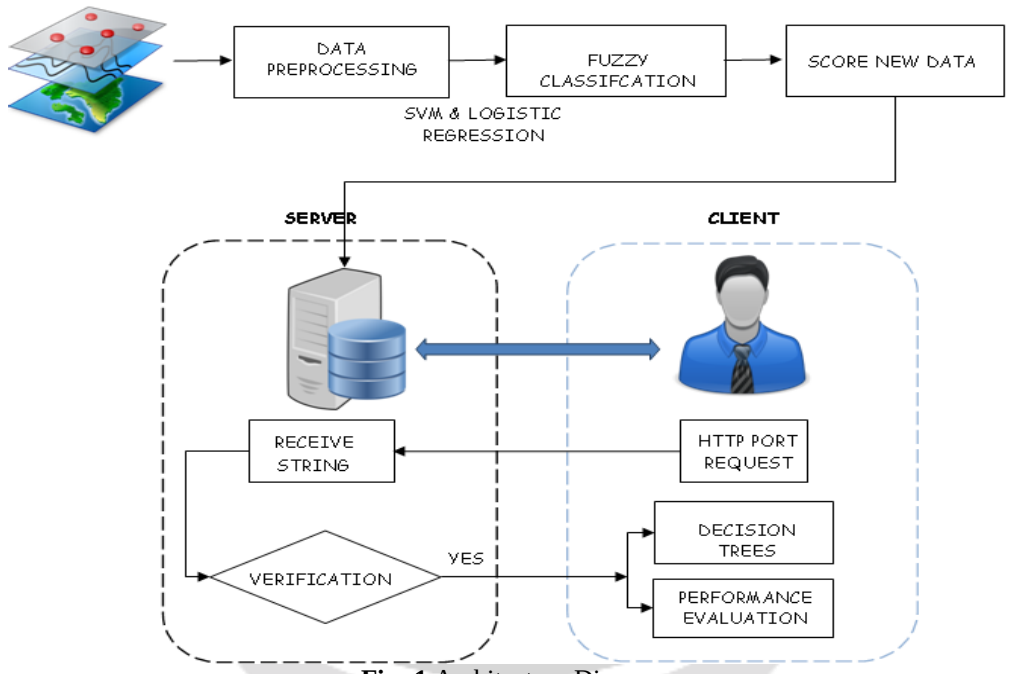

Fig -1 Architecture Digram

3. Modules Description

3.1 DATA LOADING

Data mining techniques are part of the business intelligence domain and apply specific algorithms for privacy-preserving classification of decision trees. We have taken two UCI benchmark datasets for classification protocols.

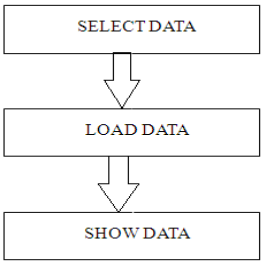

Fig -2 Data Loading

3.2 DATA PREPROCESSING

The initial step in mining analysis is data Pre-processing, here the raw data are pre-processed to discard null values, to identify and to prepare the data to enable its analysis. In the categorization sub-phase each record is analysed to identify high-level data and to extract meaningful information.

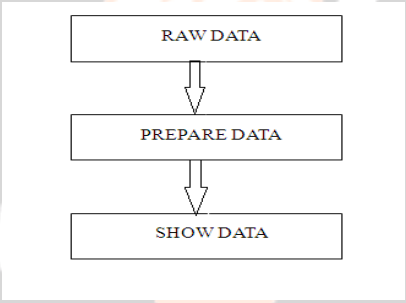

Fig -3 Data Preprocessing

3.3 DATA CLUSTERING

We cluster the data and analyse i.e., the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning.

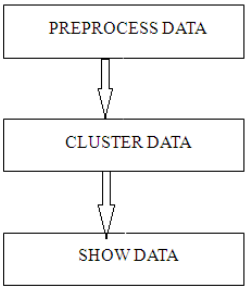

Fig -4 Data Clustering

3.4 DATA CLASSIFICATION

Data classification is the process of organizing data into categories for its most effective and efficient use. Here we have also implemented SVM and Logistic Regression for classification.Support vector machine (SVM) learning is a method for training classifiers based on different types of kernel functions – polynomial functions, radial basis functions etc. Support vector machines are a particular case of hyper plane-based classifiers.Logistic regression is a classifier that models the posterior probability of the class given the input features by fitting a logistic curve to the relationship between them. We predict that the instance belongs to the positive class, and otherwise we predict the instance belongs to the negative class

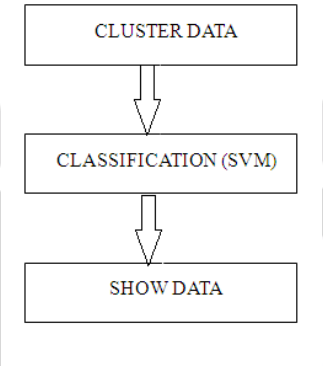

Fig -5 Data Classification

3.5 SERVER SIDE PROCESS

The Server receive the HTTP port request from the client. Based on the request, there is port number verification for data transfer. After verification, the server send the corresponding data to client for classification protocols. Then the fuzzy decision tree get constructed based on the classification result.

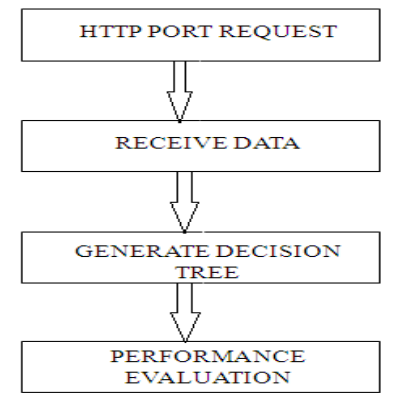

Fig -6 Server Side Process

3.6 CLIENT SIDE PROCESS

The Client send the HTTP port request to the Server to receive the data. After port number verification, the client received the data from the server. Based on the received data, the client generates Decision trees and Performance evaluation for classification protocols. Here the Client wants to score her data against a model in possession of

another party (Server). At the end of the protocol, Server learns nothing about Client data and Client learns as little as possible about Server model.We used an implementation of the classification and regression tree algorithm. Each internal node of the tree structure tests the value of a particular feature against a corresponding threshold and branches according to the result.

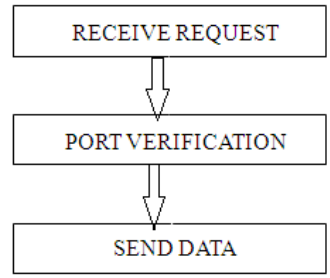

Fig -7 Client Side Process

4. CONCLUSIONS

We have proposed a distributed fuzzy decision tree (FDT) learning scheme shaped according to the Map Reduce programming model for generating both binary (FBDT) and multi-way (FMDT) FDTs from big data. We have first introduced a novel distributed fuzzy discretize , which generates strong fuzzy partitions for each continuous attribute based on fuzzy information entropy .Then, we have discussed a distributed implementation of an FDT learning algorithm, which employs the fuzzy information gain for selecting the attributes to be used in the decision nodes.

5. REFERENCES

[1]. Taking control of your health in the new era of personalized medicine.
[2]. Data mining: Concepts and techniques.
[3]. Data mining with decision trees: Theory and applications.
[4]. Recursive Partitioning and Regression Trees.