On Distributed Fuzzy Decision Trees for Big Data

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 MapReduce programming model for generating both binary and multi-way FDTs from big data. The scheme relies on a novel distributed fuzzy discretizer 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 behavior of the scheme along three dimensions: i) performance in terms of classification accuracy, model complexity and execution time, ii) scalability varying the number of computing units and iii) 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. Finally, we have used the distributed decision tree learning algorithm implemented in the MLLib library and the Chi-FRBCS-BigData algorithm, a MapReduce distributed fuzzy rule-based classification system, for comparative analysis.

CONCLUSIONS

We have proposed a distributed fuzzy decision tree (FDT) learning scheme shaped according to the MapReduce programming model for generating both binary (FBDT) and multiway (FMDT) FDTs from big data. We have first introduced a novel distributed fuzzy discretizer, 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. We have implemented the FDT learning scheme on the Apache Spark framework.

**SYSTEM REQUIREMENTS:**

**HARDWARE REQUIREMENTS:**

- System : Pentium IV 2.4 GHz.
- Hard Disk : 40 GB.
- Floppy Drive : 1.44 Mb.
- Monitor : 15 VGA Colour.
- Mouse : Logitech.
- Ram : 512 Mb

**SOFTWARE REQUIREMENTS:**

- Operating system : Windows 7/UBUNTU.
- Coding Language : Java 1.7 ,Hadoop 0.8.1
- IDE : Eclipse
- Database : MYSQL

**REFERENCES**
