Comparing the Computational Complexity and Accuracy of Classification Algorithms

S Beatrice
Xavier Institute of Engg.,
Mahim, Mumbai,
Maharashtra
09323258373
beatricejeevaraj@gmail.com

R Thirumahal
TSEC,
Bandra, Mumbai,
Maharashtra
9109820938174
r_thirumahal@yahoo.com

S P Raja
M.S University,
Tirunelveli,
Tamil Nadu
9109944259295
avemariaraja@gmail.com

ABSTRACT
Decision trees have been widely used for classification in Data mining. Number of decision tree algorithms has been developed in the past. In order to reduce the computational time the On Improving the efficiency of SLIQ (OIESLIQ) algorithm has been developed with an aim to reduce diversity of the decision tree at each split. In order to improve the accuracy, the paper proposes a novel approach (Pioneer classifier algorithm) to embark upon the other two algorithms.

Categories and Subject Descriptors
H Information systems
H.2 DATABASE MANAGEMENT
H.2.8 Database Applications
Subject: Data mining

General Terms
Algorithms, Performance

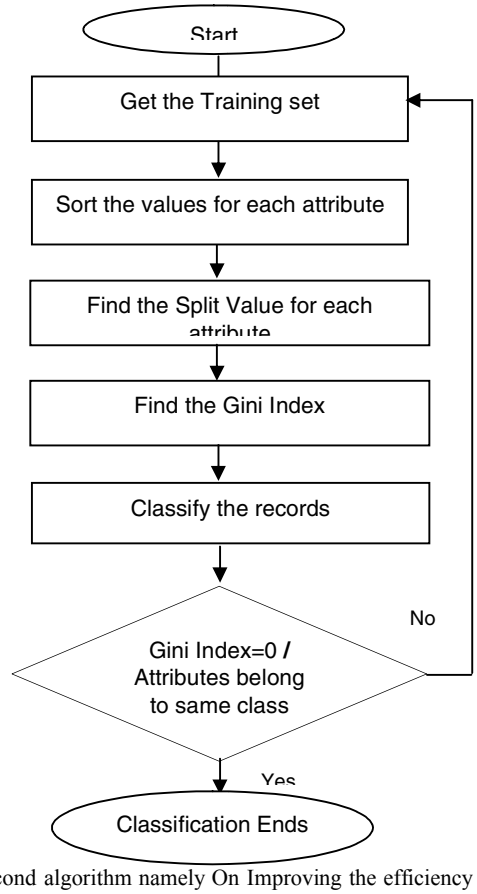
Keywords
Classification, Decision Trees;

1. INTRODUCTION
The success of computerized data management has resulted in the accumulation of huge amounts of data in several organizations. There is growing perception that analyses of these large data bases can turn "passive data" into active "actionable information". The recent emergence of Data Mining, or Knowledge Discovery in Databases, is a testimony of this trend.

Classification of data is one of the important tasks in data mining. Decision trees have been widely used for classification in Data mining. The three algorithms viewed for our paper aims at classifying the data using decision trees in which decision tree is optimal, simple to understand and interpret.

2. THEORY
In the first algorithm SLIQ [3] [Supervised Learning In Quest, where Quest is the Data Mining project at the IBM Almaden Research Center] is a decision tree [2] classifier developed by the Quest team to handle both numeric and categorical attributes. In this algorithm, split value is evaluated by ruling the average of attribute values. In SLIQ [3] algorithm, while evaluating the best split for each numeric attribute having n values, the Gini Index [1] has to be computed at n-1 possible splits. The total number of split points to be evaluated at a node with m attributes is m*(n-1), m being the number of attributes. This makes SLIQ [3] computationally complex for numeric attributes.

Design or flowchart is given below for all three algorithms



The second algorithm namely On Improving the efficiency of SLIQ algorithm (OIESLIQ) [1] has been developed to overcome this drawback. In this algorithm, split value is evaluated only at attribute values where the class information changes (after presorting the attribute values along with class information in a descending order) with the constraint that the attribute values also must be different. So the total number computations required per node would be equal to m*(n-1)

ICWET'10, February 26–27, 2010, Mumbai, Maharashtra, India.

only if the class information alternates at every value of the attribute (Worst Case) which is very unlikely in real time. In the best case the number of computations required per node would be equal to the product of number of class c and the number of attributes m (c*m much less than n*m). We ourselves have suggested an algorithm Pioneer Classifier Algorithm.

3. PROPOSED WORK

3.1 Description

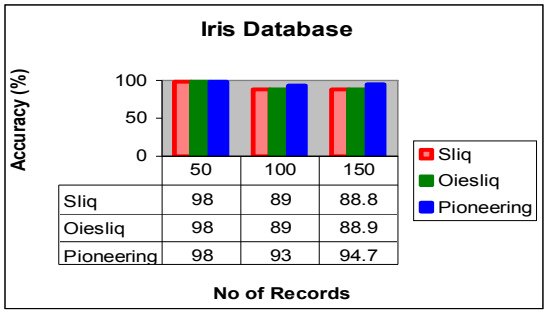
Pioneer Classifier Algorithm classifies the larger data sets more efficiently than smaller data sets. The classification is by classifying larger data sets into small groups using the greatest prime number divisor and Standard deviation of each group.

3.2 How does it works

The sample data set is taken into consideration. The data in the data sets are subjected into Pre Sorting as done in SLIQ algorithm. The total number of data in the data set is calculated. The Greatest Prime Number Divisor (GPND) of that number is found. The GPND which decides how many records have to be taken in a group. The standard deviation of a collection of numbers is a measure of the dispersion of the numbers from their expected (mean) value. The measure of deviation of numerical data within the group is calculated. The split value is estimated as the sum of the least data value in the group and the Standard deviation of the group. Then for each split value gini index is calculated.

4. RESULTS & COMPARISON

In order to prove Accuracy & computational time, the Iris database has been taken.

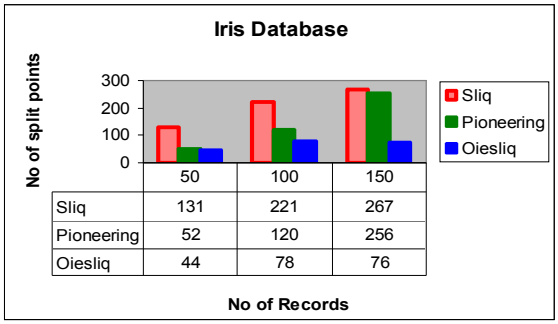


	50	100	150
Sliq	98	89	88.8
Oiesliq	98	89	88.9
Pioneering	98	93	94.7

Figure 2. Comparison of Three algorithms by using Iris database with all set of Records in terms of Accuracy

From the above graph, we can observe that the percentage of accuracy has been improved for all set of records for Pioneer Classifier algorithm as compared with other two algorithms.

To prove Computational time, the following graph shows the relationship between Number of records and the number of split points for all three algorithms.



	50	100	150
Sliq	131	221	267
Pioneering	52	120	256
Oiesliq	44	78	76

Figure 3. Comparison of Three algorithms by using Iris database with all set of Records in terms of split points

From the above graph, we can observe that the number of split points has been decreased for all set of records for OIESLIQ algorithm as compared with other two algorithms.

5. CONCLUSION

"Change is the only thing that never changes" is a prominent say. So, we would like to craft innovative and ground-breaking changes to our classification algorithm that would still stimulate the efficiency of the classification algorithm and trim down the number of split points as much as possible, so as to reduce the computational overhead to a greater extent.

6. REFERENCES

[1] Chandra. B and Varghese P. 2007. On Improving Efficiency of SLIQ Decision Tree Algorithm. Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, IEEE, August 12-17.

[2] Han J., Kamber M.. 2006. Data Mining: Concepts and Techniques", Morgan Kaufmann.

[3] Mehta.M, Agrawal.R and Rissanen.J, 1996. SLIQ: A Fast Scalable Classifier for Data Mining. In Proceedings of the 5th International Conference on Extending Database Technology, Avignon, France, Mar. 1996.