Generating Classification Trees for Small Disjuncts using Incremental GAs

Magda Fayek
Assoc. Prof.
Cairo Univ, Faculty of Eng.
Computer Eng. Dept., Egypt
+2 0101589411
magdafayek@gmail.com

Amira Samy
Technical Inspector, Engineer
Academy for Scientific Research
Kasr El Ainy Str,Cairo, Egypt
+2 0122518202
amtaleb1@yahoo.com

Nevin Darwish
Professor
Cairo Univ. Faculty of Eng.
Computer Eng. Dept., Egypt
+2 0122247364
ndarwish@eng.cu.edu.eg

ABSTRACT

In this paper an Incremental GA technique is proposed to solve the problem of small disjuncts in classification trees. It is once applied on the disjuncts sorted in ascending order and once in descending order with respect to their coverage.

Both versions of the technique have been tested using benchmark datasets and the results are compared with those of other classification techniques.

Categories and Subject Descriptors

I.2.8 [**Computing Methodologies**]: Artificial Intelligence – *Heuristic methods*

General Terms: Algorithms

Keywords

Incremental Genetic Algorithms, small disjuncts, data mining, C4.5. classification tree, rule pruning

1. THE PROPOSED TECHNIQUE

The C4.5 decision tree induction algorithm [1] is applied to induce a pruned tree which is then transformed into a set of rules (disjuncts). A disjunct is considered small if its coverage is less than a certain threshold value S. To evolve the corresponding classification rules using the proposed incremental GA small disjuncts are collected and sorted in ascending / descending order for the incremental ascending / incremental descending technique, respectively. Then the **incremental GA** proceeds as follows: Examples of the first disjunct are used to evaluate the fitness function during the GA run that evolves the classification rules of its corresponding sub-classes. One GA run is performed for each sub-classes. At the end of those GA runs the best classification rules are integrated to form the set of rules used in building the C4.5 tree for that small disjunct. Then the examples of the next small disjuncts are added at each following iteration to evolve classification rules for each of those small disjuncts. For each iteration after the first, the initial GA population consists of 50% of the best rules (chromosomes) produced by the previous GA and 50% of randomly generated chromosomes.

Finally, at testing, if a test sample is covered by some large disjunct it is classified accordingly, otherwise it is classified according to the C4.5 tree deduced from the corresponding classification rules evolved by the incremental GA.

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.

2. EXPERIMENTAL SETTINGS

The data sets used in **training** are public domain UCI' s data repository. The instances that had some missing values were removed from the data sets.

A chromosome in the GA includes a set of genes, each consisting of a condition on the designated attribute and an active bit that activates a gene with probability relative to its information gain.

The **fitness** is given by the accuracy [1]. **Testing** is performed using 10 fold Cross-validation. The max of ten independent runs is determined. The given results are the average for all datasets.

3. RESULTS

Fig. 1 and 2 compare the results of applying the proposed technique with previous techniques [3], for S= 3, 5, 7, 10 and 15 :

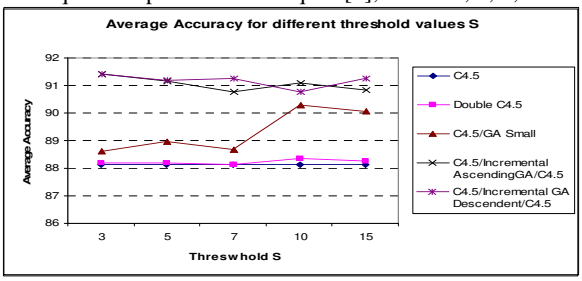

Figure 1: Average Accuracies (%)

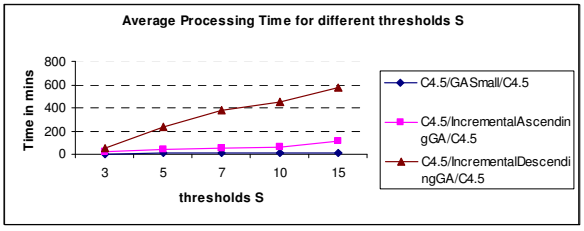

Figure 6: Average-Processing times

4. CONCLUSION

Results show that the proposed technique gives better accuracies, however, at the cost of higher processing. In addition, the descending technique is most stable giving average standard deviation of 2.49 relative to 3.90 for ascending incremental, 4.8, 4.9 and 4.5 for C4.5, double C4.5 and C4.5/GAsmall/C4.5.

5. REFERENCES

[1] J. R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publisher, 1993.

[2] Deborah R. Carvalho and Alex A. Freitas "A Hybrid Decision Tree/Genetic Algorithm Method for Data Mining", 2003.