RULE-MINER ALGORITHM: SHORTEST ROUTE JUDGING ALGORITHM USING ANT COLONY OPTIMIZATION

S. Rajesh Kumar[1], Dr. S. Murugappan[2]

[1]*Ph.D Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, (India)*

[2]*Associate Professor & Head, School of Computer Sciences, Tamil Nadu Open University, Chennai, Tamil Nadu,(India)*

ABSTRACT

This document is prepared to suggest an algorithm for more consistent rule discovery called Rule-Miner. Rule-Miner's goal is expected to extract and classifying and cataloguing rules from data. This algorithm is developed with notion observed from behavior of real world colonies of ant. We have compared the Rule-Miner's Performance against C4.5 algorithm on five domain data sets. The comparison results suggests the following outcomes, a) Rule-Miner shows more accuracy than C4.5 and b) Rule-Miner creates simple rules.

Keywords: Ant Colony Optimization, Data Set, Rule-Miner

I. OVERVIEW

In core, the cataloguing process contains a class for each rule as a class out of a set of predetermined classes, based on the values of few traits (called forecaster traits) for the case. From analysis, it is evidentially shown that the area of data mining is gaining lot of interest to show comprehensive and simple information to user. So that they can decide more effectively based on the information provided by the system.

Generally in data mining, the discovery of information is mentioned in IF THEN estimation as follows: IF <criteria> THEN <new rule>. The <criteria> section (precursor) of the new rule consist a analytical mixture of forecaster traits, in the shape of: term1 & term2 &.... Each term consist of three values <trait, worker, result>, such as <Relation = Father>. The <class> part (in THEN part) of the new rule Consist of the class expected for cases (records) whose forecaster traits obeys the <criteria> part of the new rule.

Based on our analysis, the usage of Ant Colony Algorithms as one of the way to classify new rules. There are closed areas in data mining with respect to methods for classification of rules. Essentially, data mining algorithm done by Ant colony optimization is different from procedure proposed in this document. There is few other proposals using Ant Colony Optimization which actually learns fuzzy control rules but it is not in the scope of data mining.

We hope the creation of Ant Colony algorithms for data mining is a significant research area due to the following reasons. An Ant Colony organization comprises dwarf workers (ants) that work together with one

another to complete an evolving, combined behavior for the organization as a whole, constructing a robust system proficient of judging quality solutions for issues in bigger search space.

II. ASSESSMENT RESULTS – 1

We have evaluated Rule-Miner across five public-domain data sets. The main characteristics of the data sets used in our experiment are summarized in Table 1. The first column of this table identifies the data set, whereas the other columns indicate, respectively, the number of cases, number of categorical traits, number of continuous traits, and number of classes of the data set.

As mentioned earlier, Rule-Miner discovers rules referring only to categorical traits. Therefore, continuous traits have to be discretized as a pre-processing step. This discretization was performed by the C4.5-Disc discretization algorithm. This algorithm simply uses the C4.5 algorithm for discretizing continuous traits.

Data set	#cases	#categ. trait.	#contin. trait.	#classes
Lung cancer (Chicago)	383	10	0	2
lung cancer (Dallas)	572	0	10	3
Tic-tac-toe	847	10	0	2
Dentistry	247	33	0	5
Diarrhoea	144	15	7	3

Table 1: Data Sets Used in Our Experiments

We have evaluated the performance of Rule-Miner by comparing it with C4.5 (Quinlan, 1993), a well-known rule induction algorithm. Both algorithms were trained on data discretized by the C4.5-Disc algorithm, to make the comparison between Rule-Miner and C4.5 fair.

The comparison was carried out across two criteria, namely the predictive accuracy of the discovered rule sets and their simplicity, as discussed in the following.

Predictive accuracy was measured by a 10-fold cross-validation procedure (Weiss & Kulikowski, 1991). In essence, the data set is divided into 10 mutually exclusive and exhaustive partitions. Then a classification algorithm is run 10 times. Each time a different partition is used as the test set and the other 9 partitions are used as the training set. The results of the 10 runs (accuracy rate on the test set) are then averaged and reported as the accuracy rate of the discovered rule set.

III. ASSESSMENT RESULTS – 2

The results comparing the accuracy rate of Rule-Miner and C4.5 are reported in Table 2. The numbers after the "±" symbol are the standard deviations of the corresponding accuracy rates. As shown in this table, Rule-Miner discovered rules with a better accuracy rate than C4.5 in four data sets, namely Chicagolung cancer, Dallaslung

cancer, Diarrhoea and Liver disease. In two data sets, Chicagolung cancer and Liver disease, the difference was quite small. In the other two data sets, Dallaslung cancer and Diarrhoea, the difference was more relevant. Note that although the difference of accuracy rate in Dallaslung cancer seems very small at first glance, this holds only for the absolute value of this difference. In reality the relative value of this difference can be considered relevant, since it represents a reduction of 20% in the error rate of C4.5. ((96.04 – 95.02)/(100 – 95.02) = 0.20)

On the other hand, C4.5 discovered rules with a better accuracy rate than Ant Miner in the other two data sets. In one data set, Dentistry, the difference was quite small, whereas in the Tic-tac-toe the difference was relatively large. (This result will be revisited later.) Overall one can conclude that Rule-Miner is competitive with C4.5 in terms of accuracy rate, but it should be noted that Rule-Miner's accuracy rate has a larger standard deviation than C4.5's one.

Data Set	Rule-Miner's accuracy rate (%)	C4.5's accuracy rate (%)
Lung cancer (Chicago)	76.53±11.88	73.34 ± 3.21
Lung cancer (Dallas)	96.04 ± 2.80	95.02 ± 0.31
Tic-tac-toe	73.04 ± 7.60	83.18 ± 1.71
Dentistry	86.55 ± 6.13	89.05 ± 0.62
Diarrhoea	90.00 ± 9.35	85.96 ± 1.07
Liver disease (Toronto)	59.67 ± 7.52	58.33 ± 0.72

Table 2: Accuracy Rate of Rule-Miner vs. C4.5

We now turn to the results concerning the simplicity of the discovered rule set. This simplicity was measured, as usual in the literature, by the number of discovered rules and the total number of terms (criteria's) in the antecedents of all discovered rules.

The results comparing the simplicity of the rule set discovered by Rule-Miner and by C4.5 are reported in Table 3. Again, the numbers after the "±" symbol denote standard deviations. As shown in this table, in five data sets the rule set discovered by Rule-Miner was simpler – i.e. it had a smaller number of rules and terms – than the rule set discovered by C4.5. In one data set, Chicagolung cancer, the number of rules discovered by C4.5 was somewhat smaller than the rules discovered by Ant Miner, but the rules discovered by Rule-Miner was simpler (shorter) than the C4.5 rules. To simplify the analysis of the table, let us focus on the number of rules only, since the results for the number of terms are roughly analogous. In three data sets the difference between the number of rules discovered by Rule-Miner and C4.5 is quite large, as follows.

In the Tic-tac-toe and Dentistry data sets Rule-Miner discovered 8.5 and 7.0 rules, respectively, whereas C4.5 discovered 83 and 23.2 rules, respectively. In both data sets C4.5 achieved a better accuracy rate. So, in these two data sets Rule-Miner sacrificed accuracy rate to improve rule set simplicity.

This seems a reasonable trade-off, since in many data mining applications the simplicity of a rule set tends to be even more important than its accuracy rate. Actually, there are several rule induction algorithms that were explicitly designed to improve rule set simplicity, even at the expense of reducing accuracy rate (Bohanec & Bratko, 1994; Brewlow & Aha, 1997; Catlett, 1991).

In the Liver disease data set Rule-Miner discovered 9.5 rules, whereas C4.5 discovered 49 rules. In this case the greater simplicity of the rule set discovered by Rule-Miner was achieved without unduly sacrificing accuracy rate – both algorithms have similar accuracy rates, as can be seen in the last row of Table 1.

There is, however, a caveat in the interpretation of the results of Table 3. The rules discovered by Rule-Miner are organized into an ordered rule list

	No. of rules		No. of terms	
Data set	Rule-Miner	C4.5	Rule-Miner	C4.5
lung cancer (Chicago)	7.20 ± 0.60	6.2 ± 4.20	9.80 ± 1.47	12.8 ± 9.83
lung cancer (Dallas)	6.20 ± 0.75	11.1 ± 1.45	12.2 ± 2.23	44.1 ± 7.48
Tic-tac-toe	8.50 ± 1.86	83.0 ± 14.1	10.0 ± 6.42	384.2 ± 73.4
Dentistry	7.00 ± 0.00	23.2 ± 1.99	81.0 ± 2.45	91.7 ± 10.64
Diarrhoea	3.40 ± 0.49	4.40 ± 0.93	8.20 ± 2.04	8.50 ± 3.04
Liver disease (Toronto)	9.50 ± 0.92	49.0 ± 9.4	16.2 ± 2.44	183.4 ± 38.94

Table 3: Simplicity of Rule Sets Discovered by Rule-Miner vs. C4.5

This means that, in order for a rule to be applied to a test case, the previous rules in the list must not cover that case. As a result, the rules discovered by Rule-Miner are not as modular and independent as the rules discovered by C4.5. This has the effect of reducing a little the simplicity of the rules discovered by Rule-Miner, by comparison with the rules discovered by C4.5. In any case, this effect seems to be quite compensated by the fact that, overall, the size of the rule list discovered by Rule-Miner is much smaller than the size of the rule set discovered by C4.5. Therefore, it seems safe to say that, overall, the rules discovered by Rule-Miner are simpler than the rules discovered by C4.5, which is an important point in the context of data mining.

Taking into account both the accuracy rate and rule set simplicity criteria, the results of our experiments can be summarized as follows.

In three data sets, namely Dallaslung cancer, Diarrhoea and Liver disease, Rule-Miner discovered a rule set that is both simpler and more accurate than the rule set discovered by C4.5. In one data set, Chicagolung cancer, Rule-Miner was more accurate than C4.5, but the rule sets discovered by Rule-Miner and C4.5 have about the same level of simplicity. (C4.5 discovered fewer rules, but Rule-Miner discovered rules with a smaller number of terms.)

Finally, in two data sets, namely Tic-tac-toe and Dentistry, C4.5 achieved a better accuracy rate than Rule-Miner, but the rule set discovered by Rule-Miner was simpler than the one discovered by C4.5.It is also

important to notice that in all six data sets the total number of terms of the rules discovered by Rule-Miner was smaller than C4.5's one, which is a strong evidence of the simplicity of the rules discovered by Rule-Miner.

IV. CONCLUSION

These results were obtained for a Pentium II PC with clock rate of 333 MHz and 128 MB of main memory. Rule-Miner was developed in C++ language and it took about the same processing time as C4.5 (on the order of seconds for each data set) to obtain the results. It is worthwhile to mention that the use of a high-performance programming language like C++, as well as an optimized code, is very important to improve the computational efficiency of Rule-Miner and data mining algorithms in general. The current C++ implementation of Rule-Miner is about three orders of magnitude (i.e., thousands of times) faster than a previous Mat Lab implementation.

REFERENCES

[1] Association Rule Mining: Models and Algorithms by Chengqi Zhang, Shichao Zhang - 10 Apr 2002

[2] Ant Colony Optimization by Marco Dorigo, Thomas Stützle - MIT Press, 2004

[3] Data Mining: Concepts and Techniques by Vikram Pudi - 12 Jan 2009

[4] MarcoDorigo, ChristianBlum: Ant Colony Optimization Theory: A Survey. Elsevier, Theoretical Computer Science 344(2005) pp:243-278

[5] Nada. M.A. AI Salami: Ant Colony Optimization Algorithm. Ubicc Journal, Volume 4, Number 3, August 2009 pp: 823-826

[6] David Martens, Manu De Backer, Raf Haesen: Classification with Ant Colony Optimization. IEEE Transactions on Evolutionary Computation, Vol.11, No.5, October 2007 pp: 651- 665

[7] Parpinelli R.S., Lopes H.S., Freitas, A.A.: Data Mining with an Ant Colony Optimization Algorithm. Evolutionary Computation, IEEE Transactions on Vol: 6 , Issue: 4 pp: 321-332

[8] Bing Liu, Wynne Hsu, Yiming Ma: Integrating Classification and Associaciation Rule Mining. American Association for Artificial Intelligence, Vol.1, September 1998

[9] Nan Jiang, Le Gruenwald: Research issues in data stream association rule mining. ACM SIGMOD Record, Volume 35 Issue 1, March 2006 pp:14 -19

[10] Ernst Kretschmann, Wolfgang Fleischmann, Rolf Apweiler: Automatic rule generation for protein annotation with t he C4.5 data mining algorithm applied on SWISS-PROT. Oxford Journals Bioinformatics, Vol.17 (10), July 2001 pp:920-926

[11] Jochen Hipp, Ulrich Guntzer, Gholamreza Nakhaeizadeh: Algorithms for association rule mining- a general survey and comparison. ACM SIGKDD Explorations, Vol 2 Issue 1, June, 2000 pp: 58-64

[12] Christian Hidber: Online association rule mining. ACM SIGMOD Record, Vol.28 Issue 2, June 1999 pp: 145-156

[13] Hanbing Liu, Baisheng Wang: An Association Rule Mining Algorithm Based on a Boolean Matrix. Data Science Journal, vol. 6 September 2007 pp:559-565