International Journal of Mathematical Engineering and Science
ISSN : 2277-6982 Volume 1 Issue 1
http://www.ijmes.com/ https://sites.google.com/site/ijmesjournal/

Mining Interesting Classification Rules: An Evolutionary Approach

Basheer Mohamad Al-Maqaleh

Faculty of Computer Sciences & Information Systems
Thamar University, Yemen.
basheer.almaqaleh.dm@gmail.com

Abstract. Automated discovery of rules is, due to its applicability, one of the most fundamental and important method in Knowledge Discovery in Databases(KDD). It has been an active research area in the recent past. This paper presents a classification algorithm based on Evolutionary Approach(EA) that discovers interesting classification rules in the form **If** P **Then** D. A flexible encoding scheme, genetic operators and a suitable fitness function to measure the goodness of rules are proposed for effective evolution of rule sets. The proposed algorithm is validated on several datasets of UCI data set repository and the experimental results are presented to demonstrate the effectiveness of the proposed scheme for automated rule mining.

Keywords: Knowledge Discovery in Database, Data Mining, Evolutionary Algorithms.

1 Introduction

With rapidly increasing capabilities of accessing, collecting, and storing data, Knowledge Discovery in Databases (KDD) has emerged as a new area of research in computer science. The objective of KDD systems is to extract implicitly hidden, previously unknown, and potentially useful information and knowledge from databases. Data Mining is a core stage in the entire process of KDD which applies an algorithm to extract interesting patterns[7],[9]. Typically, the number of patterns generated from massive datasets is quite large, but only some of them are likely to be useful for the domain expert analyzing the data. The most effective way of reducing volume of discovered pattern is so called interestingness measures[6]. There are two types of such measures namely, objective and subjective measures. Objective measures are those that depend only on the structure of a pattern and which can be quantified by using statistical methods. On the other hand, subjective measures are based on the subjectivity and understandability of the user who examine the patterns[5],[11]. Classification systems are useful techniques in data mining, which

are supervised learning methods that induce a classification model from a database[1],[8],[10],[12],[13],[23]. The discovered knowledge is usually presented in the form of **If** P **Then** D classification rules because this method presents a high-level, symbolic knowledge presentation and contributes the comprehensibility of the discovered knowledge[10]. In any data mining task involving prediction discovered knowledge should have high predictive accuracy. Discovered knowledge should also be interesting to the user. Discovered knowledge may be highly accurate and comprehensibility, but it is uninteresting if it states the obvious or some pattern that was previously-known by the user[2],[4].

The Evolutionary Algorithms (EAs) are adaptive techniques that can be successfully used to solve complex search and optimization problems. They are based on the principles of genetics and Darwin's natural selection theory[8],[16],[17]. In essence, an EA maintains a population of "individuals", where each individual represents a candidate solution to the target problem. EAs are iterative generate-and-test procedures, where at each "generation" a population of individuals is generated and each individual has its "fitness" computed. The fitness of an individual is a measure of the quality of its corresponding candidate solution. The higher the fitness of an individual, the higher the probability that the individual will be selected to be a "parent" individual. Certain operators are applied to the selected parent individuals in order to produce "children" individuals. The important point is that, since the children are in general produced from parents selected based on fitness, the children (new candidate solutions) tend to inherit parts of the good solutions of the previous generation, and the population as a whole gradually evolves to fitter and fitter individuals (better and better solutions to the target problem)[14],[15]. Traditional rule generation methods, are usually accurate, but have brittle operations. EAs on the other hand provide a robust and efficient approach to explore large search space[25].

A Numerous attempts have been made to apply EAs in data mining to tackle the problem of knowledge extraction and classification. Several EA designs, for discovering classification rules, have been proposed in the literature[1],[2],[3],[4],[25]. In recent studies, a lot of variation in the basic structure of classification rules are suggested by researchers for the knowledge discovery [18],[19],[20],[21][22],[24],[27] .

This paper presents a classification algorithm based on Genetic Algorithm(GA) that discovers accurate, simple and interesting classification rules in the form **If** P **Then** D from datasets.

2 The Proposed GA Design

In this section GA approach is presented for the automated discovery of interesting classification rules as the underlying knowledge representation. The evolutionary system is able to acquire information from datasets and extract interesting classification rules for each available class, given the values of some attributes, called predicting attributes. The basic idea is to consider a population composed by

individuals each representing a single candidate rule, and to gradually improve the quality of these rules by constructing new fitter rules until a fixed maximum number of generations have been reached.

2.1 Genetic Representation

To solve an optimization problem, GAs start with the chromosome(string) representation of a parameter set. The search space, or the population, is a set of chromosomes on which genetic operators can perform. The proposed GA in this work follows Michigan's approach to represent rules which each individual is encoded as a single classification rule of the form **If** P **Then** D, where P is the rule antecedent and D is the rule consequent. The antecedent of this rule can be formed by a conjunction of at most n-1 attributes, where n is the number of attributes being mined. Each condition is of the form P*i* = V*ij*, where P*i* is the *i*th attribute and V*ij* is the *j*th value of the *i*th attribute's domain. The consequent consists of a single condition of the form D*k* = V*kl*, where D*k* is the *k*th goal attribute and V*kl* is the *l*th value of the *k*th goal attribute's domain[25]. A string of fixed size encodes an individual with n genes representing the values that each attribute can assume in the rule. This encoding is shown in Fig. 1.

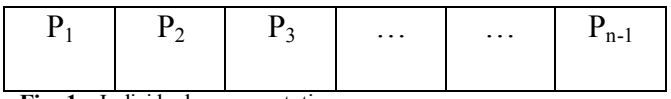

Fig. 1. Individual representation.

The genes are positional, i.e. the first gene represents the first attribute, the second gene represents the second attribute and so on. If an attribute is not present in the rule antecedent, the corresponding value in gene is ''#''. This value is a flag to indicate that the attribute does not occur in the rule antecedent. Hence, this encoding effectively represents a variable-length individual (rule). The decision part does not need to be coded into the chromosome, as it would be explained later.

2.2 Genetic Operators

Genetic operators are one of the most important components of GAs. They are being used to manipulate or recombine the genetic material of candidate rules and introduce new genetic material[3]. The proposed algorithm uses a well-known roulette wheel selection method as the fitness proportionate selection operator. An elitist reproduction strategy is used, where the best individual of each generation was passed unaltered to the next generation.

Crossover is the process by which genetic material from one parent is combined with genetic material from the second parent to produce potentially promising new offspring. The intention of this operation is to avoid the solution process to converge to local optimum. In the present work one point crossover is used as recombination

operator. Mutation operator randomly transformed the value of an attribute into another value belonging to the domain of that attribute.

2.3 Fitness Function

Fitness functions are used to measure the goodness of rules. The choice of the fitness function is very crucial as it leads the search by evolutionary algorithms towards the optimal solution. For the proposed system the Rule Interestingness (RI) measure[26] is used as a fitness function. RI is based on statistical properties of the rules(objective). As already mentioned, the discovered rules are of the form,
If P **Then** D, where
N_P Number of instances matching P.
N_D Number of instances matching D.
N_{BOTH} Number of instances matching both P and D.
N_{TOTAL} Total number of instances.
The fitness function is computed as per the following formula:

$$\text{Fitness} = N_{BOTH} - (N_P \times N_D / N_{TOTAL}) \quad \textbf{(1)}$$

Fitness measures the difference between the actual number of matches and the expected number if the left- and right-hand sides of the rule were independent. Generally the value of Fitness is positive. A value of zero would indicate that the rule is no better than chance. A negative value would imply that the rule is less successful than chance[9].

In the chess dataset if the rule **If** inline=1 $\wedge$ wr_bears_bk=2 **Then** class=safe is discovered. For this rule N_P =162, N_D=613, N_{BOTH}=162 and N_{TOTAL}=647. So the Fitness is computed using the formula(1), as under:

Fitness= 162- (162 × 613 / 647) =8.513. The Fitness value indicates that the rule can be expected to correctly predict 8.513 more correct classifications (on average) than would be expected by chance[9].

3 Computational Results

This section reports the results of computational experiments with some public domain data sets[28]. In the following experiments each decision (class) in a dataset is dealt with separately. During each run, the same decision (class) under consideration is assigned to all the individuals in the population. Assuming that the application domain has four classes we need to run GA four times i.e. in the first run GA would search for class1, in the second run for the class 2 and so on. The **Then** (decision) part of the rule does not need to be encoded into the individual. In effect, in a given run of the GA all individuals are searching for rules predicting the same class.

This approach simplifies the design of the proposed scheme and it is particularly natural when the user is not interested in a complete classification rule set (where different rules predict different classes), but is rather only interested on rules predicting a predetermined class [14]. Each data set was randomly partitioned into two parts with 2/3 of the instances used for training and 1/3 of the instances used for testing the quality of the discovered rules. The data-specific parameters and the parameters, which are encountered during the rule discovery are listed in the Table 1.

Table 1. Parameters used for the experiments.

Dataset	Population size	Maximum number of generations	Crossover rate	Mutation rate
Zoo	100	100	0.75	0.01
Balance	50	200	0.75	0.01
Nursery	80	500	0.75	0.01

The performance of the proposed algorithm on different datasets is demonstrated below:

Experiment 1:

The Zoo data set was used for this experiment. This dataset has 101 examples, 17 predicting attributes and a goal attribute, which can take 7 classes. The predicting attributes were nominal. Table 2 presents the final 7 rules discovered by the GA one rule for each class.

Table 2. Result for the Zoo dataset.

No.	Discovered Rules	Fitness
1	**If** Milk =1 **Then** Class=1	17.64
2	**If** Milk= 0 Venomous=0 **Then** Class=2	9.00
3	**If** Aquatic =1 ∧ Fins=1 **Then** Class=4	6.05
4	**If** Venomous=0 ∧ Tail=0 **Then** Class=6	3.49
5	**If** Aquatic=1 ∧ Breath=0 **Then** Class =7	3.03
6	**If** Catsize=0 ∧ Predator = 1 **Then** Class = 3	2.65
7	**If** Hair=0 ∧ Tail=0 **Then** Class = 5	1.70

It is also important to evaluate the performance of the rule set as a whole. As usual in the literature, this evaluation was done by measuring the accuracy rate on the test set, i.e. the ratio of the number of instances correctly classified over the total of instances in the test set. The accuracy rate for the discovered rule set was 99%.

Experiment 2:

This experiment was carried out on the Balance data set. This data set has 625 instances (49 **B**alanced, 288 **L**eft, 288 **R**ight), 4 predicting attributes and a goal attribute. The proposed scheme would discover the following classification rules(Table 3).

Table 3. Result for the Balance dataset.

No.	Discovered Rules	Fitness
1	**If** Left-Weight =1 **Then** Class=R	34.85
2	**If** Right-Distance= 1 **Then** Class=L	29.37
3	**If** Left-Distance=3 ∧ Right-Weight= 3 **Then** Class=B	2.46

Any way, it is interesting to evaluate the performance of the set of discovered rules as a whole, by measuring the accuracy rate, as done in the previous experiment. The accuracy rate of the discovered rule set was 89.99%.

Experiment 3:

The Nursery data set was used for this experiment This data set contains 12960 instances, 8 attributes and a goal attribute which can take 5 classes(not_recom, recommend, very_recom, priority, spec_prior). The attributes are all categorical. Table 4 shows the rules generated from this data set.

Table 4. Result for the Nursery dataset.

No.	Discovered Rules	Fitness
1	**If** health=not_recom **Then** Class=not_recom	21.48
2	**If** health=priorty **Then** Class=spec_prior	16.94
3	**If** has_nurs=proper **Then** Class=priority	11.62
4	**If** social=non_prob ∧ health= recommrnded **Then** Class= very_recom	9.67
5	**If** housing=convenient ∧ childern=1 **Then** Class=recommend	1.83

In this experiment the accuracy rate of the discovered rule set was 99.9%.

4 Conclusion and Future Work

In this paper, a GA approach to automated discovery of interesting classification rules is presented and evaluated. The proposed algorithm is an alternative to find classification rules with simple, concise, high predictive accuracy and interesting are based on the experimental results. The present work seems to be particularly effective in finding a concise set of comprehensible and interesting rules, since it discovers only a single rule for each class. Other data mining algorithms often discover several rules for a single class, which makes it difficult for the user to understand the numerous discovered rules. One direction for future research of developing a method of the distributed-population GA where each subpopulation is associated with a goal attribute value.

References

1. Barros, R.C., Basgalupp, M.P., Ferreira, A.C., Frietas, A.A.: Towards the Automatic Design of Decision Tree Induction Algorithms. In: GECCO(Companion Material),pp.567--574, Dublin, Ireland(2011)
2. Carvalho, D.R., Frietas, A.A.: A Genetic Algorithm for Discovering Small-Disjunct Rules in Data Mining. Applied Soft Computing, 2(1),75--88 (2002)
3. Vashishtha, J., Kumar, D., Ratnoo, S., Kundu, K.: Mining Comprehensible and Interesting Rules: A Genetic Algorithm Approach. International Journal of Computer Applications(09725-8887). 31(1), 39--46 (2011)
4. Carvalho D.R., Frietas A.A.: A Genetic Algorithm with Sequential Niching for Discovering Small-Disjunct Rules. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2002), pp. 1035--1042. New York (2002)
5. Frietas, A.A.: On Rule Interestingness Measures. Knowledge-Based System. 12(5-6), 309--315(1999)
6. Hilderman, R.J., Hamilton, H.J.: Applying Objective Interestingness Measures in Data Mining Systems. In: Principles of Data Mining and Knowledge Discovery Proceedings of the 4th European Conference PKDD'2000, Lecture Notes in Artificial Intelligence, vol. 1704, pp. 232 --241. Springer (2000)
7. Fayyad, U.M., Piatetsky-Sharpio, G., Smyth, P.: From Mining to Knowledge Discovery : An Overview. In: Fayyad, U.M. Piatetsky-Sharpio, G. Smyth. P., Uthurusany, R. (eds.)Advances in Knowledge Discovery and Data Mining , pp. 1--34. AAAI/MIT Press(1996)
8. Mitchell,T.: Machine Learning. McGraw-Hill(1997)
9. Bramer, M.: Principles of Data Mining. Springer-Verlag London Limited(2007)
10. Witten, I.H., Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations. Morgan Kaufmann(2000)
11. Liu, B., Hsu, W., Chen, S., Ma.Y.: Analysis the Subjective Interestingness of Association Rules. IEEE Intelligent Systems. 15(5), 47--55(2000)
12. Pazzani, M. J.: Knowledge Discovery from Data. IEEE Intelligent Systems, 10--12(2000)
13. Quinlan, R. :C4.5: Programs for Machine Learning. Morgan Kaufmann(1993)
14. Frietas, A.A.: Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer-Verlag, Berlin, Heidelberg (2002)
15. Deb, K. :Multi-Objectives Optimization using Evolutionary Algorithms. Wiley(2001)
16. Goldberg, D.E. :Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley(1989)
17. Triantaphyllou, E. Felici, G., (eds.). : Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques. Massive computing series, Springer, Heidelberg, Germany (2006)
18. Al-Maqaleh, B.M, Bharadwaj, K.K.: Genetic Programming Approach to Hierarchical Production Rule Discovery. In: Proceedings of the 5th International Conference on Databases and Data Mining (DBDM2005),vol.6, pp. 271--274. Istanbul, Turkey(2005)
19. Bharadwaj, K.K., Al-Maqaleh, B.M.: Evolutionary Approach for Automated Discovery of Censored Production Rules. In: Proceedings of the 8th International Conference on Cybernetics, Informatics and Systemics(CIS-2005), vol.10, pp.147--152.Krakow, Poland (2005)
20. Al-Maqaleh, B.M, Bharadwaj, K.K.: Genetic Programming Approach for Automated Discovery of Production Rules with Fuzzy Hierarchy. In: Proceedings of the National Conference on Methods and Models in Computing (NCM2C'2006), pp. 127--134. Jawaharlal Nehru University, New Delhi, India(2006)

21. Bharadwaj, K.K., Al-Maqaleh, B.M.: Evolutionary Approach for Automated Discovery of Augmented Production Rules. International Journal of Computational Intelligence. 3(4), 267--275(2006)
22. Al-Maqaleh, B.M, Bharadwaj, K.K.: Evolutionary Approach to Automated Discovery of Censored Production Rules with Fuzzy Hierarchy. In: Proceedings of the International Conference on Data Mining and Applications (ICDMA'2007), vol. 1, pp. 716-721. Hong Kong, China(2007)
23. Al-Maqaleh, B.M.: An Inductive Learning Algorithm for Censored Production Rule(CPR) Discovery. In: Proceedings of the International Conference on Informatics Systems, Technology, and Management, ICISTM, pp.231--236. IMT-Ghaziabad, India (2007)
24. Saroj, Bharadwaj, K.K. :A Parallel Genetic Algorithm Approach for Automated Discovery of Censored Production Rules. In: Proceedings of International Conference on Artificial Intelligence and Application (lASTED), pp. 435--441. ACTA Press, Innsbruck, Austria (2007)
25. Dehuri, S., Mall, R.: Predictive and Comprehensible Rule Discovery using a Multi Objective Genetic Algorithms. Knowledge Based Systems. 19, 413--421(2006)
26. Piatetsky-Shapiro, G.: Discovery, Analysis and Presentation of Strong Rules", In Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 229--248, AAAI Press(1991)
27. Yogita, Saroj, Kumar, D.: Rule +Exceptions: Automated Discovery of Comprehensible Decision Rules. IEEE International Advance Computing Conference(IACC2009), pp. 1479--1483. Patiala, India(2009)
28. UCI Repository of Machine Learning Databases, Department of Information and Computer Science University of California, 1994, http://www.ics.uci.edu/ ~mlearn/MLRepositry.html