A Genetic Algorithm for Discovering Classification Rules in Data Mining

Basheer M. Al-Maqaleh
Faculty of Computer Sciences and Information Systems, Thamar University, Yemen.

Hamid Shahbazkia
Faculty of Science and Technology, University of Algarve, Faro, Portugal.

ABSTRACT
Data mining has as goal to discover knowledge from huge volume of data. Rule mining is considered as one of the usable mining method in order to obtain valuable knowledge from stored data on database systems. In this paper, a genetic algorithm-based approach for mining classification rules from large database is presented. For emphasizing on accuracy, coverage and comprehensibility of the rules and simplifying the implementation of a genetic algorithm. The design of encoding, genetic operators and fitness function of genetic algorithm for this task are discussed. Experimental results show that genetic algorithm proposed in this paper is suitable for classification rule mining and those rules discovered by the algorithm have higher classification performance to unknown data.

General Terms
Knowledge Discovery in Databases (KDD), Data Mining, Machine Learning, Genetic Algorithm.

Keywords
Classification Rule, Genetic Operators, Fitness Function, Predictive Accuracy.

1. INTRODUCTION
Recent development in computer technology has significantly advanced the generation and consumption of data in our daily life. As a consequence, challenges, such as the growing data warehouses, the needs of intelligent data analysis and the scalability for large or continuous data volumes, are now moving to the desktop of business managers, data experts or even end users. Knowledge Discovery in Databases (KDD) and data mining, grounded on established disciplines such as machine learning, artificial intelligence and statistics, is dedicated to solving the challenges by exploring useful information and interesting knowledge from a massive amount of data, providing valuable support for decision making in industry, business, government, and science[1], [2]. Data mining can be used to solve a variety of tasks, including classification, regression, clustering, and association discovery [3]. A classification rule is a high-level knowledge representation of the form **If** P **Then** D , where P is a conjunction of predicting attribute values and D is the predicted class. This kind of knowledge representation has the advantage of being intuitively comprehensible to the user[4]. Mining classification rules usually utilizes supervised learning techniques that consist in discovering patterns in training data so that the resulting rules can be applied in the classification of other data. Classification rule mining aims to discover a small set of rules in the database that forms an accurate classifier[4]. Recently, techniques and algorithms for classification rule mining have been intensively studied due to the large variety of practical applications for them[1],[3], [5],[6],[7]. Genetic Algorithms (GAs) are methods based on biological mechanisms, such as, Mendel's laws and Darwin's fundamental principle of natural selection. GA process in an iteration manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. An evaluation function associates a fitness measure to every string indicating its fitness for the problem. Individuals evolve towards better and better individuals via selection procedure based on natural selection, i.e., survival of the fittest, and genetic operators, e.g., crossover and mutation. Hence, the better the quality of an individual, the higher the probability that its "genetic material" will be passed on to later generations of individuals[8]. The main advantage of GAs is their robustness, that is, once the problem is correctly modelled, the algorithm is able to explore the feasible region within the space of problem solutions, looking for the best global solution. In general the main motivation for using GAs in the discovery of high-level classification rules is that they perform a global search and cope better with attribute interaction than the greedy rule induction algorithms often used in data mining[4],[9]. In the recent years, there has been an increasing interest in involving the use of GA in data mining to tackle the problem of knowledge extraction and classification. Several GA designs, for discovering classification rules, have been proposed in the literature. A short review is presented here. A fixed encoding scheme is applied to the chromosomes and specific design is used for the mutation operator for GA designed by Fidelis et al.[10] to discover comprehensible classification rules. A GA with entropy based filtering bias to initial population for automated rule mining proposed by Kapila et al. [11]. A classification algorithm based on evolutionary approach that discovers comprehensible rules with exceptions in the form of censored production rules proposed by Bharadwaj & Al-Maqaleh [12], [13]. Dehuri & Mall [9] proposed a multi-objective algorithm for mining highly predictive & comprehensible classification rules from large databases. Goplan et al. [14] proposed a GA approach as a post-processing stage to discover accurate and interesting classification rules. Carvalho & Frietas [15] proposed a hybrid approach for rule discovery that combine decision trees and GA to automated discovery of small disjuncts rules. An accuracy-based learning system called DTGA(decision tree and GA) that aims to improve prediction accuracy over many classification problem proposed by Sarkar et al. [16]. Al-Maqaleh [17], [18] proposed a GA approach to automated

* Basheer Mohamad Al-Maqaleh –Corresponding author- is currently doing his Post-Doctoral Research at Faculty of Science and Technology, University of Algarve, Faro, Portugal. (e-mail: basheer.almaqaleh.dm@gmail.com).

discovery of classification rules from nominal datasets. The work present in this paper proposes a new GA that can be applied in classification rule mining from continuous datasets. A crucial aspect of data mining is that the discovered knowledge should be somehow interesting. Rule interestingness has both an objective (data-driven) and a subjective (user-driven) aspect[19]. This paper focuses on the objective aspect of rule interestingness. The proposed algorithm is used to mine classification rule with three measures like accuracy, coverage (completeness) and comprehensibility. Using these three measures, some previously unknown, easily understandable and compact rules can be generated.

The rest of the paper is organized as follows: Section 2 presents new GA design that includes encoding scheme, genetic operators applied and fitness function. The datasets used for validation are presented in Section 3 besides the analysis and summarization of the computational results. Conclusions and future directions are given in Section 4.

2. THE PROPOSED GA APPRAOCH

This section describes a GA that evolves a population of individuals, where each individual represents a classification rule. More precisely, each individual represents the antecedent (**If** part) of a classification rule. The consequent (**Then** part) of the rule is not encoded in the genome. Rather, it is fixed for a given GA run, so that in each run all the individuals represent rules with the same consequent. The remaining details of GA are as given below.

2.1 Individual Representation

In the proposed algorithm the genome of an individual consists of a conjunction of conditions composing a given rule antecedent. An individual is encoded as a set of *n* conditions, where *n* is the number of predictor attributes(Figure 1). Each gene represents a rule condition of the form A_i Op_i V_{ij}, where[4]:

- A_i denotes the *i*-th predictor attribute;
- Op_i denotes the comparison operators-e.g., "=" for categorical attributes; ",<=" or ">" for continuous attributes- used in the *i*-th conditions;
- Val_{ij} denotes the *j*-th value of the domain of $Attr_i$.

A_i Op_i Val_{ij}	...	A_n Op_n Val_{nj}

Figure 1. Structure of the genome of an individual

Not that this approach implicitly assumes a positional encoding of attributes in the genotype, i.e. 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. This kind of representation leads to relatively short individuals and simplifies the design of genetic operators. Figure 2 shows an example of an individual whose genotype has the general structure of Figure 1.

Marital-status= married	*#*	*Age > 30*

Figure 2. Genotype

In this example the data being mined has only three attributes, so three conditions are encoded in the genotype . This genotype is decoded into the following rule antecedent **If** *Marital-status= married* $\wedge$ *Age > 30*. The consequent (**Then** part) of the rule, which specifies the predicted class, is not represented in the genome. Rather, it is fixed for a given GA run, so that all individuals have the same rule consequent during all that run.

2.2 Genetic Operators

Genetic operators are being used in GAs to maintain genetic diversity by introducing new genetic material and to manipulate or recombine the genetic material of candidate rules. In the proposed system tournament selection, with tournament size of 5, two-point crossover with crossover probability of 90%, and mutation probability of 1% are used. Furthermore, we use elitism with an elitist factor of 1 - i.e. the best individual of each generation is passed unaltered into the next generation. The crossover (or recombination) operator essentially swaps genetic material between two "parents" creating two new "child" individuals. These two crossover points are randomly generated such that the genes between these two points are swapped between the two individuals, yielding the new child individuals. Note that crossover points can fall only between genes, and not inside a gene. Hence, crossover swaps entire rule conditions between individuals, but it cannot produce new rule conditions. The mutation is an operator that acts on a single individual at a time. It helps to find the global optimal solution of the problem, searching new areas. This operator randomly transforms the value of an attribute into another (different) value belonging to the domain of that attribute.

2.3 Fitness Function

The fitness function is used to evaluate how well an individual solves the target problem, and it is responsible for determining which individuals will reproduce and have parts of their genetic material passed onto the next generation. The better the fitness of an individual, the higher the probability of that individual being selected for reproduction, crossover, and mutation operations. Before we can define the fitness function, it is necessary to recall a few basic concepts on classification rule evaluation. The discovered rule(R) in the form **If** P **Then** D; where P is antecedent part and D is the consequent part.

Confidence: The confidence(conf) of the rule R can be defined as:

$$\text{conf} = \frac{|P \wedge D|}{|P|} \quad (1)$$

In classification problems, confidence measure is defined as the ratio of the number of examples in P that are correctly classified as decision class of D and the number of examples in P.

Coverage: The coverage(cove) of the rule R can be defined as:

$$\text{cove} = \frac{|P \wedge D|}{|D|} \quad (2)$$

In classification problems, coverage measure is defined as the ratio of the number of examples in P that are correctly classified as decision class of D and the number of examples satisfying D. The coverage of the rules is shown for completeness.

Comprehensibility(simplicity): The comprehensibility(comp) of the rule R can be defined by number of attributes on the left

hand side of the rule. If a rule has at most L condition, the comprehensibility of the rule R can be defined as[20]:

$$\text{comp} = \frac{L-h}{L-1} \quad (3)$$

where h is the length of the rule R.

A rule is considered more complex if it contains more attributes. The fitness function is computed as the arithmetic weighted mean of conf, cove and comp as under.

$$\text{Fitness} = w1 \times \text{conf} + w2 \times \text{cove} + w3 \times \text{comp} \quad (4)$$

where w1 , w2 and w3 are user-defined weights. In the proposed algorithm we have chosen the weight values of w1=0.5, w2=0.3 and w3=0.2 according to our own evaluation about relative importance of confidence, coverage and comprehensibility. Not that the value of Fitness is normalized in the range of [0..1].

3. COMPUTIONAL RESULTS

All the datasets used in the experimental evaluation are available from the University of California, Irvine, (UCI) Repository of Machine Learning Databases [21]. The examples that had some missing value were removed from these data sets. Each GA run consisted of a population of 100 individuals evolving during 300 generations. The proposed algorithm was terminated when the best Fitness did not change continually throughout 20 generations. Maximum number of conditions in a decoded rule antecedent = 5. In our experiments, the predefined division of the Adult dataset into a training and a test set is used. The other datasets were, randomly partitioned into a training set and a test set used for testing the quality of the discovered rules. For each class, the algorithm was run five times. Those rules having higher Fitness are collected. The performance of the proposed algorithm on different datasets is demonstrated below:

3.1 Experiments

3.1.1 Experiment 1

The Adult dataset (USA census) contains 48,842 data examples(records) and 15 attributes, nine of them continuous and six discrete. The last attribute is the class attribute which can take two values('>50k' or '<=50k'). The proposed algorithm discovered the following two rules(the best rule for each class) as shown in Table 1.

Table 1. Result for the Adult dataset

No	Mined rules	Conf	Cove	Comp	Fitness
1	**If** Marital-status= Never-married **Then** class= '<=50k'	0.97	0.89	1.00	0.95
2	**If** sex=male **Then** class = '>50k'	0.82	0.85	1.00	0.87

3.1.2 Experiment 2

Iris plants dataset was used in this experiment. This data set contains 3 classes (Iris-setosa, Iris-versicolour and Iris- virginica) of 50 examples each, 4 numeric predictive attributes. Table 2 shows the rules generated from this dataset.

Table 2. Result for the Iris dataset

No	Mined rules	Conf	Cove	Comp	Fitness
1	**If** Petal-length <= 1.9 **Then** class= Iris-setosa	0.98	1.00	1.00	0.99
2	**If** Petal-width > 1.7 **Then** class= Iris-virginica	0.96	0.90	1.00	0.95
3	**If** Petal-length <= 4.9 ∧ Petal-width <= 1.7 **Then** class = Iris-versicolour	1.00	0.94	0.75	0.93

3.1.3 Experiment 3

Heart disease dataset was used in this experiment. This dataset contains 13 continuous attributes with two classes (1-absences and 2- presence) of heart disease. From this dataset, the proposed algorithm discovered two rules, which are given in Table 3.

Table 3. Result for the Heart disease dataset

No	Mined rules	Conf	Cove	Comp	Fitness
1	**If** Rest-BP <=145 ∧ Vessels <=0 ∧ Thal <=3 **Then** class =1	0.92	0.60	0.50	0.74
2	**If** Sex > 0 and Chest-Pain > 3 ∧ Vessels > 0 **Then** class =2	0.95	0.53	0.50	0.73

3.1.4 Experiment 4

Dermatology dataset was used in this experiment. This dataset contains 366 examples and 34 non-categorical attributes (33 of which are linear valued and one of them is nominal. It has six classes (1- psoriasis, 2- seboreic dermatitis, 3- lichen planus, 4- pityriasis rosea, 5- cronic dermatitis, and 6- pityriasis rubra pilaris). Table 4 shows the mined rules from this dataset.

3.2 Comparative Study

In the context of classification rules, it is important to evaluate the quality of the discovered rules with respect to their predictive accuracy. This evaluation must be measured on a separate test set, containing data examples that not seen during training, i.e. the ratio of the number of examples correctly classified over the total of examples in the test set. The results comparing the performance of the proposed algorithm with two other competent learners namely, C4.5 and DTGA(decision tree and genetic algorithm) are reported in Table 5.

Table 4. Result for the Dermatology dataset

No.	Mined rules	Conf	Cove	Comp	Fitness
1	**If** vacuolisation ∧ damage of basal layer = 2 **Then** class=3	0.98	0.60	1.00	0.87
2	**If** perifollicular parakeratosis = 2 **Then** class =6	0.93	0.65	1.00	0.86
3	**If** fibrosis of the papillary dermis = 3 **Then** class =5	0.96	0.44	1.00	0.81
4	**If** fibrosis of the papillary dermis = 0 ∧ elongation of the rete ridges= 2 ∧ spongiosis = 0 **Then** class=1	0.98	0.56	0.50	0.76
5	**If** koebner phenomenon = 1 ∧vacuolisation and damage of basal layer = 0 ∧ spongiosis = 2 **Then** class =4	0.95	0.35	0.50	0.68
6	**If** koebner phenomenon = 0 ∧ fibrosis of the papillary dermis = 0 ∧ spongiosis = 3 ∧ perifollicular parakeratosis = 0 **Then** class=2	0.91	0.34	0.25	0.61

In order to make this comparison fairer, the predictive accuracy is considered here. We have compared the predictive accuracy of the rules discovered by the proposed system with the accuracy of the rules discovered by C4.5 [22] and DTGA[16].The C4.5 is a suite of algorithms for classification problems in machine learning and data mining[22]. It derived from ID3 rule induction algorithm but it improves ID3, as it handles continuous as well as missing data, and follows pruning strategies to remove the party of the tree that don't contribute to classification accuracy. The DTGA consists of two rule inducing phases. In the first phase, a base classifier, C4.5 is used to produce rules from training data set, whereas GA in the next phase refines them with the aim to provide more accurate and high-performance rules for prediction. First of all, just looking into the performance Table 5, it is clear that the proposed algorithm achieves better average predictive accuracy results than the other two competent learners. The proposed algorithm outperformed C4.5 to due to the following difference in the search for rules performed by the two algorithms. The proposed algorithm tends to discover rules for each class and as such, both large frequency classes and small frequency classes are taken into consideration. Thus, it does not bias the search towards any of these kinds of classes. In contrast, C4.5 discovers rules for all classes in each run. Also C4.5 assigns more importance to classes with a large frequency and therefore has a bias towards rules for prediction of common classes. Additionally, due to the presence of deeper nodes in the tree, C4.5 usually discovers rules which are better on the training set than on the test set. The proposed algorithm performed significantly better than default C4.5 in four data sets. The proposed algorithm performed considerably better (somewhat better) than DTGA because the DTGA discovered rules are used much more often to classify test examples belonging to small disjuncts than the large disjuncts. The predictive accuracy which small disjuncts predict the class of unseen examples is lower than that of their larger brethren [23]. The proposed algorithm has tendency to discover large disjuncts easier than small disjuncts. This is due to the ease in generalization of the rule extracted from large disjunct of the training set than the test set. It is evident that the bigger the number of examples covered by a disjunct, the better is accuracy of estimation associated to that disjunct [24]. Therefore, overall one can say that the two algorithms are roughly competitive in terms of predictive accuracy.

Table 5. Summary of predictive accuracy results

Dataset	C4.5	DTGA	The proposed algorithm
Adult	80.56	84.09	89.95
Iris	95.58	97.38	98.85
Heart disease	70.87	77.21	81.65
Dermatology	95.58	97.38	97.87

Figure 3 depicts the comparative performance of the three approaches.

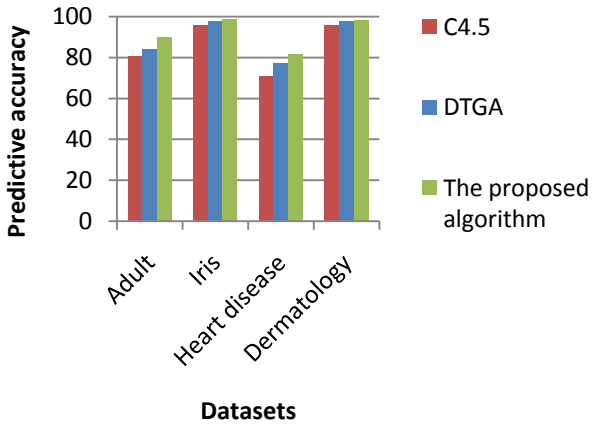

Figure 3. Predictive accuracy of mined rules by C4.5, DTGA and the proposed algorithm.

4. CONCLUSION AND FUTURE WORK

In the present work, the use of GA for discovering predictive, complete and comprehensible rules is explored. The GA scheme has been devised with flexible individual representation, appropriate genetic operators and effective fitness function. The suggested approach is tested on four

datasets from UCI machine learning repository and the rule sets discovered based on three interestingness measures (objective): accuracy, converge and comprehensibility. The proposed GA was compared with respect to the predictive accuracy of the discovered rules with C4.5 and DTGA algorithms. The results proved that none of the selected learners improved the predictive accuracy on any dataset, as much as the proposed algorithm did. An alternative approach worth investigating is to find the novel and unexpected rules using multi-objective genetic algorithms. The goal is to optimize two parameters, namely, the novelty and the unexpectedness metrics for the classification rules simultaneously.

5. REFERENCES

[1] Zhu, X. and Davidson, I. 2007. Knowledge Discovery and Data Mining Challenges and Realities. IGI Global.

[2] Fayyad, U. M., Piatetsky-Sharpio, G. and Smyth, P.1996. From mining to knowledge discovery : An overview. In: Fayyad, U .M., Piatetsky-Sharpio, G. Smyth. P. and Uthurusany, R. (eds.)Advances in knowledge discovery and data mining , AAAI/MIT Press, pp. 1-34.

[3] Han, J., Kamber, M. and Pei, J. 2011. Data Mining: Concepts and Techniques. Third Edition, Morgan Kaufmann.

[4] Freitas, A. A. 2002. Data Mining and Knowledge Discovery with Evolutionary Algorithms. Springer-Verlag, Berlin Heidelberg.

[5] Yogita, Saroj and Kumar, D. 2009. Rule +Exceptions: Automated discovery of comprehensible decision Rules. IEEE International Advance Computing Conference (IACC2009), Patiala, India, pp. 1479-1483.

[6] Barros, R.C., Basgalupp, M.P., Ferreira, A.C. and Frietas, A.A. 2011. Towards the automatic design of decision tree induction algorithms. In: GECCO (Companion Material), Dublin, Ireland, pp. 567-574.

[7] Bramer, M. 2007. Principles of Data Mining. Springer-Verlag London Limited.

[8] Goldberg, D. E. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley.

[9] Dehuri, S. and Mall, R. 2006. Predictive and comprehensible rule discovery using a multi objective genetic algorithms. Knowledge Based Systems, vol. 19, pp. 413-421.

[10] Fidelis, M.V., Lopes, H.S., Freitas, A.A. and Grossa, P. 2000. Discovering comprehensible classification rules with a genetic algorithm. In Proceedings of the 2000 Congress on Evolutionary Computation, La Jolla, CA, USA, IEEE, vol. 1, pp. 805-810.

[11] Kaplia, Saroj, Kumar D. and Kanika. 2010. A genetic algorithm with entropy based initial bias for automated rule mining. In Proceeding of the IEEE International Conference on Computer & Communication Technology(ICCCT'10), pp. 491-495.

[12] Bharadwaj, K. K. and Al-Maqaleh, B.M. 2005. 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, Krakow, Poland, pp.147-152.

[13] Bharadwaj, K. K. and Al-Maqaleh, B.M. 2006. Evolutionary approach for automated discovery of augmented production rules. International Journal of Computational Intelligence. vol. 3, Issue 4, pp. 267-275.

[14] Goplan J., Alhajj R. and Barker, K. 2006. Discovering accurate and interesting classification rules using genetic algorithm. In Proceedings of the International Conference on Data Mining(DMIN06), Las Vegas, Nevada, USA , pp. 389-395.

[15] Carvalho, D. R. and Freitas, A. A. 2002. A genetic-algorithm for discovering small-disjunct rules in data mining. Applied Soft Computing, vol. 2, pp.75-88.

[16] Sarkar, B. K., Sana, S.S. and Chaudhuri, K. 2012. A genetic algorithm-based rule extraction system. Applied Soft Computing. vol. 12, pp. 238-254.

[17] Al-Maqaleh, B.M. 2012. Genetic algorithm approach to automated discovery of comprehensible production rules. In Proceeding of the IEEE 2nd International Conference on Advanced Computing & Communication Technologies (ACCT2012), Rohtak, India, pp.69-71.

[18] Al-Maqaleh, B.M. 2012. Mining interesting classification rules: An evolutionary approach. International Journal of Mathematical Engineering and Science. vol. 1, Issue 1, pp. 13-20.

[19] Frietas, A.A. 1999. On rule interestingness measures. Knowledge-Based System. 12(5-6), pp. 309-315.

[20] Shi, X-J. and Lei, H. 2008. A genetic algorithm-based approach for classification rule discovery. In Proceeding of the IEEE International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII08), Taipei, Taiwan, pp. 175-178.

[21] UCI Repository of Machine Learning Databases, Department of Information and Computer Science University of California, 1994. [http://www.ics.uci.edu/ ~mlearn/MLRepositry.html]

[22] Quinlan. J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

[23] Quinlan, J. R. 1991. Improved estimates for the accuracy of small disjuncts. Journal of Machine Learning, Kluwer Academic Publishers Hingham, MA, USA, vol. 6, Issue 1, pp. 93-98.

[24] Holte, R. C., Acker, L. E. and Porter, B. W. 1989. Concept learning and the problem of small disjuncts. In Proceedings of IJCAI – 89, pp. 813-818.