Knowledge Acquisition Tool for Classification Rules using Genetic Algorithm Approach

Fadl Mutaher Ba-Alwi
Faculty of Computer and IT
Sana'a University – Yemen

ABSTRACT

Classification Rule Mining (CRM) is a data mining technique for discovering important classification rules from large dataset. This work presents an efficient genetic algorithm for discovering significant IF-THEN rules from a given dataset. The proposed algorithm consists of two main steps. First step generates set of classification rules and the second step deletes the weak rules and selects only the significant rules. Since weak rules are deleted and significant rules are selected, the proposed algorithm can be considered as knowledge acquisition tool for classification problems. Experimental results are presented to demonstrate the contribution of the proposed algorithm for discovering the significant rules.

General Terms

Knowledge Discovery in Databases (KDD), Data Mining, Genetic Algorithm, Machine Learning, Pattern Recognition.

Keywords

Classification rules; Genetic algorithm; Significant rule.

1. INTRODUCTION

Data mining is a rapidly evolving art and science of discovering and exploiting new, useful, and profitable relationships in data that is awaking great interest in topics such as decision making, performance prediction, and many other applications [1]. Classification Rule Mining (CRM) is a data mining technique for discovering important classification rules from large dataset that is coupled with a set of pre-defined classes [2]. In the classification task, the discovered knowledge can be represented in different forms. The intuitively clear form for most users is the IF-THEN prediction rule. The IF-part (called the rule antecedent) contains a conjunction of m conditions on values of predictor attributes. The THEN-part (called the rule consequent) contains a prediction about the value of a class attribute. Several approaches for classification rules mining were in the machine learning literature [3-10]. Genetic Algorithm (GA) is a search technique that has been heavily used in different areas where the size of the search space is large. GA is based on the mechanics of natural selection and inspired on the principle of survival of the fittest, where the fittest individuals are selected to produce offspring for the next generation [11]. Selection, crossover and mutation are the basic genetic operators for generating offspring from the fittest individuals. Several GA approaches have been designed for discovering classification rules [11-19]. The contribution of this paper is the discovery of significant (novel) rules from large dataset using genetic algorithm approach. The proposed algorithm consists of two main steps. First step generates set of classification rules and the second step removes the weak rules and selects only the significant rules. Since weak rules are deleted and significant rules are selected, the proposed algorithm can be considered as knowledge acquisition tool for

classification problems. This work is organized as follows. Section 2 describes the related works about significant rules. Section 3 is the detailed description of proposed method. Section 4 describes how the significant rules can be selected from a set of classification rules. Section 5 describes the computational results for the used dataset in the experiment and comparative study with existing techniques. Finally, section 6 concludes the paper.

2. RELATED WORK

The major drawback of CRM is the large number of rules that may be generated [2]. Researchers use different measurements to select only important rules from all possible rules. In [4] the authors extract unexpectedness classification rules. In [5] the authors considered statistical quantitative rules (SQ rules) as a new category of rules. They proposed a permutation-based algorithm for discovering significant SQ rules. The problem addressed in [6] is how to efficiently select a limited number (k) of significant Association Rules (ARs) from the full set of classification ARs. The proposed algorithm in [7] applies a statistical significance test before accepting a pattern. A Numerous attempts have been made to apply GAs in data mining for knowledge discovery and classification. The CRM model with GA in [11] considers the characteristics of cloud computing. In [14] a classification algorithm based on GA approach presented to discover production rules in Conjunctive Normal Form (CNF) where a conjunctive relationship exists between two attributes and disjunction is there among the values of the same attribute. In [17, 20] the proposed GA for classification IF-THEN rules tried to avoid the drawbacks of creating randomly an initial population by creating initial population in a systematic way using the generalized Uniform Population (UP) method.

3. THE PROPOSED GA APPROACH

The accuracy of the discovered classification rules by GA are more accurate than the rules obtained by the other classification algorithms [13]. GA works similar to what happens in nature as species evolve by natural selection. An initial population of individuals is generated randomly or in a systematic way. The individuals in the current population are encoded and evaluated according to fitness function. Individuals are selected according to their fitness to form a new population.

3.1 Individual Encoding

GA has two styles for rule encoding namely Michigan and Pittsburgh. In the Michigan style each individual encodes a single rule, whereas in the Pittsburgh style each individual encodes a set of rules. The Michigan style is better to be used where the goal is to find a small set of accurate classification rules [3]. Therefore, the Michigan style is adopted for encoding in this paper.

Each individual corresponds to the entire IF-part of the rule and composed of n genes, where n is the number of predicted attributes in the given dataset. Each gene corresponds to one condition in this IF-part. The genes are positional, i.e. the first gene represents the first attribute, the second gene represents the second attribute and so on [3]. Each i-th gene contains value (Vi) for the i-th attribute as shown in Figure 1.

Att$_1$	…..	Att$_n$
Gene$_1$	…..	Gene$_n$
V$_1$	…..	V$_n$

Fig 1: Individual structure

If the i-th attribute does not occur in the IF-part of the rule, the corresponding value in gene is blank (i.e. Vi = space character). Otherwise, Vi will be assigned one of the values belonging to the domain of the i-th attribute. Further, the attribute values are coded as follows:'1' for first value, '2' for second value, and so on. The individuals do not involve the class predicted by a rule. Therefore, the GA is run at least once for each class and in each run, all individuals are searching for rules having the same class. Although each individual has a fixed length, the genes are interpreted in such a way that the rule has a variable length [3].

Example 1. Consider the "Weather" dataset in Table 1. The number of predicted attributes is 4. Therefore, 4 genes are used to encode the IF- part of a particular rule.

Table 1. Description of the weather dataset

Attribute	Domain of att.	Code
Outlook	sunny, rain, overcast	1, 2, 3
Temperature	hot, cool, mild	1, 2, 3
Humidity	high, normal	1, 2
Windy	weak, strong	1, 2
Play (class)	yes, no	1, 2

The antecedent part: "IF Outlook=overcast AND Windy=strong" is corresponding to the individual in Figure 2.

Outlook	Temperature	Humidity	Windy
Gene1	Gene2	Gene3	Gene4
3			2

Fig 2: An individual corresponds to a single rule.

3.2 Initialize Population

As mention earlier, the systematic way of creating initial population by the Uniform Population (UP) method [20] avoids the drawbacks of creating initial population in a random way. Therefore, the UP method is adopted for creating initial population in this paper.

3.3 Fitness Function

In GA fitness is used to evaluate the individuals in the population. Individuals with higher fitness value will have higher probability of being selected as candidates for next generation. The proposed fitness function is based on some definitions (explained below) for evaluating single rule Ri of the form: IF A THEN Ci (A is antecedent part and the i-th

class Ci is the consequent part). Rule Ri classify an instance in a given dataset **D** with **n** predicted attributes:

Definition 1. The Precision of the rule Ri can be defined as:

$$\text{Precision (Ri)} = \frac{|A \ \& \ C_i|}{|A|}$$

where |A| is the number of instances satisfying the conjunction conditions in the antecedent part A of rule Ri and |A & Ci| is the number of instances that satisfy both the antecedent A and the consequent Ci. In classification problems, precision used to measure predictive accuracy. This measure has alternative names such as: predictive accuracy, strength, consistency, and confidence.

Definition 2. The Coverage of the rule Ri can be defined as:

$$\text{Coverage (Ri)} = \frac{|A \ \& \ C_i|}{|D|}$$

where |D| is cardinality of the dataset D. In classification problems, coverage measures the percentage of instances covered by both antecedent and consequent parts of a rule. This measure has different names such as: recall, completeness, and support.

Definition 3. The Simplicity of the rule Ri can be defined as:

$$\text{Simplicity (Ri)} = \frac{|\sim A|}{n}$$

where |~A| is the number of attributes not presented in the antecedent part A.

Example 2. The length of the individual in Figure 3 represents the number of predicted attributes. Therefore, n=16, ~A=12 (i.e. 16 - 4=12), and Simplicity of the rule = 0.75.

Gen1				Gen11	…		Gen16	
1		0	…	0	…	0		

Fig 3: An individual for example 2

Definition 4. Let Compl(V_{ij}^{Ci}) denotes the complement of the i-th value of the j-th attribute for class Ci. More precisely, the number of instances in D, excluding V$_{ij}$. This measure can be defined by counting the number of instances with V$_{ij}$ in class Ci, and the number of instances in D, as follows:

$$\text{Compl}(V_{ij}^{Ci}) = 1 - \frac{|V_{ij} \ \& \ Ci|}{|D|}$$

Where $|V_{ij} \ \& \ Ci|$ is the number of instances belongs to class Ci and satisfy the condition value V$_{ij}$.

Example 3. Consider the zoo dataset in [21] with 101 instances. The number of instances with value 'hair=1' in the class 'type=1' is 39. Therefore, the Compl('hair=1') for class 'type =1' is 1- 39/101 = 0.614.

Definition 5. Let cont(att_j^{Ci}) denotes the contribution of the j-th attribute for class Ci, which represents how significantly att$_j$ determines the antecedent part of Ri. The calculation of this measure is given as follows:

$$\text{cont}(\text{att}_j^{Ci}) = \frac{\text{Freq}\left(V_{ij}^{Ci}\right) * \text{FreqClass}\left(V_{ij}^{D}\right) * \text{Compl}\left(V_{ij}^{Ci}\right) * |V_{ij}|}{|D|}$$

where,

$$\text{Freq}\,(V_{ij}^{Ci}) = \frac{\text{number of instances in D with } V_{ij} \text{ and class Ci}}{\text{number of instances in D with class Ci}},$$

$$\text{FreqClass}\,(V_{ij}^{D}) = \frac{\text{number of classes in D}}{\text{number of classes in D which contain } V_{ij}}.$$

Definition 6. Let $\text{cont}(R_i^{Ci})$ denotes the contribution of the rule Ri for class Ci, which represents how significantly Ri determines Ci. This metric is computed as follows:

$$\text{cont}(R_i^{Ci}) = \frac{\sum_{j=1}^{|A|} \text{cont}\left(\text{att}_j^{Ci}\right)}{|A|}$$

where |A| is the number of attributes (conjunction conditions) in the antecedent part of rule Ri.

From the above definitions the proposed fitness function F(Ri) can be formulated as follows:

F(Ri) = w1 * Precision(Ri) *

$$\frac{w2*\text{cont}(R_i^{Ci}) + w3*\text{Coverage(Ri)} + w4*\text{Simplicity(Ri)}}{w1+w2+w3+w4}$$

where w1, w2, w3, and w4 are user-defined weights $\in$ [0,1].

3.4 Genetic Operators
The genetic operators are the core of the GA. They maintain the genetic diversity by introducing new genetic material. Selection, crossover, and mutation are the conventional genetic operators.

3.4.1 Selections
The selection operator is used to improve the chances of the survival of the fittest individuals. The fitter the individual, the more times it is likely to be selected to reproduce. A number of selection methods: Proportionate Selection (e.g. roulette wheel), Ranking Selection (e.g. truncate selection), Tournament Selection (e.g. binary tournament selection), etc. are described and used in literature so far. In the present work the roulette selection method is used. First, the total of fitness value of all the individuals in the population is calculated (say 'S') and then a random number 'r' is generated such that r $\in$ [0...S]. Finally, the individual whose fitness, added to the previous one, is greater or equal to 'r', is selected.

3.4.2 Crossover
Crossover operator exchanges subsequences of two individuals (parents) to create two new individuals (offspring). Different types of crossover operations are single-point crossover, two-point crossover, uniform crossover etc. In this paper, single-point crossover is used with crossover probability (Pc) = 0.85.

Example 4. The dotted line in Figure 4 determines the randomly chosen single-point crossover for the parents (P₁ and P₂). The genes of parents to the right of the crossover point are swapped between the two individuals, yielding the new offspring O₁ and O₂.
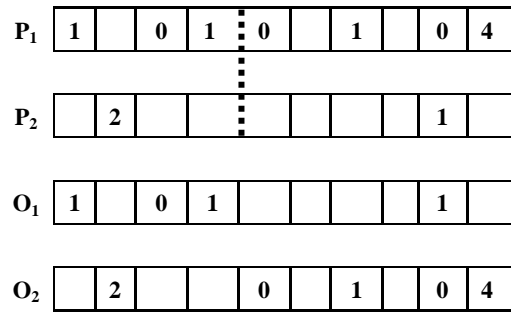

Fig 4: Single-point crossover

3.4.3 Mutation
Mutation operation is used to guarantee that the search is done in the whole population by maintaining the diversity of gene. Higher mutation probability can turn the genetic algorithm in a random search algorithm [14]. In the present work, the mutation probability (Pm) = 0.01.

4. SELECTING SIGNIFICANT RULES
The discovered set of classification rules may contain weak rules (i.e. noisy rule) which are unnecessary rules for the user. Removing such these rules needs to recognize them with respect to the user subjective. The weak rules can be defined as follows:

Definition 7. Weak rule is a rule that covers more instances from the other classes than from its own. Let R_W^{Ci} is the set of weak rules for class Ci such that every rule $\in R_W^{Ci} < \delta_{Ci}$, where δ_{Ci} is small user-defined constant $\in$ [0…1] for class Ci.

Selection of significant classification rules is based on the definitions (6) and (7). Rule Ri is significant for class Ci if the contribution of Ri is greater than the sum of the contribution of all weak rules in R_W^{Ci}. Formally written as:

$$\text{cont}(R_i^{Ci}) > \sum_{j=1}^{|R_W^{Ci}|} \text{cont}\left(R_j^{Ci} \in R_W^{Ci}\right)$$

where $|R_W^{Ci}|$ is the cardinality of the set R_W^{Ci} (i.e. number of weak rules for class Ci). Since weak rules are deleted and significant rules are selected, the proposed algorithm can be considered as knowledge acquisition tool for classification problems.

5. DATASET USED IN EXPIRMENT
The performance of the proposed algorithm is investigated on the public domain dataset, the so-called zoo dataset which obtained from the UCI-Machine Learning Repository [21]. This dataset have been used heavily for classification tasks. It contains 101 instances and 18 attributes (17 predicting attributes and one target attribute, which can take 7 classes). All attributes are categorical, and no missing data. In the pre-processing phase the irrelevant attribute "animal-name" was removed because each instance contains unique value for this attribute, so it has no generalization power.

5.1 Computational Result
In the initial population at the first run for class 'type=1', more than 65534 individuals are generated using UP method. Figure 5. shows some of these individuals. The discovered classification rules of three classes 'type=1', 'type=3' and

'type=5', for instance, at the first step of the proposed algorithm are shown in Table 2. The symbol '&' in the 'IF-part' column means logical 'AND'. The user-defined weights (w1, w2, w3, w4) are assigned value 1 and the user-defined constant δ_{Ci}=0.5 (i=1,3,5). Rule R_1 is selected as significant rule for class 'type=1' because:

$$\text{cont}(R_1^{type=1}) > \sum_{j=1}^{3} \text{cont}\left(R_j^{type=1} \in R_W^{type=1}\right)$$

where $R_W^{type=1}$ = {R_2 , R_3 , R_4}. i.e. 1.688 > (0.858 + 0.302 + 0.425 = 1.585).

Further, rule R_5 is selected as significant rule for class 'type=3' because:

$$\text{cont}(R_5^{type=3}) > \sum_{j=1}^{2} \text{cont}\left(R_j^{type=3} \in R_W^{type=3}\right)$$

where $R_W^{type=3}$ = {R_6 , R_7}. i.e. 0.53 > (0.054 + 0.436 = 0.49).

Furthermore, rule R_8 is selected as significant rule for class 'type=5' because:

$$\text{cont}(R_8^{type=5}) > \sum_{j=1}^{2} \text{cont}\left(R_j^{type=5} \in R_W^{type=5}\right)$$

where $R_W^{type=5}$={R_9 , R_{10}}. i.e. 0.571 > (0.312 + 0.219 = 0.531).

Finally, all the selected significant classification rules for each class at the second step of the proposed algorithm are shown in Table 3.
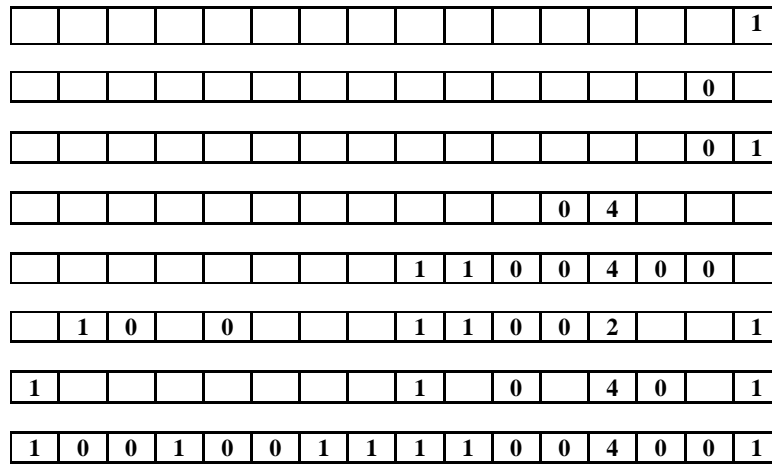

Fig 5: Some of individuals at the first run using UP method

Table 2. Discovered classification rule set for Zoo dataset at the first step.

Rule#	Antecedent part of the discovered rules	Class(THEN part)	Precession (Ri)	Cont (Ri)	Simplicity (Ri)	Coverage (Ri)	F(Ri)
R_1	milk = 1	1	1	1.688	0.938	0.406	0.758
R_2	milk=1 & aquatic=0 & backbone=1 & fins=0	1	1	0.858	0.75	0.347	0.489
R_3	legs=4 & Catsize = 1	1	0.962	0.302	0.875	0.248	0.343
R_4	toothed=1 & legs=4 & domestic=0 & catsize=1	1	1	0.425	0.75	0.188	0.341
R_5	eggs= 1 & predator=1 & toothed=1& catsize=0	3	0.273	0.53	0.75	0.03	0.089
R_6	venomous = 1	3	0.25	0.054	0.938	0.02	0.063
R_7	eggs= 1 & predator=1 & catsize=0	3	0.125	0.436	0.813	0.03	0.04
R_8	milk=0 & aquatic=1 & breathes=1 & legs=4 & catsize=0	5	1	0.571	0.688	0.04	0.325
R_9	hair=0 & airborne=0 & aquatic=1 & toothed=1& breathes=1& legs=4 & catsize=0	5	1	0.312	0.563	0.04	0.229
R_{10}	airborne=0 & aquatic=1 & toothed=1 & breathes=1 & catsize=0	5	0.444	0.219	0.688	0.04	0.105

Table 3. Discovered significant classification rule set for Zoo dataset at the second step.

Rule#	Significant Rules	Precession (Ri)	Cont (Ri)	Simplicity (Ri)	Coverage (Ri)	F(Ri)
R_1	IF milk=1 THEN type=1	1	1.688	0.938	0.406	0.758
R_2	IF feathers=1 THEN type=2	1	1.112	0.938	0.198	0.562
R_3	IF eggs= 1 & predator=1 & toothed=1& catsize=0 THEN type=3	0.273	0.53	0.75	0.03	0.089
R_4	IF milk=0 & fins=1 THEN type=4	1	0.558	0.875	0.129	0.391
R_5	IF milk=0 & aquatic=1 & breathes=1 & legs=4 & catsize=0 THEN type=5	1	0.176	0.688	0.04	0.226
R_6	IF aquatic=0 & legs=6 THEN type=6	1	0.588	0.875	0.079	0.386
R_7	IF Airborne=0 & backbone=0 THEN type=7	0.833	0.624	0.875	0.099	0.333

5.2 Comparative Study

The discovered rules by the proposed algorithm are compared with the results obtained by genetic algorithms used in [18, 22]. Algorithm is run separately for each class and the corresponding rule set is obtained for this predicted class. The discovered rules by PSO algorithm in [22] (see Table 4.) are better than the discovered by [18] in terms of accuracy. The proposed algorithm has found better classification rules (in terms of rule contribution) than PSO algorithm.

Table 4. The discovered rules by PSO algorithm

Discovered rules	Class
IF milk=1	1
IF feathers=1 & toothed=0	2
IF eggs=1 & predator=1 & catsize=0	3
IF aquatic=1 & breathes=0 & tail=1	4
IF airborne=0 &aquatic=1 &toothed=1&catsize=0	5
IF airborne=1 & fins=0 & tail=0	6
IF predator=1 & breathes=0 & domestic=0	7

Table 5. shows the rule contribution for proposed and PSO algorithms.

Table 5. Rule contribution

Rule#	PSO	Proposed
R_1	1.688	1.688
R_2	0.779	1.112
R_3	0.436	0.53
R_4	0.513	0.558
R_5	0.273	0.176
R_6	0.567	0.588
R_7	0.564	0.624

Figure 6. shows the comparisons in terms of rule contribution between the results obtained by the proposed algorithm and PSO algorithm.
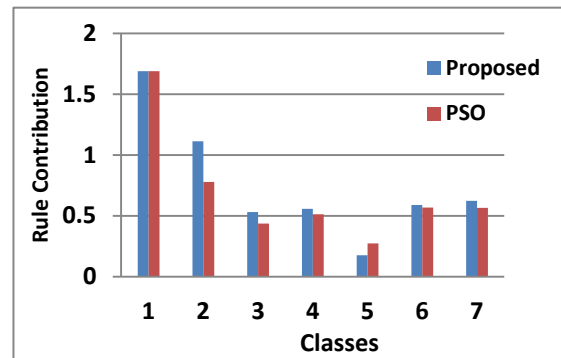

Fig 6: Comparison between proposed and PSO algorithms

6. CONCLUSION AND FUTURE WORK

Classification is a data mining technique which explores data from the database. The genetic algorithms are adaptive techniques that can be used in solving complex search problems. A database can be viewed as a very large search space and a genetic algorithm can be viewed as a search technique in a database. In this paper, the problem of discovering significant classification rules has been investigated. The proposed algorithm consists of two phases. In the first phase, set of classification rules are generated using genetic algorithm approach. In the second phase, among the discovered classification rules, the novel rules are selected and the weak rules are removed. Since weak rules are removed and novel rules are selected, the proposed algorithm can be considered as a knowledge acquisition tool for classification problems. The zoo dataset has been used for experiment and the proposed approach is giving satisfactory results. Fuzzy genetic algorithm may be introduced for

selecting fuzzy significant classification rules as an ulterior development direction.

7. REFERENCES

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

[2] Wang, Y.J., Xin, Q., and Coenen, F., 2008 Mining Efficiently Significant Classification Association rules. Data Mining: Foundations and Practice, 443-467.

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

[4] Liu, B., Hsu, W., and Chen, S., 1997 Using General Impressions to Analyze Discovered Classification Rules. In Proc. KDD97 3rd International Conference on Knowledge Discovery and Data Mining, AAAI Press, 31–36.

[5] Zhang, H., Padmanabhan, B., and Tuzhilin, A., 2004 On the discovery of significant statistical quantitative rules. In Proc. of the 10th international conference on knowledge discovery and data mining, KDD-2004, 374-383.

[6] Wang, Y.J., Xin, Q., and Coenen, F., 2005 Selection of Significant Rules in Classification Association Rule Mining. In Proc. of ICDM2005 WORKSHOP on Foundations of Semantic Oriented Data and Web Mining (ICDM-FDM'2005), 106-108.

[7] Webb, G.I., 2006 Discovering significant rules. In Proc. of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD-2006, 434-443.

[8] Tzacheva, A., and Raś, Z. W., 2007 Constraint Based Action Rule Discovery with Single Classification Rules. Intelligent Data Engineering and Automated Learning, IDEAL-2007. Lecture Notes in Computer Science, vol. 4482/2007, 322-329.

[9] Qin, B., Xia, Y., Prabhakar, S., and Tu, Y., 2009 A Rule-Based Classification Algorithm for Uncertain Data. Data Engineering, ICDE '09. IEEE, 25th International Conference.

[10] Hills, J., Davis, L.M., and Bagnall, A., 2012 Interestingness Measures for Fixed Consequent Rules. Intelligent Data Engineering and Automated Learning, IDEAL-2012. Lecture Notes in Computer Science, vol. 7435/2012, 68-75.

[11] Ding, J., and Yang, S., 2012 Classification Rules Mining Model with Genetic Algorithm in Cloud Computing. International Journal of Computer Applications , vol. 48–No.18, 0975–888.

[12] Noda, E., Freitas, A.A., and Lopes, H.S. 1999 Discovering interesting prediction rule with a genetic algorithm. In Proc. of the 1999 Congress on Evolutionary Computation, vol. 2.

[13] Yang, L., Widyantoro, D.H., Ioerger, T., and Yen, J., 2001 An Entropy-based Adaptive Genetic Algorithm for Learning Classification Rules. In Proc. of the 2001 Congress on Evolutionary Computation, 790–796.

[14] Vashishtha, J., Kumar, D., Ratnoo, S., and Kundu, K., 2011 Mining Comprehensible and Interesting Rules: A Genetic Algorithm Approach. International Journal of Computer Applications , vol. 31, No.1, 0975 – 8887.

[15] G¨undogan, K.K., Alata_s, B., Karc, A., Tatar, Y., 2002 Comprehensible Classification Rule Mining With Two-Level Genetic Algorithm. 2nd FAE International Symposium, TRNC, 373-377.

[16] Fidelis, M.V., Lopes, H.S., and Freitas, A.A., 2000 Discovering Comprehensible Classification Rules with a Genetic Algorithm. In Proc. of Congress on Evolutionary Computation.

[17] Koray, K., ALATAS, B., and KARCI, A., 2004 Mining Classification Rules by Using Genetic Algorithms with Non-random Initial Population and Uniform Operator. Turk J Elec Engin, vol. 12, No.1.

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

[19] Dehuri S., Patnaik S., Ghosh A., and Mall R., 2008 Application of elitist multi-objective genetic algorithm for classification rule generation. Applied Soft Computing, vol. 8, 477–487.

[20] Karc, A., and Arslan, A., 2002 Uniform population in genetic algorithms. I.¨U. Journal of Electrical & Electronics, vol.2, 495-504.

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

[22] Bilal A., and Erhan A., 2009 Multi-objective rule mining using a chaotic particle swarm optimization algorithm. Int. Journal of Knowledge-Based Systems, vol.22, 455–460