International Journal of Innovative
Computing, Information and Control
Volume **8**, Number **9**, September **2012**

ICIC International ©2012 ISSN 1349-4198
pp. 6303–6315

A NEW APPROACH TO CLASSIFICATION RULE EXTRACTION PROBLEM BY THE REAL VALUE CODING

Murat Koklu*, Humar Kahramanli and Novruz Allahverdi

Department of Electronic and Computer Education
Selcuk University
Konya 42003, Turkey
*Corresponding author: mkoklu@selcuk.edu.tr
{ hkahramanli; noval }@selcuk.edu.tr

Received June 2011; revised October 2011

Abstract. *In this study a new method that uses artificial immune system (AIS) algorithm has been presented to extract rules from medical related dataset. Four real life problems data were investigated for determining feasibility of the proposed method. The data were obtained from machine learning repository of University of California at Irvine (UCI). The datasets were obtained from Iris Dataset which is the multi-class problem, Pima Indian Diabetes Dataset and two different Wisconsin Breast Cancer datasets. The proposed method achieved prediciton accuracy ratios of 100%, 77.2%, 98.54% and 95.61% for the Iris, Pima Indians Diabetes, Wisconsin Breast Cancer (original) and Wisconsin Breast Cancer (diagnostic) datasets, respectively. It has been observed that these results are better than the results obtained from related previous studies.*
Keywords: Rules extraction, Artificial immune systems, CLONALG algorithm

1. **Introduction.** Information technology development over the last two decades has moved rapidly from centralized single purpose systems to the distributed, multipurpose systems [1]. As a result of this development the wide range of data has occurred. In recent years there are growing interest to analyze examples in such repositories of data. Since there is a huge collection of data to be explored, data mining became an attractive field for scientists. It aims to discover a useful knowledge from a large amount of data. Thus, data mining becomes one of the most important tools used for solving most of today's problems that are related to different sectors of our life. The wide range of applications from business tasks (credit risk analysis) to scientific tasks (prediction of mutagenicity of organic compounds) have led to the development of a huge variety of learning methods and algorithms for rule extraction and prediction. The general tasks of classification, regression, clustering, or deviation analysis have a large number of solutions such as neural networks, decision tree learners, rule learners or Bayesian networks [1]. Among them rule extraction studies have the rising popularity because of providing explanation capability, data mining and knowledge discovery and knowledge acquisition for the expert system. Rule extraction has some advantages [2]:

- Providing a mechanism that can interpret the network input/output mappings in the form of use.
- Ability to identify deficiencies in the original training.
- Identification of unnecessary network parameters for removal would enhance network performance.
- Analysis of previously unknown relationships in the data.

- Provide reasoning and explanation capabilities.
- Support cross-referencing and verification capabilities.
- Alleviate the knowledge acquisition problem and refine initial domain knowledge.

A rule extraction algorithm should meet several important requirements for practical use. Extracted rules need to be simple and comprehensive; otherwise, a human will not be able to comprehend them. It is also important to discover accurate knowledge that is [3]. In the most of studies researchers use Artificial Neural Network (ANN) and Support Vector Machine (SVM) for classification and then extract rules from this system.

Ozbakir et al. proposed a rule-based classifier algorithm which makes use of ant colony optimization. The proposed rule extraction algorithm works on the trained ANNs in order to discover the hidden knowledge which is available in the form of connection weights within ANN structure [4].

Zárate et al. used Formal Concept Analysis (FCA) in order to extract and represent knowledge from previously trained ANN. The new FCANN approach permits to obtain a complete canonical base, non-redundant and with minimum implications, which qualitatively describes the process being studied. The proposed approach has a sequence of steps such as the generation of a synthetic dataset [5].

Kusunoki et al. introduced a two stage approach to construct the hierarchical rule classifier for multiclass problems. In the first stage an agglomerative hierarchical clustering algorithm is used to obtain a hierarchical structure of groups of decision classes. In the second stage they apply a rule induction algorithm for each branching node in this hierarchy – these rules assign a classified object to groups of classes [6].

Nayak proposed a methodology named as Gyan that represents the knowledge of a trained network in the form of restricted first-order predicate rules. The successful application and competitive results obtained by Gyan for various problem domains demonstrate its effectiveness in real-life problems (such as Queensland Rail and Remote Sensing), in fairly large size problems (in terms of number of attributes, such as Breast Cancer, Moral Reasoner, Voting and Mushroom), and in continuous-valued problem domains (such as Cleveland heart disease) [7].

Elalfi et al. presented an algorithm for extracting accurate and comprehensible rules from databases by trained artificial neural network (ANN) and genetic algorithm (GA) [8].

Rodríguez et al. presented a distributed enhanced genetic algorithm for classification rules extraction that is based on the island model and also scalable for data training partitioning. To be able to generate an accurate classifier with data partition, two techniques were proposed: an elitist pool for rule selection and a novel technique of data distribution (DLF) that uses heuristics based on the local data to dynamically redistribute the training data in the node neighborhood [9].

Martens et al., introduced two methods for extracting rules from SVM which is taken from the artificial neural networks domain, being Trepan and G-REX. The described techniques are compared using publicly available datasets, such as Ripley's synthetic dataset and the multi-class iris dataset [10].

Farquad et al. presented a hybrid approach for extracting rules from SVR. The proposed hybrid rule extraction procedure has two phases: (1) to obtain the reduced training set in the form of support vectors using SVR (2) to train the machine leaning techniques (with explanation capability) using the reduced training set. The proposed hybrid rule extraction procedure is compared with stand-alone CART, ANFIS and DENFIS. Extensive experiments are conducted on five benchmark data sets viz namely Auto MPG, Body

Fat, Boston Housing, Forest Fires and Pollution, to demonstrate the effectiveness of the proposed approach in generating accurate regression rules [11].

Chen et al. proposed classification system which develops rules by analyzing examples. This study therefore proposes a new procedure using feature selection and entropy-based rough sets [12].

Ang et al. proposed an Evolutionary Memetic Algorithm (EMA), which uses a local search intensity scheme to complement the global search capability of Evolutionary Algorithms (EAs), for rule extraction. They studied two schemes for local search, namely EMA-μGA, which uses a micro-Genetic Algorithm-based (μGA) technique, and EMA-AIS, which is inspired by Artificial Immune System (AIS) and uses the clonal selection for cell proliferation [13].

Avner introduces a system that extracts comprehensible symbolic rules from a multilayer perceptron. Once the network has been trained in the usual manner, the training set is presented again, and the actual activations of the units recorded. Logical rules, corresponding to the logical combinations of the incoming signals, are extracted at each activated unit [14].

Omkar used Ant Colony Optimization (ACO) to obtain rules that can classify the data into pre-defined classes. It can be used to classify acoustic emission (AE) signals with respect to the sources [15].

Odajima et al. proposed a GRG (Greedy Rule Generation) algorithm, a method for generating classification rules from a data set with discrete attributes. The algorithm is "greedy" in the sense that at every iteration, it searches for the best rule to generate [16].

Wong et al. proposed a method based on genetic algorithms to automatically extract fuzzy rules to identify a system where only its input-output data are available. This method can determine a fuzzy system with fewer fuzzy rules as well as the antecedent and consequent parameters of the fuzzy rules at the same time. A nonlinear system is utilized to illustrate the efficiency of the proposed method in the rule extraction for fuzzy modeling [17].

As it is seen from literature most of researchers have preferred to extract rules from ANN or SVM. First difference of this study from the previous ones is to extract rules directly from the data instead of ANN or SVM. Moreover, researchers have used binary data coding method. To implement binary coding researcher splits data into the custom intervals. The second difference is the way of presenting the data since partitioning of the data has been implemented by optimization. Let a parameter of a data take value from the interval $(0, 5)$. Assume that the researcher wants to split this interval into 3 subintervals. Widely used method in the literature is to divide this interval into three equal subintervarls such as $(0, 1.66)$, $[1.66, 3.33]$ and $[3.33, 5]$. However, there is only little possibility that this split will have subintervals as $(0, 1.222)$, $[1.222, 2.789)$ and $[2.789, 5)$. These subintervals may belong to the correct splitting and the optimization can do it as in this study. Using optimization the interval will be split into subintervals which cannot be foreseen by researchers and this partitioning may generate more accurately rules.

For optimization purpose CLONALG which is one of the Artificial Immune System (AIS) algorithms has been used in this study. As the CLONALG is an optimization algorithm it is necessary to use the fitness function. Accuracy function has been used as fitness function for the optimization algorithm.

The rest of paper is organized as follows. In Section 2, the material and methods have been introduced. In Section 3, the proposed rule extraction method has been described. In Section 4, experimental results have been presented. In Section 5, the paper is concluded.

2. Material and Methods.

2.1. **Artificial immune systems.** Artificial Immune System (AIS) is a computational technique inspired from natural immune system. AIS used to solve different problems such as clustering/classification, anomaly detection, computer security, numeric function optimization, combinatory optimization, learning, bio-informatics, image-processing, robotics, control, virus detection and web mining [18]. The AIS has focused on three main immunological theories: clonal selection, immune networks and negative selection [19].

CLONALG is the abbreviation of the clonal algorithm and has been inspired by the following elements of the clonal selection theory [20]:

- Maintenance of a specific memory set
- Selection and cloning of most stimulated antibodies
- Death of non-stimulated antibodies
- Affinity maturation (mutation)
- Re-selection of clones proportional to affinity with antigen
- Generation and maintenance of diversity

The CLONALG algorithm can be described as follows [20].

2.1.1. *Initialization.* The first step of the CLONALG technique is initialization, which involves preparing an antibody pool of fixed size N. This pool is then partitioned into two components, a memory antibody section m that eventually becomes representative of die algorithms solution and a remaining antibody pool r used for introducing additional diversity into the system.

2.1.2. *Loop.* The algorithm then proceeds by executing a number of iterations of exposing the system to all known antigens. A single round of exposure or iteration is referred to as a generation. The number of generations G the system executes is user configurable, though the system can use a problem specific stopping condition.

Select Antigen: A single antigen is selected at random without replacement (for the current generation) from the pool of antigens.

Exposure: The system is exposed to the selected antigen. Affinity values are calculated for all antibodies against the antigen. Affinity is a measure of similarity, and is problem dependent.

Selection: A set of n antibodies are selected from the entire antibody pool that have the highest affinity with the antigen.

Cloning: The set of selected antibodies are then cloned in proportion to their affinity (rank based).

Affinity Maturation (mutation): The clone (set of duplicate antigen) is then subjected to an affinity maturation process to better match the antigen m that is subject to question. Here the degree of maturation is inversely proportional to their parents affinity (rant based): meaning that the greater the affinity, the lower the mutation.

Clone Exposure: The clone is exposed to the antigen, and affinity measures are calculated.

Candidature: The antibody or antibodies with the highest affinity from the clones are then selected as candidate memory antibodies for m. If the affinity of a candidate memory cell is higher than that of the highest stimulated antigen from the memory pool m then it replaces with that antigen. Group replacements can occur in a similar, but batched manner for m.

Replacement: Finally, the d individuals in the remaining r antigen pool with the lowest affinity are replaced with new random antibodies.

2.1.3. *Finish.* After die completion of the training regime, the memory m which is one of the component of the antigen pool is then taken as the algorithms solution. Depending on the problem domain, the solution may be a single best individual antigen or the collective of all antigens in the pool [20].

2.2. **The fitness function.** In this study CLONALG has been used for classification rule mining aim. Also the fitness function used in this study is defined as follow [21]:

$$F = \frac{TP * TN}{(TP + FN) * (FP + TN)} \tag{1}$$

where N is the total number of samples. TP (true positives) is the number of samples detected by the rule that have the class predicted by the rule. FP (false positives) is the number of samples detected by the rule that have a class different from the class predicted by the rule, FN (false negatives) is the numbers of samples that are not detected by the rule but that have the class predicted by the rule. TN (true negatives) is the number of samples that are not detected by the rule and that do not have the class predicted by the rule.

3. **Rule Extraction Using CLONALG.** In this paper, we propose a novel approach for extracting rules from data. The idea behind suggested approach is to use artificial immune systems for optimization of the classification accuracy function. The Rule Mining algorithm is as follows:

Step 1. Code the data

Step 2. Create the population randomly

Step 3. Execute CLONALG

Step 4. Decode the rules

In this study as the first step of this algorithm the dataset was normalized in the interval $[0, 1]$. The normalization has been made using following Equation (2):

$$x_n = \frac{x_r - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

where, x_r is the original value of data attribute, $x_{\min}$ is the minimum value of attribute of whole data, $x_{\max}$ is the maximum value of attribute of whole data and x_n is the normalized value of data attribute.

In the second step of the algorithm the population was created. The population dimension was calculated with respect to attribute number. Each individual in population represents the intervals of attributes. Due to the fact that each interval symbolizes with two numbers, the individual's dimension is equal to 2∗attribute number. Suppose that the dataset has N attributes. In our presentation the first N numbers represent the midpoints of interval and other N means the amount of expansions of these intervals. After optimization the $2 * N$ length rules have been produced. Let $X = \{x_1, x_2, \ldots, x_N, x_{N+1}, \ldots, x_{2N}\}$ is the rule vector and $F = \{f_1, f_2, \ldots, f_N\}$ is the futures. Here $\{x_1, x_2, \ldots, x_N\}$ means the midpoints of the intervals, where $\{x_{N+1}, \ldots, x_{2N}\}$ means the expansions due to the fact that the the future f_i takes the value from the interval $(x_i - x_{N+i}, x_i + x_{N+i})$. Hence, the rule is shown as Equation (3).

$$R = \bigcup_{i=1}^{N} f_i \in (x_i - x_{i+N}, x_i + x_{i+N}) \tag{3}$$

For example, suppose that the dataset has four attributes that means $N = 4$. Hence, the individual's dimension must be $2 * N = 8$. Let the individual be $\{2.5, 3, 4, 1, 1.5, 4, 2, 2\}$. Thus, the first four values $\{2.5, 3, 4, 1\}$ mean the midpoints and others $\{1.5, 4, 2, 2\}$ mean

the expansions. Since the midpoint of interval for the first feature is 2 and amount of expansion is 1.5 this attribute takes the value from the interval $(1, 4)$. Due to the same reason the second attribute takes the values from the interval $(-1, 7)$, the third attribute takes the values from the interval $(2, 6)$ and the fourth attribute takes the values from the interval $(-1, 3)$.

In third stage of algorithm CLONALG was executed for all classes separately. As result of this third stage, the vectors which have $2 * N$ length were produced. Each vector means a rule. To see the rules these vectors must be decoded as the last stage of study. For extracting a rule that belongs to each class the best antibody must be decoded as follows:

- The best antibody is denormalized.
- The obtained vector is divided into 2 segments. First segment represents the midpoints of intervals while second represents the expansions.
- The attribute intervals are formed using this midpoints and expansions.
- The operator "AND" is used to correlate the composed intervals of the same data.

For example, consider the Iris dataset from UCI machine learning repository [22], which is widely used botanical problem related classification and rule extraction studies. The dataset has four attributes namely Sepal length (SL), Sepal width (SW), Petal length (PL) and Petal width (PW). The SL takes the value from interval $(4.3, 7.9)$, SW takes value from interval $(2, 4.4)$, PL takes value from interval $(1, 6.9)$ and PW takes value from interval $(0.1, 2.5)$. Dataset has three classes as Iris Setosa, Iris Versicolour and Iris Virginica.

As mentioned previously researchers have splitted these intervals by custom decsion. For example, in literature [4] SL has been splitted in to 3 subintervals as $[4.3, 5.55]$, $(5.55, 6.15]$, $(6.15, 7.9]$, SW as $[2, 2.95]$, $(2.95, 3.35]$, $(3.35, 4.4]$, PL as $[1, 2.45]$, $(2.45, 4.75]$, $(4.75, 6.9]$ and PW $[0.1, 0.8]$, $(0.8, 1.75]$, $(1.75, 2.5]$. Different from the previous studies this study makes use of CLONALG optimization algorithm for splitting. As a result the intervals for the Setosa class have been extracted as SL$\in (3.5858, 9.4853)$, SW$\in (2.2960, 4.9254)$, PL$\in (0.2074, 2.9959)$ and PW$\in (-0.3674, 1.4770)$. It can be observed that it is hard to form such intervals by custom decision.

Using these intervals this rule is written as follows:

IF SL$\in (3.5858, 9.4853)$ & SW$\in (2.2960, 4.9254)$ & PL$\in (0.2074, 2.9959)$ & PW$\in (-0.3674, 1.4770)$ THEN the Class Setosa (with custom splitting).

Due to SL takes value from the interval $(4.3, 7.9)$ and $(4.3, 7.9) \subset (3.5858, 9.4853)$, it is not required to write SL$\in (3.5858, 9.4853)$. So we can eliminate SL from the rule. Due to the fact that $2 < 2.2960$ and $4.4 < 4.9254$ it can be written that SW> 2.2960. Carring out similar logic, it can be written PL< 2.2995 and PW< 1.4770. Consequently the supposed rule is decoded in a brief way as follows:

IF SW> 2.2960 & PL< 2.2995 & PW< 1.4770 THEN Class Setosa (with COLONALG splitting)

It should be noted that the first representation of the rule is not false but difficult to read where the second is more comprehensible. Detailed explanation of IRIS dataset can be seen in Section 4.1. To show the performance of models, majority vote was used for determining the class of samples.

4. Experimental Results.

Proposed method was applied to four different datasets in this study. The used datasets were Wisconsin Breast Cancer (original) dataset, Wisconsin Breast Cancer (diagnostic) dataset, Pima Indian Diabetes and Iris dataset.

4.1. Iris dataset.

The dataset chosen for first experiment was the Iris dataset from UCI machine learning repository [22]. The dataset has four attributes. These are

- Sepal length
- Sepal width
- Petal length
- Petal width

The classes were coded as 1 (Iris Setosa), 2 (Iris Versicolour) and 3 (Iris Virginica). All classes have 50 samples. The attributes have different range values in the database and these ranges can be seen in Table 1.

Table 1. Iris dataset attribute and range values

Attribute	**Range**
Sepal length (SL)	4.3-7.9
Sepal width (SW)	2-4.4
Petal length (PL)	1-6.9
Petal width (PW)	0.1-2.5

Then CLONALG was executed on this data. As a result 7 rules (1 rules for class 1, 3 rules for class 2, and 3 rules for class 2) were created. To see the rules, extracting vectors must be decoded. The extracted set of rules are presented in Table 2.

Table 2. Iris dataset rules

Rule Numbers	**Rules**
Rule 1	IF SW $>$ 2.2960 & PL $<$ 2.2995 & PW $<$ 1.4770 THEN Class 1.
Rule 2	IF SW $>$ 2.2960 & PL $<$ 2.2995 & PW $<$ 1.4770 THEN Class 1.
Rule 3	IF SW $\in$ (2.8232, 3.3693) & PL $\in$ (4.6651, 5.0688) THEN Class 2.
Rule 4	IF SW $\in$ (2.3993, 2.7015) & PL $<$ 5.2246 & PW $\in$ (0.6249, 1.6499) THEN Class 2
Rule 5	IF PL $>$ 4.7428 THEN Class 3.
Rule 6	IF SL $<$ 7.6553 & SW $<$ 2.9679 & PL $>$ 3.3016 & PW $>$ 1.4345 THEN Class 3
Rule 7	IF SL $\in$ (5.9947, 6.7276) & SW $<$ 3.7341 & PW $\in$ (1.7034, 2.6570) THEN Class 3.

In this database 100% of samples were classified correctly. The classification accuracy of the proposed method is shown in Table 3 with the accuracies obtained from previous studies in the literature [23]. As can be seen from Table 3 the proposed method has the best performance for rule extracting.

4.2. **Pima Indian diabetes dataset.** The dataset chosen for second experiment was the Pima Indian Diabetes dataset from the same repository [22]. The dataset has eight attributes. These are

- Number of times pregnant.
- Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
- Diastolic blood pressure (mm Hg).
- Triceps skin fold thickness (mm).
- 2-Hour serum insulin (mu U/ml)
- Body mass index (weight in kg/(height in m)^2)
- Diabetes pedigree function
- Age (years)

Table 3. Iris dataset experimental results

Method	Accuracy	Reference
Our Method	100	This study
Grobian (rough)	100	Browne
PVM 2 rules	98.0	Weiss
C-MLP2LN	98.0	Duch et al.
SSV	98.0	Duch et al.
PVM 1 rule	97.3	Weiss
NEFCLASS	96.7	Nauck et al.
FuNe-I	96.7	Halgamuge
CART (dec. tree)	96.0	Weiss
FuNN	95.7	Kasabov

The classes were coded as 0 and 1. The class 0 (500 samples) and class 1 (268 samples) contain the patients with tested negative and positive for diabetes respectively. The attributes have different range values in the database and these ranges can be seen in Table 4.

Table 4. Pima Indian diabetes attribute and range values

Attribute	Range
Number of times Pregnant (NTP)	0-17
Plasma glucose concentration a 2 hours in an oral glucose tolerance test (PGC)	0-199
Diastolic blood pressure (mm Hg) (DBP)	0-122
Triceps skin fold thickness (mm) (TSFT)	0-99
2-Hour serum insulin (mu U/ml) (HSI)	0-846
Body mass index (weight in kg/(height in m)^2) (BMI)	0-67.1
Diabetes pedigree function (DPF)	0.078-2.42
Age (years) (A)	21-81

Then CLONALG was executed on this data. As a result 26 rules (13 rules for class 0 and 13 rules for class 1) were created. To see the rules, extracting vectors must be decoded. The examples of extracted set of rules were presented in Table 5.

Table 5. Pima Indian diabetes rules

Rule Numbers	Rules
Rule 1	IF NTP $<$ 14 & PGC $<$ 126.6508 & DPF $<$ 4.4057 & A $<$ 68.7796 THEN Class 0.
Rule 2	IF NTP $<$ 7 & PGC $<$ 140.7444 & TSFT $<$ 62.2444 & A $<$ 34.3863 THEN Class 0.
Rule 3	IF PGC $>$ 123.4773 & DBP $<$ 118.4450 & TSFT $<$ 89.9492 & DPF $<$ 1.5145 THEN Class 1.
Rule 4	IF PGC $>$ 123.0103 & TSFT $>$ 60.5536 & A $<$ 64.9686 THEN Class 1.

For this dataset 77.2% of samples were classified correctly. The classification accuracy of the proposed method is shown in Table 6 with the accuracies obtained from previous

studies in the literature [23]. As can be seen from Table 6 the proposed method has the best performance for rule extracting.

TABLE 6. Pima Indian diabetes dataset experimental results

Method	Accuracy %	Reference
Our Method	77.2	This study
SSV 5 nodes/BF	75.3 ± 4.8	WD, Ghostminer
SSV opt nodes/3CV/BF	74.7 ± 3.5	WD, Ghostminer
SSV opt prune/3CV/BS	74.6 ± 3.3	WD, Ghostminer
SSV opt prune/3CV/BF	74.0 ± 4.1	WD, Ghostminer
SSV opt nodes/3CV/BS	72.9 ± 4.3	WD, Ghostminer
SSV 5 nodes/BF	74.9 ± 4.8	WD, Ghostminer
SSV 3 nodes/BF	74.6 ± 5.2	WD, Ghostminer
CART	74.5	Stalog
DB-CART	74.4	Shang & Breiman
ASR	74.3	Ster & Dobnikar
CART	72.8	Ster & Dobnikar
C4.5	73.0	Stalog

4.3. **Wisconsin breast cancer (original) dataset.** The dataset chosen for third experiment was the Wisconsin Breast Cancer (original) Dataset from the same repository [22].

The dataset has nine attributes. There are

- Clump Thickness
- Uniformity of Cell Size
- Uniformity of Cell Shape
- Marginal Adhesion
- Single Epithelial Cell Size
- Bare Nuclei
- Bland Chromatin
- Normal Nucleoli
- Mitoses

The classes were coded as 0 and 1. The class 0 (458 samples) and class 1 (241 samples) contain the patients with benign and malignant type cancers respectively. Due to 18 of this data has missing value they have been elected from dataset. As a result 444 and 239 remaining samples were used.

The attributes have the same range values in the database and these attributes are altered between 1 and 10. The attributes have different range values in the database and these ranges can be seen in Table 7.

Then CLONALG was executed on this data. As a result 20 rules (10 rules for class 0 and 10 rules for class 1) were created. To see the rules, extracting vectors must be decoded. The examples of the extracted set of rules are presented in Table 8.

To show the performance of model, majority vote was used for determining the class of samples. 98.54% of samples were classified correctly. The classification accuracy of the proposed method is shown in Table 9 with the accuracies obtained from previous studies in the literature [23].

As can be seen from Table 9 the proposed method is the one of the best rule extractors regarding to performance.

TABLE 7. Wisconsin breast cancer (original) attribute and range values

Attribute	Range
Clump Thickness (CT)	1-10
Uniformity of Cell Size (UCSI)	1-10
Uniformity of Cell Shape (UCSH)	1-10
Marginal Adhesion (MA)	1-10
Single Epithelial Cell Size (SECS)	1-10
Bare Nuclei (BN)	1-10
Bland Chromation (BC)	1-10
Normal Nucleoli (NN)	1-10
Mitoses (M)	1-10

TABLE 8. Wisconsin breast cancer dataset (original) rules

Rule Numbers	Rules
Rule 1	IF CT $<$ 7 & UCSI $<$ 5 & BN $<$ 4 & BC $<$ 8 THEN Class 0.
Rule 2	IF CT $<$ 7 & UCSH $<$ 5 & MA $<$7 & SECS $<$ 10 & BN $<$ 6 THEN Class 0.
Rule 3	IF CT $>$ 2 & UCSH $>$ 1 & SECS $>$ 1 THEN Class 1.
Rule 4	IF UCSI $>$2 & UCSH $>$ 2 THEN Class 1.

TABLE 9. Wisconsin breast cancer dataset (original) experimental results

Method	Accuracy %	Reference
C-MLP2LN	99.0	
Our Method	98.54	This study
FSM	98.3	RA
C4.5 (decision tree)	96.0	Hamilton et al.
RIAC (prob. inductive)	95.0	Hamilton et al.

4.4. **Wisconsin breast cancer (diagnostic) dataset.** The dataset chosen for the fourth experiment was the Wisconsin Breast Cancer (diagnostic) Dataset from same repository [22].

The dataset has ten attributes. There are

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension ("coastline approximation" - 1)

The mean, standard error, and largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. The classes were coded as 0 and 1. The class 0 (357 samples) and class 1 (212 samples) contain the patients with benign and malignant type cancers respectively.

The attributes have different range values in the database and these ranges can be seen in Table 10.

TABLE 10. Wisconsin breast cancer dataset attribute and range values

Attribute	Range
Mean Radius (MR)	6.981-28.11
Standard Error of Radius (SER)	9.71-39.28
Largest Radius (LR)	43.79-188.5
Mean Texture (MT)	143.5-2501
Standard Error of Texture (SET)	0.0526-0.1634
Largest Texture (LT)	0.0194-0.3454
Mean Perimeter (MP)	0-0.4268
Standard Error of Perimeter (SEP)	0-0.2012
Largest Perimeter (LP)	0.1060-0.3040
Mean Area (MA)	0.05-0.0974
Standard Error of Area (SEA)	0.1115-2.8730
Largest Area (LA)	0.3602-4.8850
Mean Smoothness (MS)	0.7570-21.98
Standard Error of Smoothness (SES)	6.8020-542.2
Largest Smoothness (LS)	0.0017-0.0311
Mean Compactness (MC)	0.0023-0.1354
Standard Error of Compactness (SEC)	0-0.3960
Largest Compactness (LC)	0-0.0528
Mean Concavity (MC)	0.0079-0.0790
Standard Error of Concavity (SECO)	0-0.0298
Largest Concavity (LCO)	7.93-36.04
Mean Concave Points (MCP)	12.02-49.54
Standard Error of Concave Points (SECP)	50.41-251.2
Largest Concave Points (LCP)	185.2-4254
Mean Symmetry (MSY)	0.0712-0.2226
Standard Error of Symmetry (SESY)	0.0273-1.0580
Largest Symmetry (LSY)	0-1.2520
Mean Fractal Dimension (MFD)	0-0.2910
Standard Error of Fractal Dimension (SEFD)	0.1565-0.6638
Largest Fractal Dimension (LFD)	0.055-0.2075

CLONALG was executed on this data. As a result 32 rules (16 rules for class 0 and 16 rules for class 1) were created. To see the rules, extracting vectors must be decoded. The examples of the extracted set of rules are presented in Table 11.

To show the performance of model, majority vote was used for determining the class of samples. 95.61% of samples were classified correctly.

5. **Conclusions.** Mining classification rules is an important task of data mining. In this paper, a new algorithm for extracting rules has been presented. The rules have been extracted directly from the data that have been coded as real. The main contribution of this study is to determine the sensitive attribute intervals of the rules using optimization. For this aim CLONALG algorithm has been used.

The approach for extracting rules consists of four phases:

1. Code the data
2. Create the population randomly

TABLE 11. Wisconsin breast cancer dataset rules

Rule Numbers	Rules
Rule 1	IF MR $\in (4.7734, 26.6794)$ & SER < 34.8582 & LR $\in (55.4837, 127.4613)$ & MT < 1415.9272 & SET < 0.1304 & LT < 0.2515 & MP < 0.1730 & SEP < 0.2081 & LP $\in (0.0115, 0.2749)$ & MA < 0.0853 & SEA < 0.9242 & LA < 4.7234 & MS $\in (6.7904, 21.9296)$ & LS < 0.0240 & MC < 0.1727 & SEC < 0.2155 & MCP $\in (0.5041, 48.8863)$ & SECP > 57.9528 & LCP $\in (264.0809, 1731.2106)$ & SEFD $\in (0.1848, 0.3390)$ & LFD $\in (0.00468, 0.1052)$ THEN Class 0.
Rule 2	IF MT < 1999.2256 & SET < 0.1580 & LT < 0.2666 & MP < 0.3149 & MA > 0.0510 & SEA < 2.6517 & LA < 4.8260 & MS < 19.9975 & SES < 532.9538 & SEC < 0.3441 & LC < 0.0441 & MC < 0.0404 & LCO $\in (11.5835, 19.4892)$ & MCP < 39.7186 & SECP < 211.9051 & MSY < 0.1623 & SESY < 0.0042 & LSY < 0.4433 THEN Class 0.
Rule 3	IF MR > 12.1572 & LR > 90.2290 & MP > 0.0085 & LP > 0.1478 & SEA < 2.4535 & SES < 509.5871 & MC < 0.1043 & SEC $\in (0.0090, 0.1882)$ & MC < 0.0734 & MCP < 37.2139 & SECP > 66.9107 & LCP > 3592.7801 & MSY < 0.2000 & SESY > 0.0957 & LSY $\in (0.1730, 1.2167)$ & MFD > 0.0193 & SEFD > 0.1726 THEN Class 1.
Rule 4	IF MR > 8.7525 & LR > 54.8137 & SET < 0.1284 & LP > 0.0740 & MA < 0.0837 & SEA < 2.6418 & LS < 0.0285 & MC < 0.1119 & SEC < 0.2194 & LC < 0.0444 & MC < 0.0694 & SECO < 0.0260 & LCO > 11.8056 & SECP > 109.9146 & LCP > 604.4602 & MSY < 0.2184 & MFD > 0.0745 THEN Class 1.

3. Execute CLONALG
4. Decode the rules

The presented approach was applied to four real world classification problems. The data were obtained from University of California at Irvine (UCI) machine learning repository. It has been observed that the obtained results are one of the best results compared with related previous studies. The three of the datasets used in this study have two classes and one of them has three classes. Moreover, all of the datasets have real vauled features. In future works, realize the classification for multi-classing databeses which have nominal and real values.

Acknowledgment. This study is supported by the Scientific Research Projects Unit of Selcuk University. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

REFERENCES

[1] P. Andreeva, M. Dimitrova and P. Radeva, *Learning Models and Algorithms for Rule Extraction and Prediction*, www.icsr.bas.bg/ICSRwebsite/Staff/Andreeva_Plamena/AndreevaPGMsClsf_final.pdf, 2011.

[2] A. S. Kuttiyil, *Survey of Rule Extraction Methods*, Master Thesis, Wayne State University, 2004.

[3] G. Lashkia, L. Anthony and H. Koshimizu, Classification rule extraction based on relevant, irredundant attributes and rule enlargement, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol.11, no.4, 2007.
[4] L. Ozbakir, A. Baykasoglu, S. Kulluk and H. Yapıcı, TACO-miner: An ant colony based algorithm for rule extraction from trained neural networks, *Expert Systems with Applications*, vol.36, pp.12295-12305, 2009.
[5] L. E. Zárate, S. M. Dias and M. A. J. Song, FCANN: A new approach for extraction and representation of knowledge from ANN trained via formal concept analysis, *Neurocomputing*, vol.71, no.13-15, pp.2670-2684, 2008.
[6] Y. Kusunoki, M. Inuiguchi and J. Stefanowski, Rule induction via clustering decision classes, *International Journal of Innovative Computing, Information and Control*, vol.4, no.10, pp.2663-2677, 2008.
[7] R. Nayak, Generating rules with predicates, terms and variables from the pruned neural networks, *Neural Networks*, vol.22, no.4, pp.405-414, 2009.
[8] A. E. Elalfi, R. Haque and M. E. Elalami, Extracting rules from trained neural network using GA for managing E-business, *Applied Soft Computing*, vol.4, no.1, pp.65-77, 2004.
[9] M. Rodríguez, D. M. Escalante and A. Peregrín, Efficient distributed genetic algorithm for rule extraction, *Applied Soft Computing*, vol.11, no.1, pp.733-743, 2011.
[10] D. Martens, B. Baesens, T. V. Gestel and J. Vanthienen, Comprehensible credit scoring models using rule extraction from support vector machines, *European Journal of Operational Research*, vol.183, no.3, pp.1466-1476, 2007.
[11] M. A. H. Farquad, V. Ravi and S. B. Raju, Support vector regression based hybrid rule extraction methods for forecasting, *Expert Systems with Applications*, vol.37, no.8, pp.5577-5589, 2010.
[12] Y. Chen, J. Chang and C. Cheng, Forecasting IPO returns using feature selection and entropy-based rough sets, *International Journal of Innovative Computing, Information and Control*, vol.4, no.8, pp.1861-1875, 2008.
[13] J. H. Ang, K. C. Tan and A. A. Mamun, An evolutionary memetic algorithm for rule extraction, *Expert Systems with Applications*, vol.37, no.2, pp.1302-1315, 2010.
[14] S. Avner, Extraction of comprehensive symbolic rules from a multi-layer perceptron, *Engineering Applications of Artificial Intelligence*, vol.9, no.2, pp.137-143, 1996.
[15] S. N. Omkar and U. R. Karanth, Rule extraction for classification of acoustic emission signals using ant colony optimisation, *Engineering Applications of Artificial Intelligence*, vol.21, no.8, pp.1381-1388, 2008.
[16] K. Odajima, Y. Hayashi, G. Tianxia and R. Setiono, Greedy rule generation from discrete data and its use in neural network rule extraction, *Neural Networks*, vol.21, no.7, pp.1020-1028, 2008.
[17] C. C. Wong and N. S. Lin, Rule extraction for fuzzy modeling, *Fuzzy Sets and Systems*, vol.88, no.1, pp.23-30, 1997.
[18] E. Hart and J. Timmis, Application areas of AIS: The past, the present and the feature, *ICARIS, LNCS*, vol.3627, pp.483-497, 2005.
[19] J. Timmis, A. Hone, T. Stibor and E. Clarck, Theoretical advances in artficial immune systems, *Theoretical Computer Science*, vol.403, pp.11-32, 2008.
[20] J. Brownlee, Clonal selection theory & CLONALG the clonal selection classification algorithm (CSCA), *Tecnical Report*, 2005.
[21] R. S. Parpinelli, H. S. Lopes and A. A. Freitas, An ant colony based system for data mining: Applications to medical data, *Proc. of the Genetic and Evolutionary Computation Conference*, Morgan Kaufmann, San Francisco, CA, pp.791-798, 2001.
[22] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, http://archive.ics.uci.edu/ml, 2010.
[23] *Logical Rules Extracted from Data*, http://www.is.umk.pl/projects/rules.html, 2011.