

# Performance Evaluation of Evolutionary Algorithms in Classification of Biomedical Datasets

Ajay Kumar Tanwani and Muddassar Farooq

Next Generation Intelligent Networks Research Center (nexGIN RC)  
National University of Computer & Emerging Sciences (FAST-NUCES)  
Islamabad, 44000, Pakistan  
{ajay.tanwani, muddassar.farooq}@nexginrc.org

## ABSTRACT

Biomedical datasets pose a unique challenge for machine learning and data mining techniques to extract accurate, comprehensible and hidden knowledge from them. In this paper, we comprehensively investigate the role of a biomedical dataset on the classification accuracy of an algorithm. To this end, we quantify the complexity of a biomedical dataset in terms of its missing values, imbalance ratio, noise and information gain. We have performed our experiments using six well-known evolutionary rule learning algorithms: XCS, UCS, GAssist, cAnt-Miner, SLAVE and Ishibuchi, on 31 publicly available biomedical datasets. The results of our experiments show that GAssist gives better classification accuracy among the compared schemes. However, the nature of a biomedical dataset – not the selection of evolutionary algorithm – plays a major role in determining the classification accuracy of a dataset. We further show that noise is a dominating factor in determining the complexity of a dataset and it is inversely proportional to the classification accuracy of all the algorithms. The complexity of biomedical dataset will prove useful to researchers in evaluating the classification potential of their dataset for automatic knowledge extraction.

## Categories and Subject Descriptors

D.2.8 [Software Engineering]: Metrics—Complexity Measures, Performance Measures; I.1.2 [Computing Methodologies]: Algorithms—Analysis of algorithms

## General Terms

Algorithms, Experimentation, Performance

## Keywords

Classification, Evolutionary Rule Learning Algorithms, Biomedical Datasets

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO'09, July 8–12, 2009, Montréal Québec, Canada.  
Copyright 2009 ACM 978-1-60558-505-5/09/07 ...\$5.00.

## 1. INTRODUCTION

Recent advancements in the field of *bioinformatics* and *computational biology* are increasing the complexity of underlying biomedical datasets. The use of sophisticated equipments like mass spectrometers and magnetic resonance imaging (MRI) scanners generate enormous amounts of data that pose a number of issues regarding their electronic storage and efficient processing. One of the major challenges in this context is to automatically extract accurate, comprehensible, and hidden knowledge from large amounts of raw data. The discovered knowledge can then help medical experts in classification of anomalies for these datasets.

The well-known data mining techniques for knowledge extraction and classification include probabilistic methods, neural networks, support vector machines, decision trees, instance based learners, rough sets and evolutionary algorithms. The evolutionary algorithms – inspired from the evolution process in biological species – show a number of desirable properties like self-adaptation, robustness, collective learning etc., which make them suitable for challenging real world problems. The Evolutionary Computation (EC) paradigm has been successfully used in several data mining techniques including but not limited to genetic based machine learning systems (GBML), learning classifier systems (LCS), ant colony inspired classifiers, and hybrid variants of evolutionary fuzzy systems and neural networks. The evolutionary classifiers are becoming popular for data mining of medical datasets because of their ability to find hidden patterns in electronic records that are not otherwise obvious even to physicians [1].

However, the choice of a classifier is not obvious to a researcher working on the classification of biomedical datasets. Consequently, the common methodology adopted by researchers is to experiment their dataset with a few well-known machine learning techniques and select the one that gives better results. As a result, no attempt is yet made to systematically investigate the factors that define the accuracy of a classifier. In this paper, we define the complexity of a dataset in terms of missing values, imbalance ratio, noise and information gain. Moreover, we evaluate the performance of six well-known evolutionary rule learning classifiers: XCS, UCS, GAssist, cAnt-Miner, SLAVE and Ishibuchi on 31 publicly available biomedical datasets. The results of our experiments provide two valuable insights: (1) classification accuracy strongly depends on the complexity of a biomedical dataset, (2) noise in a dataset dominates its complexity. To conclude, we propose that researchers should first evaluate the complexity of their biomedical dataset using our proposed parameters in order to gain an insight about its overall classification potential.

The remaining paper is organized as follows: we introduce the evolutionary algorithms used in our study in Section 3. In Section

4, we quantify the complexity of the biomedical datasets. We report the results of our experiments which are followed by discussions in Section 5. Finally, we conclude the paper with an outlook to our future work.

## 2. RELATED WORK

We now present a brief overview of different studies that analyze the performance of evolutionary algorithms on various biomedical domains. In [3], Wong et al. applied evolutionary algorithms to discover knowledge in the form of rules and casual structures from fracture and scoliosis databases. Their results suggest that evolutionary algorithms are useful in finding interesting patterns. John Holmes in [4] presented his stimulus response learning classifier system, EpiCS, to enhance classification accuracy in an imbalanced class dataset. He, however, used artificially created liver cancer dataset. Bernado-Mansilla in [5] characterize the complexity of the classification problem by a set of geometrical descriptors and analyze the competence of XCS in this domain. The authors in [6] compared XCS with Bayesian network, SMO and C4.5 for mining breast cancer data and showed that XCS provides significantly higher accuracy along with C4.5. However its rules are considered more comprehensible and descriptive by the domain medical experts. The work in [7] evaluates two competitive learning classifier systems, XCS and UCS, for extracting knowledge from imbalanced data using both fabricated and real world problems. The results of their study prove the robustness of these algorithms compared with IBk, C4.5 and SMO. In [8], the authors compared the Pittsburgh and Michigan style classifier using XCS and GAssist on 13 publicly available datasets to reveal important differences between the two systems. The comparative study performed in [9] between evolutionary algorithms (XCS and Gale) and non-evolutionary algorithms (instance based, decision trees, rule-learning, statistical models and support vector machines) on several datasets suggest evolutionary algorithms as more suitable for data mining and classification. The results of the experiments carried in [10] show better classification accuracy for well-known ant colony inspired Ant Miner compared with C4.5 on 4 biomedical datasets. The authors in [11] have analyzed several strategies of evolutionary fuzzy models for data mining and knowledge discovery.

A common theme observed in various studies is that they are inclined towards particular classifier(s) instead of the biomedical dataset(s). In contrast, our study uses a novel methodology to quantify the complexity of a dataset, which we show, governs the accuracy of a classifier.

## 3. EVOLUTIONARY ALGORITHMS

We have selected a diverse set of well-known evolutionary rule learning algorithms for our empirical study. The selected algorithms are: (1) reinforcement learning based Michigan style XCS [18], (2) supervised learning based Michigan style UCS [19], (3) Pittsburgh style GAssist [20], (4) Ant Colony Optimization (ACO) inspired cAnt-Miner [21], (5) genetic fuzzy iterative learner SLAVE [22], and (6) genetic fuzzy classifier Ishibuchi [23].

### 3.1 XCS

XCS is an accuracy based Michigan-style classifier, that evolves a set of rules as a population of classifiers ( $P$ ). Each rule consists of a condition and an action. Each rule has three main performance parameters: (1) payoff prediction ( $p$ ), (2) prediction error ( $\epsilon$ ), and (3) fitness ( $F$ ). The first step in classification of an instance is to build a match set ( $M$ ) that consists of rules whose conditions are satisfied. The payoff prediction of each rule is computed and

its corresponding action set ( $A$ ) is created. The online learning is made possible with a reward ( $r$ ), returned by the environment, that is subsequently used to tune the performance parameters of the rules in the action set. The updated fitness is inversely proportional to the prediction error. Finally a genetic algorithm  $GA$ , with crossover and mutation probabilities  $\chi$  and  $\mu$  respectively, is applied to the rules in the action set and consequently new rules are added to the population. Some rules are also deleted from the population depending on their experience.

The parameter configuration of XCS used in our experiments is as follows: population size  $N = 6400$ , learning rate  $\beta = 0.2$ ,  $\theta_{sub} = \theta_{del} = 50$ , tournament size = 0.4,  $\chi = 0.8$ ,  $\mu = 0.04$  and the number of explorations are kept 100,000.

### 3.2 UCS

UCS is another accuracy based, Michigan-style classifier which is in principle quite similar to XCS. However, it uses a supervised learning scheme to compute fitness instead of reinforcement learning employed by XCS. UCS like XCS also evolves a population of rules ( $P$ ). Each rule has two parameters: (1) accuracy ( $acc$ ), and (2) fitness ( $F$ ). During the training phase, for every instance a set of rules whose conditions are satisfied become part of its match set ( $M$ ). The rules that perform correct classification become part of the correct set ( $C$ ), and the others become part of the incorrect set ( $!C$ ). Finally, the genetic algorithm  $GA$  is applied to the correct set to update its population. Every instance during testing is classified through weighted voting, on the basis of fitness, to select the action.

We have used default parameter settings with  $N = 6400$ , number of iterations = 100,000 and  $acc_0 = 0.99$ . The other tuning parameters of  $GA$  are kept same as that in XCS.

### 3.3 GAssist

GAssist (Genetic Algorithms based claSSifier sySTem), in contrast to XCS and UCS, is a Pittsburgh-style learning classifier in which the rules are assembled in the form of a decision list. GAssist-ADI uses Adaptive Discretization Intervals (ADI) rule representation. In such systems, the continuous space is discretized into fixed intervals for developing rules. Generalization is introduced by deleting and selecting rule set as a function of their accuracy and length. The crossover between two rules takes place across attribute boundaries rather than attribute intervals.

GAssist parameter setting is as follows: crossover probability = 0.6, number of iterations = 500, minimum number of rules for rule deletion = 12, and set of uniform discretizers with 4, 5, 6, 7, 8, 10, 15, 20 and 25 bins.

### 3.4 cAnt-Miner

Ant Miner, inspired by behavior of real ant colonies, uses Ant Colony Optimization (ACO) to construct classification rules from the training data. The Rule Discovery process consists of 3 steps i.e. rule generation, rule pruning and rule updating. In the rule generation step, an ant starts with an empty rule list and adds one term at a time based on the probability of that attribute-value pair. It continues to add terms to the rule without duplication until all the attributes are exhausted or the new terms make the rule more specific, defined by a user specified threshold. In the rule pruning step all the terms are removed one by one from the rule that degrades the accuracy of that rule. While updating rules, the pheromone values of terms are increased or decreased on the basis of their usage in the rule discovery process. cAnt-Miner is a variant of Ant Miner for real valued attributes. The parameters of cAnt-Miner are set to: number of ants = 3000, minimum cases per rule = 5, maximum number of uncovered cases = 10 and convergence test size = 10.

### 3.5 SLAVE

SLAVE (Structural Learning Algorithm in Vague Environment) is totally different from the classical Michigan-style and Pittsburgh-style rule learning algorithms. In this approach, every entity in the population represents a unique rule. But during an iteration of a genetic algorithm, only the best individual is added to the final set of rules which is eventually used for classification. In this way, SLAVE combines its iterative learning approach with the fuzzy models. The fitness of the rules is determined by their completeness and consistency.

In our experiments, the parameter configuration of SLAVE is: number of labels = 5, population size = 100, number of iterations allowed without change = 500 and mutation probability = 0.01.

### 3.6 Ishibuchi

Ishibuchi et al. proposed a fuzzy rule learning method for multi-dimensional pattern classification problem with continuous attributes. The classification is done with the help of a fuzzy-rule base in which each fuzzy if-then rule is handled as an individual, and a fitness value is assigned to each rule. The criteria for assigning a class label is based on a simple heuristic procedure which assigns a grade of certainty for each fuzzy if-then rule. Because it uses linguistic values with fixed membership functions as antecedent fuzzy sets, a linguistic interpretation of each fuzzy if-then rule is easily obtained which greatly helps in comprehending the generated solution.

The experiments are carried with default parameters using number of labels = 5, population size = 100, number of evaluations = 10,000, along with crossover and mutation probabilities of 1.0 and 0.9 respectively.

## 4. NATURE OF BIOMEDICAL DATASETS

Biomedical datasets provide a whole spectrum of difficulties – high-dimensionality, multiple classes, imbalanced classes, missing values and noisy data – that affect the classification accuracy of algorithms. The inconsistencies and inherent complexities in biomedical datasets obtained from different sources justify the need to separately investigate the impact of the nature of biomedical dataset in classification. To this end, we have selected 31 diverse biomedical datasets publicly available from UCI machine learning repository [12]. We now introduce four parameters that we use to quantify the complexity of a biomedical dataset: (1) missing values, (2) imbalance ratio, (3) noise, and (4) information gain.

### 4.1 Missing Values

A major focus of the machine learning community has been to analyze the effect of missing data on the accuracy of a classifier. The missing data is generally classified into three types: (1) missing completely at random (MCAR), (2) missing at random (MAR), and (3) not missing at random (NMAR). The datasets obtained from clinical databases contain several missing fields which can belong to all three categories of missing values. In Table 1 we see that VA-Heart dataset contains upto 27% of missing values in its attributes.

### 4.2 Imbalance Ratio

Orriols-Puig and Bernado-Mansilla compute class imbalance as the ratio between the number of majority class instances and the number of minority class instances [7]. But, this is only suitable for two-class problems as it does not include proportion of other class instances for a multi-class dataset. For example, Thyroid0387 has a total of 32 classes with 6771 majority class instances and only 1 minority class instance. The imbalance ratio, using the above method, is 6771 which definitely does not represent the true picture

because the distribution of instances of other classes is relatively uniform. Therefore, we use following definition of imbalance ratio  $I_r$  to cater for proportion of all class distributions:

$$I_r = \frac{N_c - 1}{N_c} \sum_{i=1}^{N_c} \frac{I_i}{I_n - I_i} \quad (1)$$

where  $I_r$  is in the range ( $1 \leq I_r < \infty$ ) and  $I_r = 1$  is a completely balanced dataset having equal instances of all classes.  $N_c$  is the number of classes,  $I_i$  is the number of instances of class  $i$  and  $I_n$  is the total number of instances. Hyperthyroid is the most imbalanced dataset in our repository with an imbalance ratio of 28.81.

### 4.3 Noise

Noise is of two types: (1) attribute noise, and (2) class noise. Research has shown that the impact of class noise on classification accuracy is significantly more as compared to the attribute noise [14] and hence, we only quantify class noise in our study. The common sources of class noise are inconsistent and mislabeled instances. A number of research efforts have been made to quantify the level of noise in a dataset, but its definition still remains subjective. Brodley and Friedl characterized noise as the proportion of incorrectly classified instances by a set of trained classifiers [15]. We use a similar approach to quantify noise but utilize confusion matrices for a set of classifiers to determine noisy instances. Noise is then quantified as the sum of all off-diagonal entities where each entity is the minimum of all the corresponding elements in a set of confusion matrices. The defined criteria is based upon two assumptions: (1) an inconsistent or misclassified instance is likely to confuse every classifier, and (2) the bias of an algorithm towards particular class instances can be factored out by using a set of classifiers. The advantage of our approach is that we separately identify misclassified instances of every class and only categorize those as noisy which are misclassified by all the classifiers.

The confusion matrix of a  $n^{th}$  classifier in a set of  $n$  classifiers can in general be represented as:

$$C_n = \begin{pmatrix} i_{11}^n & i_{12}^n & \cdot & \cdot & \cdot & i_{1j}^n \\ i_{21}^n & i_{22}^n & \cdot & \cdot & \cdot & i_{2j}^n \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ i_{i1}^n & i_{i2}^n & \cdot & \cdot & \cdot & i_{ij}^n \end{pmatrix}$$

where the diagonal elements in  $C_n$  represent the correctly classified instances and off-diagonal elements are the incorrectly classified instances. The percentage of class noise in a dataset of  $I_n$  instances can be computed as below:

$$Noise = \left( \frac{1}{I_n} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \min(C_1(i, j), C_2(i, j), \dots, C_n(i, j)) \right) 100 \quad (2)$$

where  $i \neq j$  and  $\min(C_1(i, j), C_2(i, j), \dots, C_n(i, j))$  is an entity for corresponding  $i$  and  $j$  that represents minimum number of class instances misclassified by all the classifiers. We have used five well-known and diverse machine learning algorithms as a set of classifiers in our study: Naive Bayes (probabilistic), SMO (support vector machines), J48 (decision trees), Ripper (inductive rule learner) and IBk (instance based learner). We use the standard implementations of these schemes in Wakaito Environment for Knowledge Acquisition (WEKA) [13]. It is evident from Table 1 that biomedical datasets are generally associated with high percentage of noise levels.

## 4.4 Information Gain

Information gain is an information-theoretic measure that evaluates the quality of attributes in a dataset. It measures the reduction in uncertainty if the values of an attribute are known. For a given attribute  $X$  and a class attribute  $Y$ , the uncertainty is given by their respective entropies  $H(X)$  and  $H(Y)$ . Then the information gain of  $X$  with respect to  $Y$  is given by  $I(Y; X)$ , where

$$I(Y; X) = H(Y) - H(Y|X)$$

The average and total information gain of a biomedical dataset shown in Table 1 give a measure of the quality of its attributes for classification.

## 5. RESULTS AND DISCUSSIONS

We now present the results of our experiments that we have done to analyze the nature of 31 biomedical datasets with six evolutionary algorithms. We have used the standard ACO framework, MYRA [16], for cAnt-Miner and Knowledge Extraction based on Evolutionary Learning (KEEL) [17] for evolutionary classifiers to remove any implementation bias in our study. We evaluate the classification accuracy of the evolutionary algorithms using standard ten fold stratified cross-validation in order to ensure systematic and unbiased analysis. The results summarized in Table 1 show the nature of a dataset in terms of its quantified parameters, along with the resulting classification accuracies of all the algorithms. We now systematically discuss the role of evolutionary learning paradigm and the nature of biomedical dataset on classification accuracy.

### 5.1 Effect of Evolutionary Algorithm

#### 5.1.1 Pittsburgh-Style – GAssist

The results of our experiments show that **GAssist** – a Pittsburgh-style learning classifier – performs better than other evolutionary rule-learning algorithms. It provides the best overall classification accuracy of 77.33 and least standard deviation of 16.63. Moreover, it also outperformed other classifiers for 13 biomedical datasets. The greater accuracy is a result of its superior fitness function that combines the accuracy and complexity of an individual using Minimum Description Length (MDL) principle to yield optimum rules [20].

#### 5.1.2 Nature Inspired – cAnt-Miner

**cAnt-Miner** ranks second and it also closely follows **GAssist**'s policy to generate simpler rules. The ants generate rules by selecting attribute-value pairs on the basis of their entropy and pheromone values [2]. Consequently, it uses only high quality attributes (we model quality with information gain) in the formulation of its rules. Moreover, its pruning mechanism yields simpler and shorter rules, thereby, achieving greater classification accuracy.

#### 5.1.3 Michigan-Style – UCS and XCS

The Michigan-style learning classifiers – **UCS** and **XCS** – use online learning to evolve a set of condition-action rules from each training instance. Thus, they can be more useful in identifying hidden patterns and generating information rich rules compared with simple and generic rules of **GAssist** and **cAntMiner**. We therefore suggest that if medical experts are available to refine rules, Michigan-style classifier for knowledge extraction can prove to be useful.

#### 5.1.4 Genetic Fuzzy – SLAVE and Ishibuchi

The results in Table 1 show that the genetic fuzzy rule learning classifiers are not generally suitable for classification of biomedical datasets. The fuzzy rules so generated, however, can be particularly used to evaluate the uncertainty associated with the prognosis.

### 5.2 Effect of Nature of Dataset

A careful insight into the results enables a reader to draw an important conclusion: *the variance in accuracy of classifiers on a particular dataset is significantly smaller compared with the variance in accuracy of the same classifier on different datasets*. The statement holds for more than 25 datasets; with notable exceptions being Dermatology, Splice-Junction Gene Sequence, and Promoters Gene Sequence. Consequently, we can say that accuracy is strongly dependent on the nature of biomedical dataset. We now discuss important factors that determine the net classification potential of a dataset.

#### 5.2.1 Role of Multiple Classes

We conclude from Table 1 that for multi-class problems, **UCS** gives significantly better accuracy compared with other classifiers. The reason is that it evolves only those highly-rewarded classifiers of the match set in the correct set, which predict the same class as that of the training example [24]. In comparison, **GAssist** has serious problems in dealing with multi-class problems – specially when the when the number of output classes are more than 5. On these datasets (Cardiac Arrhythmia, Dermatology, E-Coli and Thyroid0387), the average classification accuracy of **UCS** is 83.49% compared with 75.52% of **GAssist**.

#### 5.2.2 Role of Instances

It is obvious in Table 1 that evolutionary algorithms over-fit on datasets with small number of instances. Consequently, accuracy of classifiers on Lung Cancer, Post Operative Patient, Promoters Gene Sequence and Switzerland Heart datasets severely degrades. The evolutionary classifiers during training create *small disjuncts* with rare cases [25]; and as a result, their accuracy significantly degrades during testing.

#### 5.2.3 Role of Attributes

The attributes of a dataset vary in three aspects: (1) number, (2) type (continuous, binary and nominal), and (3) quality. We see in Table 1 that number and type of attributes have little role in defining the classification potential of a dataset. Very poor performance of **XCS** on Splice-Junction Gene Sequence, Promoters Genes Sequence and Lung Cancer datasets came as a surprise to us. Our analysis reveal that large number of nominal attributes in these datasets – 61, 58 and 56 respectively – is the main cause of their poor performance with **XCS**. Our conclusion is that **XCS** is unable to cater for large number of nominal attributes in a dataset.

Remember, we quantify quality of attributes with information gain. The graph in Figure 1 clearly shows that classification accuracy and hence, the classification potential of a dataset increases with an increase in the information gain of its attributes.

#### 5.2.4 Role of Missing Values

The missing or incomplete data degrades the accuracy of learning algorithms. Therefore, a number of methods like *Wild-to-Wild*, mean or mode method, InGrimputation Model, listwise deletion etc. have been proposed. Figure 2 reveals that **GAssist** is relatively more resilient to missing values compared with other algorithms. **GAssist** replaces a missing value with the mean of its class for real valued attributes and with their mode for nominal attributes.

**Table 1: The Table shows: (1) Summary of used datasets in alphabetical order; number of instances, classes, attributes (continuous, binary, nominal), percentage of missing values in the attributes, noise, average information gain (Avg Info Gain) and total information gain (Net Info Gain). (2) Classification accuracies of evolutionary rule-learning algorithms; bold entries in every row represents the best accuracy.**

Dataset	Nature of Dataset										Evolutionary Rule Learning Classifiers								Mean
	Instances	Classes	Attributes		Missing Values	Noise	Imb Ratio	Avg Info Gain	Net Info Gain	XCS	UCS	GAssist	can-Miner	SLAVE	Ishibuchi				
			Con	Bin												Nom			
Am-Thyroid	7200	3	6	15	0	0	0.11	8.37	0.037	0.78	97.08	96.99	94.67	<b>99.15</b>	93.29	97.61	95.63 ± 2.52		
Breast Cancer	699	2	1	0	9	0	2.72	1.21	0.451	4.51	96.14	<b>96.57</b>	94.56	93.56	94.70	94.71	95.04 ± 1.11		
Breast Cancer Diagnostic	569	2	31	0	0	0	2.11	1.14	0.303	9.39	93.67	92.44	<b>95.43</b>	93.15	91.56	92.09	93.06 ± 1.38		
Breast Cancer Prognostic	198	2	33	0	0	0	0.06	1.76	0.004	0.15	65.76	72.82	70.29	73.82	74.29	<b>76.29</b>	72.21 ± 3.72		
Cardiac Arrhythmia	452	16	272	7	0	0	0.32	1.28	0.047	13.06	-	61.31	54.86	<b>67.92</b>	65.49	-	62.39 ± 5.72		
Cleveland-Heart	303	5	10	3	0	0	0.15	1.78	0.115	1.49	<b>58.09</b>	52.15	57.41	57.74	57.74	54.44	54.78 ± 3.71		
Contraceptive Method	1473	3	2	3	4	0	31.98	1.04	0.041	0.36	53.43	47.32	<b>55.54</b>	50.92	25.46	48.85	43.58	46.04 ± 10.95	
Dermatology	366	6	1	1	32	0	0.06	0.82	0.442	15.02	94.84	<b>96.99</b>	92.64	91.00	3.83	30.60	68.32 ± 40.53		
Echocardiogram	132	2	8	2	2	0	4.67	6.06	1.24	0.084	88.63	84.78	<b>96.21</b>	83.19	92.47	93.24	89.75 ± 5.10		
E-Coli	336	8	7	0	1	0	6.55	1.25	0.678	5.42	90.51	<b>93.73</b>	74.74	79.17	82.72	67.89	81.46 ± 9.68		
Haberman's Survival	306	3	0	0	0	0	16.67	1.57	0.023	0.07	<b>74.23</b>	74.20	69.96	71.53	73.18	73.20	72.72 ± 1.67		
Hepatitis	155	2	6	0	13	0	5.67	10.97	2.05	0.058	81.33	81.29	<b>91.50</b>	80.00	81.96	80.04	82.69 ± 4.39		
Horse Colic	368	2	8	4	15	0	19.39	11.96	1.15	1.64	84.23	81.47	<b>93.73</b>	83.97	67.33	63.05	78.96 ± 11.54		
Hungarian Heart	294	5	10	3	0	0	20.46	13.61	1.74	1.02	65.98	62.26	<b>75.14</b>	62.95	64.60	63.95	65.81 ± 4.75		
Hyper-Thyroid	3772	5	7	21	1	0	2.17	0.34	0.012	0.36	97.35	97.88	<b>98.57</b>	98.12	97.43	97.30	97.77 ± 0.51		
Hypo-Thyroid	3163	2	7	18	0	0	6.74	0.54	0.024	0.60	97.16	97.85	<b>99.43</b>	98.96	95.51	95.23	97.36 ± 1.74		
Liver Disorders	345	2	6	0	0	0	21.88	1.05	0.011	0.06	63.26	<b>67.27</b>	61.18	65.48	58.54	58.27	62.33 ± 3.67		
Lung Cancer	32	3	0	0	0	0	0.28	9.86	1.02	0.152	30.83	44.99	41.67	<b>45.83</b>	-	-	40.83 ± 6.90		
Lymph Nodes	148	4	3	9	6	0	10.81	1.46	0.138	2.48	79.19	<b>81.09</b>	78.57	77.81	69.57	72.90	76.52 ± 4.36		
Mammographic Masses	961	2	1	0	4	0	3.37	1.45	0.193	0.97	80.75	82.21	<b>83.25</b>	81.06	65.56	66.60	76.57 ± 8.18		
New Thyroid	215	3	5	0	0	0	2.79	1.78	0.602	3.01	<b>94.93</b>	92.60	92.19	90.24	91.23	86.15	91.22 ± 2.94		
Pima Indians Diabetes	768	2	8	0	0	0	20.18	1.20	0.064	0.52	73.71	74.76	72.15	<b>75.00</b>	72.67	68.62	72.81 ± 2.34		
Post-Operative Patient	90	3	0	0	8	0	0.44	30.00	1.90	0.016	70.00	63.33	61.11	60.00	70.00	<b>71.11</b>	65.93 ± 5.00		
Promoters Genes	106	2	0	0	58	0	4.72	1.00	0.078	4.51	2.82	<b>76.27</b>	62.91	75.45	27.27	-	48.94 ± 32.56		
Protein Data	21618	3	0	0	1	0	45.48	1.19	0.065	0.07	51.41	51.21	<b>54.52</b>	54.46	<b>54.52</b>	54.52	53.44 ± 1.65		
Sick	2800	2	7	21	1	0	2.24	0.71	0.013	0.37	93.89	<b>97.57</b>	97.32	97.18	93.86	93.89	95.62 ± 1.91		
Splice-Junction Gene Sequence	3190	3	0	0	61	0	4.6	1.15	0.022	3.94	5.60	57.30	<b>92.45</b>	83.80	52.55	-	58.34 ± 34.01		
Starlog Heart	270	2	7	3	3	0	15.19	1.03	0.092	1.19	80.74	<b>83.33</b>	81.11	75.19	72.22	73.33	77.65 ± 4.65		
Switzerland Heart	123	5	10	3	0	0	17.07	1.14	0.023	0.30	31.67	31.67	<b>65.83</b>	30.19	31.79	42.37	38.92 ± 13.91		
Thyroid387	9172	32	7	21	1	0	5.50	2.99	0.091	2.64	74.13	81.92	79.83	<b>85.47</b>	76.46	74.02	78.64 ± 4.59		
VA-Heart	200	5	10	3	0	0	26.85	1.04	0.023	0.30	32.00	28.99	<b>58.50</b>	29.00	20.00	33.50	33.66 ± 13.04		
<b>Mean</b>	1930	4.5	15	4	9	0	3.73	12.49	2.97	0.13	70.11 ± 26.71	74.34 ± 19.95	<b>77.33</b> (1)	74.56 ± 18.76	66.96 ± 24.92	70.87 ± 19.21			

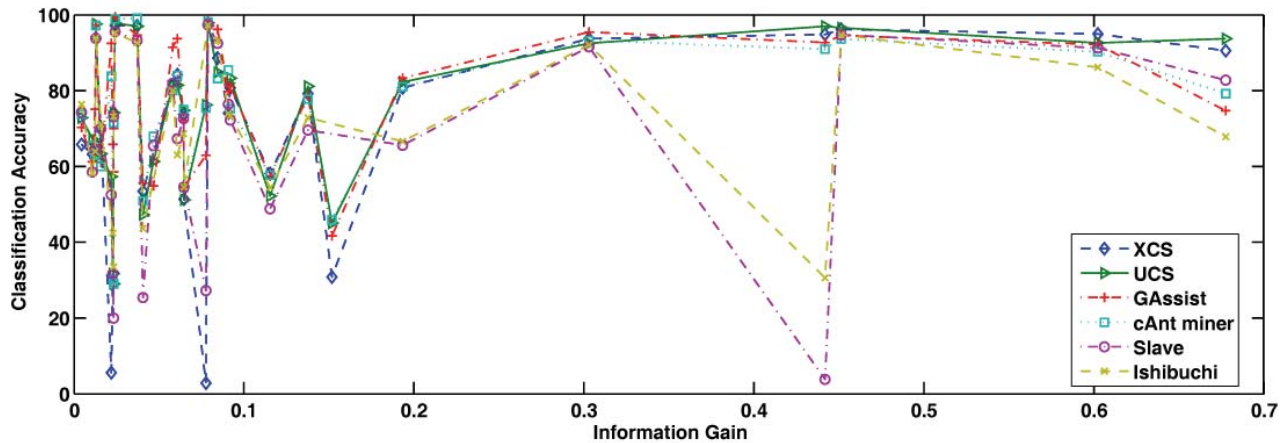


Figure 1: Average Information Gain vs Classification Accuracy

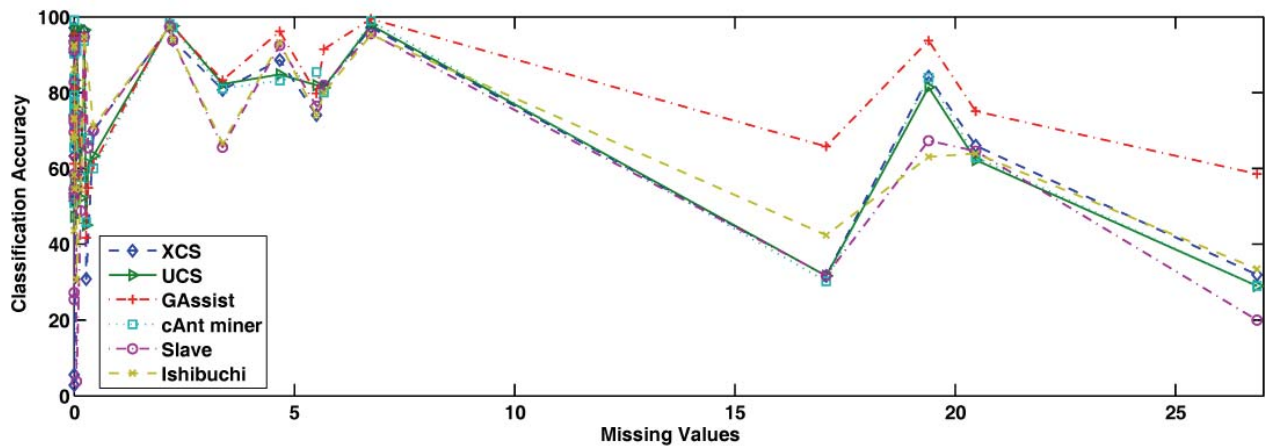


Figure 2: Missing Values vs Classification Accuracy

### 5.2.5 Role of Imbalanced Classes

A learning algorithm during classification may develop a bias towards its majority class. However, Figure 3 shows that the net accuracy of evolutionary classifiers remains unaffected even in datasets with high imbalance ratios.

### 5.2.6 Role of Noise

Classification potential of a dataset is inversely proportional to level of noise in a dataset. Consequently, accuracy of classifying noisy datasets is very small (see Figure 4). GAssist shows more resilience to noise in datasets because of its added generalization pressure with *bloat control* based on MDL principle. The MDL principle forces GAssist to reduce the size and length of its individuals. In short its ‘simple’ evolution policy makes it resilient to noise.

### 5.2.7 Combined Effect of Nature of Dataset

Our facet-wise study of dataset parameters show that noise, information gain and missing values play a significant role in defining the classification accuracy of an algorithm. We now conclude our findings in Figure 5 to have a better understanding of the combined effect. It is obvious in Figure 5 that noise in a dataset effectively de-

termines the classification accuracy and it is inversely proportional to the classification accuracy of a biomedical dataset. The figure depicts that high average information gain of a dataset yields better classification accuracy, however, low average information gain of a dataset can produce unpredictable results. The percentage of missing values has minor impact on the overall classification accuracy but the accuracy significantly degrades when the percentage of missing values in a dataset are more than 15%. The combined effect of dataset parameters reveal that the classification potential of a dataset can be estimated by analyzing its complexity using our proposed parameters. This can help researchers to take necessary preventive measures in pre-processing the dataset to ensure better classification accuracy.

## 6. CONCLUSION

In this paper, we have quantified the complexity of biomedical datasets in terms of missing values, noise, imbalance ratio and information gain. The effect of complexity on classification accuracy is evaluated using six well-known evolutionary rule learning algorithms. The results of our experiments show that GAssist – in most of the datasets – provides better classification accuracy compared with other algorithms. Our analysis reveals that the classification

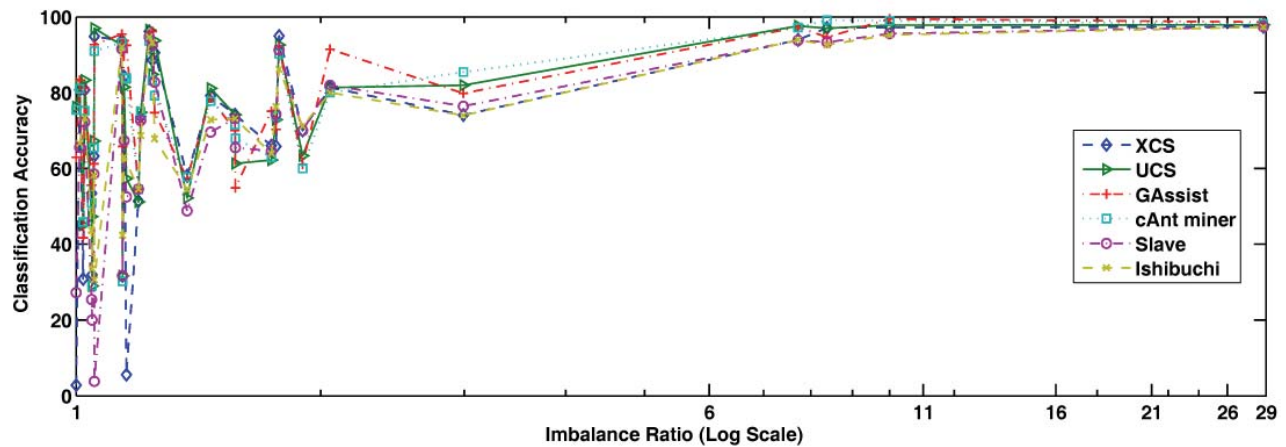


Figure 3: Class Imbalance (Log Scale) vs Classification Accuracy

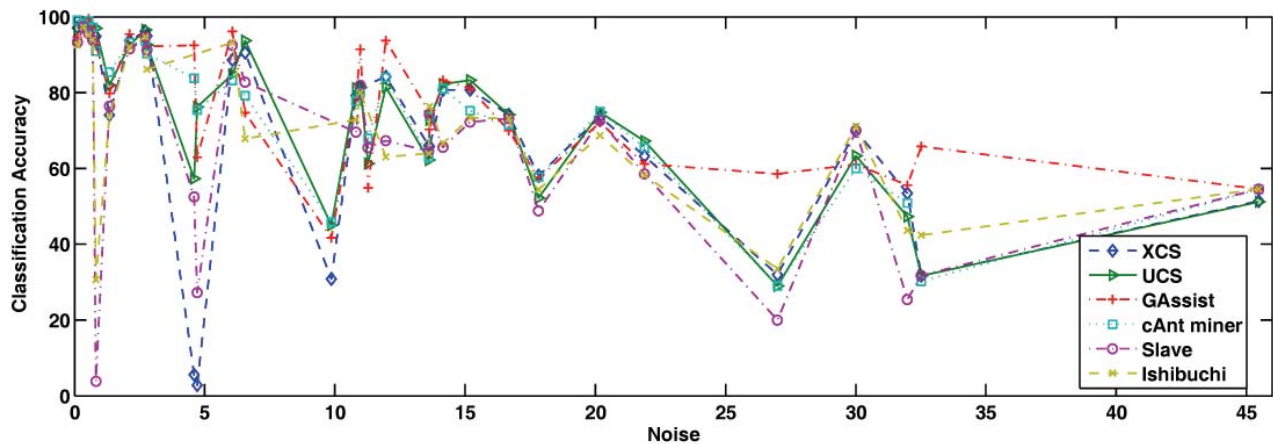


Figure 4: Noise vs Classification Accuracy

accuracy of a biomedical dataset is, however, a function of the nature of biomedical dataset rather than the choice of a particular evolutionary learner. The major contribution of this paper is a unique methodology to determine the classification potential of a dataset by analyzing the complexity of dataset. In the future, we would like to present generated rules of different classifiers to the medical specialists for their expert feedback.

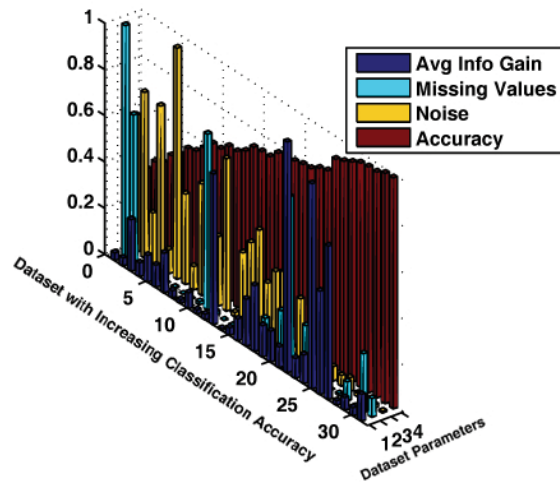
## Acknowledgements

The authors of this paper are supported, in part, by the National ICT R&D Fund, Ministry of Information Technology, Government of Pakistan. The information, data, comments, and views detailed herein may not necessarily reflect the endorsements of views of the National ICT R&D Fund.

## 7. REFERENCES

- [1] C.A. Pena-Reyes, M. Sipper, "Evolutionary computation in medicine: an overview", *Journal of Artificial Intelligence in Medicine*, 19(1), pp. 1-23, 2000.
- [2] R.S. Parpinelli, H.S. Lopes, A.A. Freitas, "Data mining with an ant colony optimization algorithm", *IEEE Transactions on Evolutionary Computation*, 6(4), pp. 321-332, 2002.
- [3] M.L. Wong, W. Lam, K.S. Leung, P.S. Ngan, J.C.V. Cheng, "Discovering knowledge from medical databases using evolutionary algorithms", *IEEE Engineering in Medicine and Biology*, 19(4), pp. 45-55, 2000.
- [4] J.H. Homes, "Learning classifier systems applied to knowledge discovery in clinical research databases", *International Workshop on Learning Classifier Systems (IWLCS)*, pp. 243-261, 2000.
- [5] E. Mansilla, "Domain of competence of XCS classifier system in complexity measurement space", *IEEE Trans. on Evolutionary Computation*, 9(1), pp. 82-104, 2005.
- [6] F. Kharbat, L. Bull, M. Odeh, "Mining breast cancer data with XCS", *Genetic and Evolutionary Computation Conference (GECCO)*, pp. 2066-2073, UK, 2007.
- [7] A.O.- Puig, E.B.-Mansilla, "Evolutionary rule-based systems for imbalanced data sets", *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 13(3), pp. 213-225, Springer, 2009.
- [8] J. Bacardit, M.V. Butz, "Data mining in learning classifier systems: comparing XCS with GAssist", *International Workshop on Learning Classifier Systems (IWLCS)*, pp. 282-290, UK, 2007.





**Figure 5: Relationship between Classification Accuracy and Nature of Dataset: x-axis contains biomedical datasets in increasing order of their classification accuracies; y-axis contains normalized parameters of datasets, 1-Average Information Gain, 2-Missing Values, 3-Noise, 4-Classification Accuracy.**

- [9] E. Bernado, X. Llorca, J.M. Garrell, "XCS and GALE: a comparative study of two learning classifier systems with six other learning algorithms on classification tasks", International Workshop on Learning Classifier Systems (IWLCS), pp. 115-132, Spain, 2002.
- [10] R.S. Parpinelli, H.S. Lopes, A.A. Freitas, "An ant colony based system for data mining: applications to medical data", Int. Conf. on Knowledge Discovery and Data mining, pp. 55-62, Boston, 2000.
- [11] M. Galea, Q. Shen, J. Levine, "Evolutionary approaches to fuzzy modelling for classification", Knowledge Engineering Review, 19(1), pp. 27-59, 2004.
- [12] UCI repository of machine learning databases, University of California-Irvine, Department of Information and Computer Science. [www.ics.uci.edu/~mllearn/MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html)
- [13] I.H. Witten et.al. "Data mining: practical machine learning tools and techniques", Morgan Kaufmann, 2nd ed., 2005.
- [14] X. Zhu, X. Wu, "Class noise vs. attribute noise: a quantitative study of their impacts, Artificial Intelligence Review, 22(3), pp. 177-210, 2004.
- [15] C.E. Brodley, M.A. Friedl, "Identifying mislabeled training data", Journal of Artificial Intelligence Research, 11, pp. 131-167, 1999.
- [16] F.E.B. Otero, Ant Colony Optimization Framework, MYRA. <http://sourceforge.net/projects/myra/>
- [17] J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesús, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera, "KEEL: a software tool to assess evolutionary algorithms for data mining problems", Soft Computing, 13, pp. 307-318, Springer, 2008.
- [18] M.V. Butz, T. Kovacs, P.L. Lanzi, S.W. Wilson, "Toward a theory of generalization and learning in XCS", IEEE Transactions on Evolutionary Computation, 8(1), pp. 28-46, 2004.
- [19] E. B. Mansilla, J.M. Garrell-Guiu, "Accuracy-based learning classifier systems: models, analysis and applications to classification tasks", Evolutionary Computation, 11(3), pp. 209-238, MIT Press, 2006.
- [20] J. Bacardit, J.M. Garrell, "Bloat control and generalization pressure using the minimum description length principle for a Pittsburgh approach Learning Classifier System", International Workshop on Learning Classifier Systems (IWLCS), pp. 59-79, UK, 2007.
- [21] F.E.B. Otero, A.A. Freitas, C.J. Johnson, "cAnt-Miner: an ant colony classification algorithm to cope with continuous attributes", Ant Colony Optimization and Swarm Intelligence, pp. 48-59, Belgium, 2008.
- [22] A. Gonzalez, R. Perez, "SLAVE: a genetic learning system based on an iterative approach", IEEE Transaction on Fuzzy Systems, 7(2), pp. 176-191, 1999.
- [23] H. Ishibuchi, T. Nakashima, T. Murata, "Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems", IEEE Transactions on Systems, Man, and Cybernetics, 29(5), pp. 601-618, 1999.
- [24] A. Orriols-Puig, E. Bernado-Mansilla, "Revisiting UCS: description, fitness sharing and comparison with XCS", International Workshop on Learning Classifier Systems (IWLCS), pp. 96-116, 2008.
- [25] T. Jo, N. Japkowicz, "Class imbalances versus small disjuncts", SIGKDD Explor. Newsl., 6(1), pp. 40-49, 2004.