Classification Potential vs. Classification Accuracy: A Comprehensive Study of Evolutionary Algorithms with Biomedical Datasets

Ajay Kumar Tanwani and Muddassar Farooq

Next Generation Intelligent Networks Research Center (nexGIN RC)
National University of Computer & Emerging Sciences (FAST-NU)
Islamabad, 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 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 and statistical analysis show that GAssist gives better classification results on majority of biomedical datasets among the compared schemes but cannot be categorized as the best classifier. Moreover, our analysis reveals that 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 evaluated algorithms. Towards the end, we provide researchers with a meta-classification model that can be used to determine the classification potential of a dataset on the basis of its complexity measures.

Keywords: Classification, Evolutionary Rule Learning Algorithms, Biomedical Datasets, Performance Measures.

1 Introduction

Recent advancements in the field of *bioinformatics* and *computational biology* are increasing the complexity of underlying biomedical datasets. The use of sophisticated equipment like mass spectrometers and magnetic resonance imaging (MRI) scanners generate large amounts of data that pose a number of issues regarding 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.

Well-known data mining techniques for knowledge extraction and classification include probabilistic methods, neural networks, support vector machines, decision trees,

J. Bacardit et al. (Eds.): IWLCS 2008/2009, LNAI 6471, pp. 127–144, 2010.

instance based learners, rough sets and evolutionary algorithms. The evolutionary algorithms – inspired from the evolution process in the 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, it is not obvious to a researcher working on the classification of biomedical datasets to choose a suitable classifier. Consequently, the common methodology adopted by researchers is to empirically evaluate their dataset with a few well-known machine learning techniques and select the one that gives better results. As a result, no attempt is made to systematically investigate the factors that define the accuracy of a classifier. An important contribution of this paper is that the accuracy of a classifier depends on the complexity of a datset. 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, and (2) noise of a dataset predominately defines its complexity. To conclude, we propose that researchers should first evaluate the complexity of their medical dataset and then use our proposed meta-model to determine its 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 statistical analysis and 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 [2], 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 [3] 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 [4] characterized the complexity of the classification problem by a set of geometrical descriptors and analyzed the competence of XCS in this domain. The authors in [5] compared XCS with Bayesian network, SMO and C4.5 for mining breast cancer data and showed that XCS provides significantly higher accuracy followed by C4.5. However its rules are considered more comprehensible and descriptive by the

domain experts. The work in [6] 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 [7], 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 [8] 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 suggests evolutionary algorithms as more suitable for data mining and classification. The results of the experiments carried in [9] show better classification accuracy for well-known ant colony inspired, Ant-Miner, compared with C4.5 on 4 biomedical datasets. The authors in [10] have analyzed several strategies of evolutionary fuzzy models for data mining and knowledge discovery. In our earlier work [11], we provide several guidelines to select a suitable machine learning scheme for classification of biomedical datasets, however, the work is limited to non-evolutionary algorithms.

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, defines the accuracy of a classifier. Moreover, we also build a meta-model of our findings that can be used to determine the classification potential of a biomedical dataset.

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 [12], (2) supervised learning based Michigan style UCS [13], (3) Pittsburgh style GAssist [14], (4) Ant Colony Optimization (ACO) inspired cAnt-Miner [15], (5) genetic fuzzy iterative learner SLAVE [16], and (6) genetic fuzzy classifier Ishibuchi [17]. In all our experiments, the parameters are selected to achieve the best operating point on the ROC (Receiver Operating Characteristic) curve [18].

3.1 XCS

XCS is a reinforcement learning based Michigan-style classifier that evolves a set of rules as a population of classifiers (P). Each rule consists of a condition, an action and three performance parameters: (1) payoff prediction (p), (2) prediction error (ϵ), and (3) fitness (F). The first step in classification 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 χ and μ 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 β = 0.2, θ_{sub} = θ_{del} = 50, tournament size = 0.4, χ = 0.8, μ = 0.04 and the number of explorations are kept at $100,000$.

3.2 UCS

UCS is an 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 following parameter settings: 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 discreteness – 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: the 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: the 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 multidimensional 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 the following parameters: the number of labels = 5, population size = 100, number of evaluations = $10,000$, along with crossover and mutation probabilities of 1.0 and 0.9.

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 [19]. 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 up to 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 [6]. 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:

$$I_r = \frac{N_c - 1}{N_c} \sum_{i=1}^{N_c} \frac{I_i}{I_n - I_i} \tag{1}$$

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 [20] 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 [21]. 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 (incorrectly classified instances) 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 nth classifier in a set of n classifiers can in general be represented as:

$$\mathbf{C_n} = \begin{pmatrix} i^n_{11} & i^n_{12} & \cdots & i^n_{1j} \\ i^n_{21} & i^n_{22} & \cdots & i^n_{2j} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ i^n_{i1} & i^n_{i2} & \cdots & i^n_{ij} \end{pmatrix}$$

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:

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

where $i \neq j$ and $min(C_1(i,j), C_2(i,j).....C_n(i,j))$ 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 IB*k* (instance based learner). We use the standard implementations of these schemes in Wakaito Environment for Knowledge Acquisition (WEKA) [22]. 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 [22]. 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

$$I(Y;X) = H(Y) - H(Y|X) \quad (3)$$

The average and total information gain of a biomedical dataset shown in Table 1 gives a direct 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 [23], for cAnt-Miner and Knowledge Extraction based on Evolutionary Learning (KEEL) [24] for other 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 provide the insights of the obtained results using statistical procedures to analyze the effect of evolutionary learning paradigm and then discuss in detail the role of nature of biomedical dataset on classification accuracy.

5.1 Statistical Analysis of Results

In this section, we provide the statistical analysis of the results obtained in Table 1 to systematically quantify the performance of evolutionary algorithms. The common approach used by many researchers in such cases is to use pairwise comparisons between all the classifiers using commonly used statistical tests such as paired t-test or wilcoxon

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	cAnt-Miner	SLAVE	Ishibuchi	
			Con	Bin	Nom												
Ann-Thyroid	7200	3	6	15	0	0	0.11	8.37	0.037	0.78	97.08	96.99	94.67	**99.15**	93.29	92.61	95.63 ± 2.52
Breast Cancer	699	2	1	0	9	0.23	2.72	1.21	0.451	4.51	96.14	**96.57**	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	**95.43**	93.15	91.56	92.09	93.06 ± 1.38
Breast Cancer Prognostic	198	2	33	0	0	0.06	13.64	1.76	0.004	0.15	65.76	72.82	70.29	73.82	74.29	**76.29**	72.21 ± 3.72
Cardiac Arrhythmia	452	16	272	7	0	0.32	11.28	1.57	0.047	13.06	-	61.31	54.86	**67.92**	65.49	-	62.39 ± 5.72
Cleveland-Heart	303	5	10	3	0	0.15	17.82	1.37	0.115	1.49	**58.09**	52.15	57.41	57.74	48.85	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	**55.54**	50.92	25.46	43.58	46.04 ± 10.95
Dermatology	366	6	1	1	32	0.06	0.82	1.05	0.442	15.02	94.84	**96.99**	92.64	91.00	3.83	30.60	68.32 ± 40.53
Echocardiogram	132	2	8	2	2	4.67	6.06	1.24	0.084	1.01	88.63	84.78	**96.21**	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	**93.73**	74.74	79.17	82.72	67.89	81.46 ± 9.68
Haberman's Survival	306	3	3	0	0	0	16.67	1.57	0.023	0.07	**74.23**	74.20	69.96	71.53	73.18	73.20	72.72 ± 1.67
Hepatitis	155	2	6	0	13	5.67	10.97	2.05	0.058	1.10	81.33	81.29	**91.50**	80.00	81.96	80.04	82.69 ± 4.39
Horse Colic	368	2	8	4	15	19.39	11.96	1.15	0.061	1.64	84.23	81.47	**93.73**	83.97	67.33	63.05	78.96 ± 11.54
Hungarian Heart	294	5	10	3	0	20.46	13.61	1.74	0.079	1.02	65.98	62.26	**75.14**	62.95	64.60	63.95	65.81 ± 4.75
Hyper Thyroid	3772	5	7	21	1	2.17	0.34	28.81	0.012	0.36	97.35	97.88	**98.57**	98.12	97.43	97.30	97.77 ± 0.51
Hypo-Thyroid	3163	2	7	18	0	6.74	0.54	9.99	0.024	0.60	97.16	97.85	**99.43**	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	**67.27**	61.18	65.48	58.54	58.27	62.33 ± 3.67
Lung Cancer	32	3	0	0	56	0.28	9.86	1.02	0.152	8.50	30.83	44.99	41.67	**45.83**	-	-	40.83 ± 6.90
Lymph Nodes	148	4	3	9	6	0	10.81	1.46	0.138	2.48	79.19	**81.09**	78.57	77.81	69.57	72.90	76.52 ± 4.36
Mammographic Masses	961	2	1	0	4	3.37	14.15	1.01	0.193	0.97	80.75	82.21	**83.25**	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	**94.93**	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	**75.00**	72.67	68.62	72.81 ± 2.34
Post Operative Patient	90	3	0	0	8	0.44	30.00	1.90	0.016	0.13	70.00	63.33	61.11	60.00	70.00	**71.11**	65.93 ± 5.00
Promoters Genes Sequence	106	2	0	0	58	0	4.72	1.00	0.078	4.51	2.82	**76.27**	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	**54.52**	54.46	**54.52**	**54.52**	53.44 ± 1.65
Sick	2800	2	7	21	1	2.24	0.71	7.72	0.013	0.37	93.89	**97.57**	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	**92.45**	83.80	52.55	-	58.34 ± 34.01
Statlog Heart	270	2	7	3	3	0	15.19	1.03	0.092	1.19	80.74	**83.33**	81.11	75.19	72.22	73.33	77.65 ± 4.65
Switzerland Heart	123	5	10	3	0	17.07	32.52	1.14	0.023	0.30	31.67	31.67	**65.83**	30.19	31.79	42.37	38.92 ± 13.91
Thyroid0387	9172	32	7	21	1	5.50	1.35	2.99	0.091	2.64	74.13	81.92	79.83	**85.47**	76.46	74.02	78.64 ± 4.59
VA-Heart	200	5	10	3	0	26.85	27.00	1.04	0.023	0.30	32.00	28.99	**58.50**	29.00	20.00	33.50	33.66 ± 13.04
Mean	1930	4.5	15	4	9	3.73	12.49	2.97	0.13	2.74	70.11 $^{(5)}$ ± 26.71	74.34 $^{(3)}$ ± 19.95	**77.33** $^{(1)}$ ± **16.69**	74.56 $^{(2)}$ ± 18.76	66.96 $^{(6)}$ ± 24.92	70.87 $^{(4)}$ ± 19.21	
Average Ranks											3.29 $^{(3)}$	3.02 $^{(2)}$	**2.71** $^{(1)}$	3.35 $^{(4)}$	4.35 $^{(6)}$	4.27 $^{(5)}$	

signed rank test and to report significant differences between the pairs [6][8]. Demsar has criticized the misuse of these approaches for multiple classifier comparisons because: (1) none of them reasons about comparing the means of more than two random variables, and (2) a certain portion of null hypothesis is always rejected due to a random chance by doing so [25]. In this paper, we use more specialized methods for comparing the average ranks of evolutionary classifiers (see Table 1) as suggested by Demsar [25] and Garcia [26].

Global Comparison of Evolutionary Classifiers. We use two most widely used non-parametric tests for comparison of multiple hypothesis among the classifiers: (1) **Friedman Test** [27], and (2) **Iman and Davenport Test** [28]. These tests utilize χ^2 and F distributions respectively to check if the distribution of observed and expected frequencies differ from each other.

Friedman and Iman and Davenport tests perform a global analysis to check whether the measured average ranks of all the classifiers are significantly different from the mean rank (3.5 in our case). The corresponding statistics χ^2_F and F_F are calculated as explained by Friedman and Iman and Davenport:

$$\chi^2_F = 19.94, F_F = 4.44$$

The critical values for corresponding tests χ^2_C and F_C obtained from the χ^2 and F distribution tables at $\alpha = 0.05$ with 5 and 150 degrees of freedom are:

$$\chi^2_C(5) = 11.07, F_C(5, 150) = 2.27$$

Since, the critical values are lower than the test statistics, the null hypothesis can be rejected and the post-hoc tests can be applied to detect significant differences between classifiers.

Comparison with the Control Classifier – GAssist. It can be seen from the results in Table 1 that GAssist 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. To compare the performance of GAssist with other evolutionary algorithms, we now establish the multiple hypothesis where every other evolutionary classifier is statistically compared with GAssist.

We use two post-hoc tests to determine the statistical significance of results: (1) **Bonferroni-Dunn Test** [29], and (2) **Holm Test** [30]. In general, these post-hoc tests vary in adjusting the threshold of significance level α in accordance with their multiple hypothesis. Bonferroni-Dunn Test controls the family-wise error rate in a single step by dividing α with the number of comparisons $(k - 1)$. Holm's Test is a step-down procedure in which the hypothesis is tested on the p-values arranged in ascending order. Starting from the lowest p-value, all the hypothesis are rejected for which $p_i <= \alpha/k-i$ while all the other remaining hypothesis are retained. Holm's Test is more powerful as it makes no assumptions about the hypothesis and in general, rejects more hypothesis than Bonferroni-Dunn's Test. The corresponding probability of the test statistic from the normal distribution table is obtained from the z-value by comparing i^{th} and j^{th} classifier. If the probability is less than the appropriate significance level, the null hypothesis is rejected. The results of comparison with control classifier GAssist are shown in Table 2.

Table 2. Test statistics for comparison with control classifier - GAssist ($\alpha = 0.05$, $k = 6$, $N = 31$ and $R_j = 2.71$). Null hypothesis is rejected for bold entries in p column.

i	Algorithms	Z-Value $(R_i - R_j)/\sqrt{k(k+1)/6N}$	p	Bonferroni-Dunn (B-D) $\alpha/(k-1)$	Holm $\alpha/(k-i)$	Critical Value B-D	Critical Value Holm
1	**SLAVE**	3.462	**5.36E-4**	0.01	0.01	**0.01**	**0.017**
2	**Ishibuchi**	3.292	**9.93E-4**	0.01	0.0125		
3	**cAntMiner**	1.358	0.174	0.01	0.017		
4	**XCS**	1.222	0.222	0.01	0.025		
5	**UCS**	0.645	0.519	0.01	0.05		

Table 3. Test statistics for pairwise comparisons ($\alpha = 0.05$, $k = 6$, $N = 31$). Null hypothesis is rejected for bold entries in p column.

i	Algorithms	Z-Value $(R_i - R_j)/\sqrt{k(k+1)/6N}$	p	Nemenyi $2 * \alpha/k(k-1)$	Holm $\alpha/(k-i)$	Critical Value Nemenyi	Critical Value Holm
1	**GAssist vs SLAVE**	3.462	**5.36E-4**	0.003	0.003	**0.003**	**0.004**
2	**GAssist vs Ishibuchi**	3.292	**9.93E-4**	0.003	0.004		
3	**UCS vs SLAVE**	2.817	**0.004**	0.003	0.004		
4	**UCS vs Ishibuchi**	2.647	0.008	0.003	0.004		
5	**XCS vs SLAVE**	2.240	0.025	0.003	0.004		
6	**cAnt-Miner vs SLAVE**	2.104	0.035	0.003	0.005		
7	**XCS vs Ishibuchi**	2.070	0.038	0.003	0.0055		
8	**cAnt-Miner vs Ishibuchi**	1.935	0.053	0.003	0.006		
9	**GAssist vs cAntMiner**	1.358	0.174	0.003	0.007		
10	**XCS vs GAssist**	1.222	0.222	0.003	0.008		
11	**UCS vs cAnt-Miner**	0.713	0.476	0.003	0.01		
12	**UCS vs Gassist**	0.645	0.519	0.003	0.0125		
13	**XCS vs UCS**	0.577	0.564	0.003	0.017		
14	**SLAVE vs Ishibuchi**	0.170	0.865	0.003	0.025		
15	**XCS vs cAnt-Miner**	0.136	0.892	0.003	0.05		

The last column gives the critical values of the used tests. If the p-value is less than or equal to this critical value, the null hypothesis is rejected for the corresponding test. It can be seen that the results of GAssist are statistically significant compared to SLAVE and Ishibuchi and hence, the null hypothesis can be rejected, while nothing much can be said about other algorithms with the given results.

Pairwise Comparisons. As GAssist cannot be termed as the best classifier against all the other classifiers in the last section, we now make the pairwise comparisons to analyze the statistical differences between all the classifiers. Along with the Holm's Test, we use the pairwise counterpart of Bonferroni-Dunn's Test called Nemenyi Test [31], for comparing all classifiers with each other. Nemenyi Test is more conservative than Benferroni-Dunn's Test as it steps-down the significance level by number of pairwise comparisons ($k(k-1)/2$ instead of $(k-1)$). The results in Table 3 show that the Nemenyi Test rejects the hypothesis of GAssist against SLAVE and Ishibuchi while the Holm's method also allows to reject the hypothesis for UCS vs SLAVE.

5.2 Effect of Evolutionary Algorithm

The use of statistical analysis provides deeper analysis to the obtained results than simply averaging the classification accuracies; a raw measure of ranking the performance

of algorithms. We now present the role of evolutionary learning paradigm in classifying biomedical datasets based on the obtained results:

Pittsburgh-Style – GAssist. The results of our experiments show that **GAssist** – a Pittsburgh-style learning classifier – performs better than other evolutionary rule-learning algorithms. 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 [14].

Nature Inspired – cAnt-Miner. cAnt-Miner 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 [32]. 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.

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.

Genetic Fuzzy – SLAVE and Ishibuchi. The results 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.

5.3 Effect of Nature of Dataset

A careful insight into the results of Table 1 enables the 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.

Role of Multiple Classes. It can be inferred 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 [33]. In comparison, GAssist has serious problems in dealing with multi-class problems – specially when the number of output classes are more than 5. On these datasets, the average accuracy of UCS is 83.49% compared with 75.52% of GAssist.

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. We argue that during training, classifiers create *small disjuncts* with rare cases [34]; as a result, their accuracy significantly degrades during testing.

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 reveals 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 increases with an increase in the information gain of its attributes.

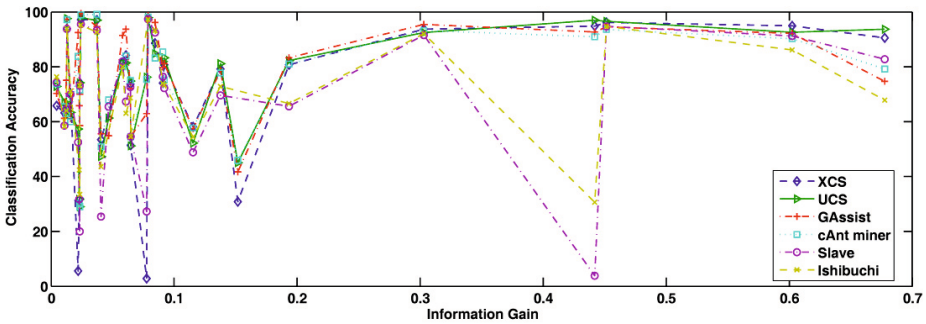

Fig. 1. Average Information Gain vs Classification Accuracy

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, random assignment, InGrimputation Model, listwise deletion etc. have been proposed for imputation to increase the accuracy of a classifier. 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. For nominal attributes it replaces missing value with their mode.

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.

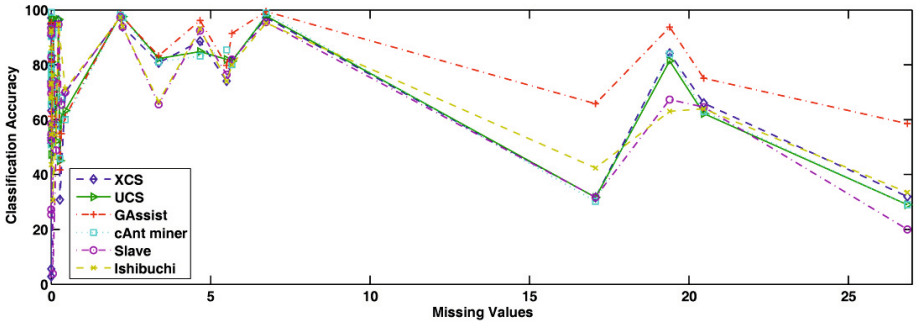

Fig. 2. Missing Values vs Classification Accuracy

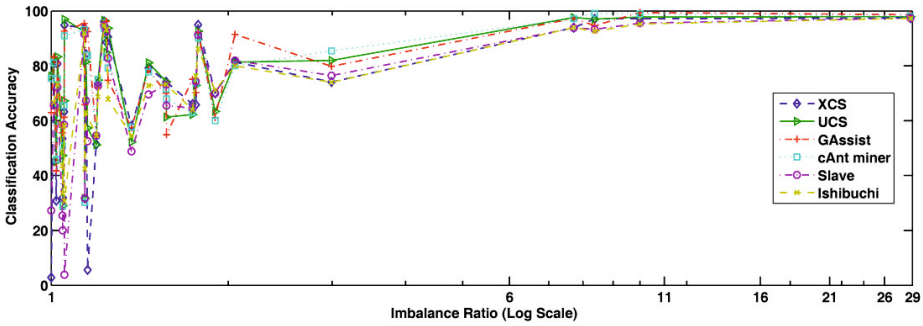

Fig. 3. Class Imbalance (Log Scale) vs Classification Accuracy

Role of Noise. The results in Table 1 show that classification potential of a dataset is inversely proportional to the 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.

5.4 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 while imbalance ratio does not dominate the resulting accuracy. We now conclude our findings in Figure 5 to have a better understanding of the combined effect of the complexity parameters.

It is obvious in Figure 5 that noise in a dataset effectively determines the classification accuracy. The high average information gain of a dataset yields better classification accuracy; while the percentage of missing values in a dataset has minor impact on the accuracy.

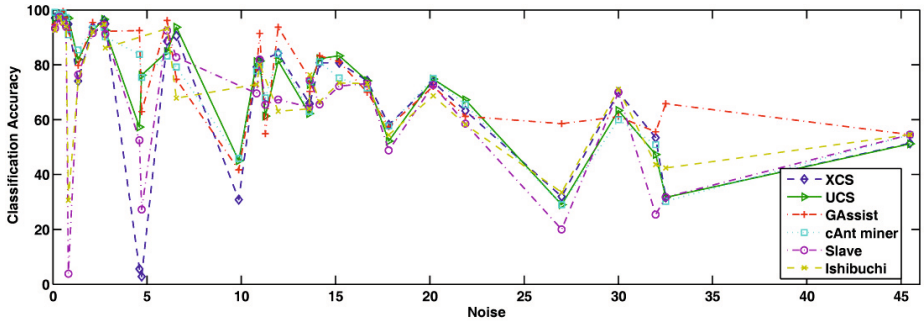

Fig. 4. Noise vs Classification Accuracy

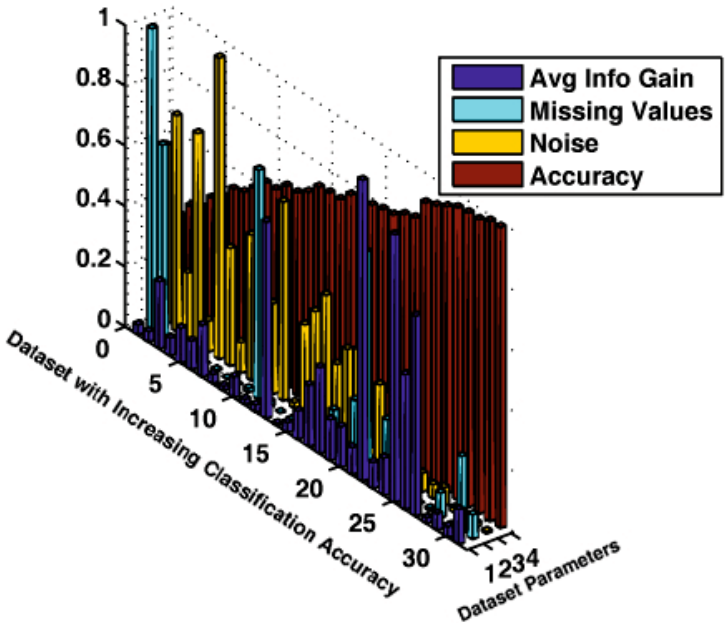

Fig. 5. Relationship between Classification Accuracy and Nature of Dataset: x-axis contains biomedical datasets in increasing order of their classification accuracies; y-axis contains normalized parameters of datasets, 1-Average Information Gain, 2-Missing Values, 3-Noise, 4-Classification Accuracy

Meta-Model for Classification Potential of a Dataset. In this section, we apply our meta-model framework [35] to get a measure of the classification potential of a dataset based on the complexity parameters. We create a meta-dataset comprising of three attributes for complexity parameters: average information gain, missing values, and noise.

We categorize the output class – classification potential – into three classes based on the classification accuracy: good (greater than 0.8), satisfactory (0.6-0.8) and bad (less than 0.6). The interesting patterns lying in this meta-dataset are extracted using two classifiers: (1) GAssist, it gives good classification results, and (2) Boosted J48 [22], to compare the results with well-known non-evolutionary algorithm.

Classification Rules of GAssist

```
0:Noise is [>0.667]|bad
1:MissingValues is [>0.905]|bad
2:MissingValues is [<0.125]|Noise is [>0.145]|satisfactory
3:Noise is [>0.287]|satisfactory
4:AvgInfoGain is [<0.29]|MissingValues is [>0.6]|bad
5:Default rule -> good
```

The classification rules generated by both classifiers prove our thesis that a noise level greater than 0.25 severely degrades the classification potential of a dataset. As expected, GAssist is able to generate more generic and comprehensible rules. For example, if noise level is above 0.667, the classification potential is bad irrespective of other parameters. The knowledge extracted by both algorithms provide same generalization. Hence, our proposed meta-model can be effectively used in determining the true classification potential of a biomedical dataset. We believe this can prove to be a very effective tool for analyzing the inherent complexities and needs for pre-processing the dataset.

Decision Tree of J48

```
Noise <= 0.26297
|   MissingValues <= 0.016387
|   |   AvgInfoGain <= 0.65192
|   |   |   AvgInfoGain <= 0.059957: good
|   |   |   AvgInfoGain > 0.059957: satisfactory
|   |   AvgInfoGain > 0.65192: good
|   MissingValues > 0.016387: good
Noise > 0.26297
|   MissingValues <= 0.002235: satisfactory
|   MissingValues > 0.002235: bad
```

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 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 using a meta-model framework. In the future, we would like to present the generated rules of different classifiers to the medical experts for their 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.

References

1. Pena-Reyes, C.A., Sipper, M.: Evolutionary computation in medicine: an overview. Journal of Artificial Intelligence in Medicine 19(1), 1–23 (2000)
2. Wong, M.L., Lam, W., Leung, K.S., Ngan, P.S., Cheng, J.C.V.: Discovering knowledge from medical databases using evolutionary algorithms. IEEE Engineering in Medicine and Biology 19(4), 45–55 (2000)
3. Holmes, J.H.: Learning classifier systems applied to knowledge discovery in clinical research databases. In: Lanzi, P.L., Stolzmann, W., Wilson, S.W. (eds.) IWLCS 2000. LNCS (LNAI), vol. 1996, pp. 243–261. Springer, Heidelberg (2001)
4. Bernado Mansilla, E.: Domain of competence of XCS classifier system in complexity measurement space. IEEE Transactions on Evolutionary Computation 9(1), 82–104 (2005)
5. Kharbat, F., Bull, L., Odeh, M.: Mining breast cancer data with XCS, Genetic and Evolutionary Computation Conference (GECCO), pp. 2066-2073, UK (2007)
6. Puig, A.O., Mansilla, E.B.: Evolutionary rule-based systems for imbalanced data sets. Soft Computing - A Fusion of Foundations, Methodologies and Applications 13(3), 213–225 (2009)
7. Bacardit, J., Butz, M.V.: Data mining in learning classifier systems: comparing XCS with GAssist. In: Kovacs, T., Llorà, X., Takadama, K., Lanzi, P.L., Stolzmann, W., Wilson, S.W. (eds.) IWLCS 2003. LNCS (LNAI), vol. 4399, pp. 282–290. Springer, Heidelberg (2007)
8. Bernadó, E., Llorà, X., Garrell, J.M.: XCS and GALE: a comparative study of two learning classifier systems with six other learning algorithms on classification tasks. In: Lanzi, P.L., Stolzmann, W., Wilson, S.W. (eds.) IWLCS 2001. LNCS (LNAI), vol. 2321, pp. 115–132. Springer, Heidelberg (2002)
9. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: An ant colony based system for data mining: applications to medical data. In: Int. Conf. on Knowledge Discovery and Data mining, Boston, pp. 55–62 (2000)
10. Galea, M., Shen, Q., Levine, J.: Evolutionary approaches to fuzzy modelling for classification. Knowledge Engineering Review 19(1), 27–59 (2004)
11. Tanwani, A.K., Afridi, J., Shafiq, M.Z., Farooq, M.: Guidelines to select machine learning scheme for classifcation of biomedical datasets. In: Pizzuti, C., Ritchie, M.D., Giacobini, M. (eds.) EvoBIO 2009. LNCS, vol. 5483, pp. 128–139. Springer, Heidelberg (2009)
12. Butz, M.V., Kovacs, T., Lanzi, P.L., Wilson, S.W.: Toward a theory of generalization and learning in XCS. IEEE Transactions on Evolutionary Computation 8(1), 28–46 (2004)

13. Bernado-Mansilla, E., Garrell-Guiu, J.M.: Accuracy-based learning classifier systems: models, analysis and applications to classification tasks. Evolutionary Computation 11(3), 209–238 (2006)
14. Bacardit, J., Garrell, J.M.: Bloat control and generalization pressure using the minimum description length principle for a Pittsburgh approach Learning Classifier System. In: Kovacs, T., Llorà, X., Takadama, K., Lanzi, P.L., Stolzmann, W., Wilson, S.W. (eds.) IWLCS 2003. LNCS (LNAI), vol. 4399, pp. 59–79. Springer, Heidelberg (2007)
15. Otero, F.E.B., Freitas, A.A., Johnson, C.J.: cAnt-Miner: an ant colony classification algorithm to cope with continuous attributes. In: Ant Colony Optimization and Swarm Intelligence, Belgium, pp. 48–59 (2008)
16. Gonzalez, A., Perez, R.: SLAVE: a genetic learning system based on an iterative approach. IEEE Transaction on Fuzzy Systems 7(2), 176–191 (1999)
17. Ishibuchi, H., Nakashima, T., Murata, T.: Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems. IEEE Transactions on Systems, Man, and Cybernetics 29(5), 601–618 (1999)
18. Fawcett, T.: ROC graphs: notes and practical considerations for researchers, TR HPL-2003-4, HP Labs, USA (2004)
19. UCI repository of machine learning databases, University of California-Irvine, Department of Information and Computer Science, `www.ics.uci.edu/~mlearn/MLRepository.html` (last accessed: June 25, 2010)
20. Zhu, X., Wu, X.: Class noise vs. attribute noise: a quantitative study of their impacts. Artificial Intelligence Review 22(3), 177–210 (2004)
21. Brodley, C.E., Friedl, M.A.: Identifying mislabeled training data. Journal of Artificial Intelligence Research 11, 131–167 (1999)
22. Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
23. Otero, F.E.B.: Ant Colony Optimization Framework, MYRA, `http://sourceforge.net/projects/myra/` (last accessed: June 27, 2010)
24. Alcala-Fdez, J., Sanchez, L., Garcia, S., del Jesus, M.J., Ventura, S., Garrell, J.M., Otero, J., Romero, C., Bacardit, J., Rivas, V.M., Fernandez, J.C., Herrera, F.: KEEL: a software tool to assess evolutionary algorithms for data mining problems. Soft Computing 13, 307–318 (2008)
25. Demsar, J.: Statistical comparisons of classifiers over multiple datasets. Journal of Machine Learning and Research 7, 1–30 (2006)
26. Garcia, S., Herrera, F.: An extension on "Statistical comparisons of classifiers over multiple datasets" for all pairwise comparisons. Journal of Machine Learning and Research 9, 2677–2694 (2008)
27. Friedman, M.: A comparison of alternative tests of significance for the problem of m rankings. Annals of Mathematical Statistics 11, 86–92 (1940)
28. Iman, R.L., Davenport, J.M.: Approximations of the critical region of the Friedman statistic. Communications in Statistics, 571–595 (1980)
29. Dunn, O.J.: Multiple comparisons among means. Journal of the American Statistical Association 56, 52–64 (1961)
30. Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70 (1979)
31. Nemenyi, P.B.: Distribution-free multiple comparisons, PhD Thesis, Princeton University (1963)

32. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an ant colony optimization algorithm. IEEE Transactions on Evolutionary Computation 6(4), 321–332 (2002)
33. Orriols-Puig, A., Bernadó-Mansilla, E.: Revisiting UCS: description, fitness sharing and comparison with XCS. In: Bacardit, J., Bernadó-Mansilla, E., Butz, M.V., Kovacs, T., Llorà, X., Takadama, K. (eds.) IWLCS 2006 and IWLCS 2007. LNCS (LNAI), vol. 4998, pp. 96–116. Springer, Heidelberg (2008)
34. Jo, T., Japkowicz, N.: Class imbalances versus small disjuncts. SIGKDD Explor. Newsl. 6(1), 40–49 (2004)
35. Tanwani, A.K., Farooq, M.: The role of biomedical dataset in classification. In: Combi, C., Shahar, Y., Abu-Hanna, A. (eds.) Artificial Intelligence in Medicine. LNCS (LNAI), vol. 5651, pp. 370–374. Springer, Heidelberg (2009)