Exploring Relationships Between Genotype and Oral Cancer Development Through XCS*

Alessandro Passaro
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
56127 Pisa, Italy
passaro@di.unipi.it

Flavio Baronti
Dipartimento di Informatica
Università di Pisa
Largo B. Pontecorvo, 3
56127 Pisa, Italy
baronti@di.unipi.it

Valentina Maggini
Dipartimento di Scienze
dell'Uomo & dell'Ambiente
Università di Pisa
56127 Pisa, Italy
v.maggini@geog.unipi.it

ABSTRACT

In medical research, being able to justify decisions is generally as important as taking the right ones. Interpretability is then one of the chief characteristics a learning algorithm must have, in order to be successfully applied to a medical data set. Other important features are seamless treatment of different data types, and ability to cope well with missing values. XCS and decision trees both appear to have this desirable characteristics; we compared them on a data set regarding Head and neck squamous cell carcinoma (HNSCC). This kind of oral cancer already been found to be associated with smoking and alcohol drinking habits. However the individual risk could be modified by genetic polymorphisms of enzymes involved in the metabolism of tobacco carcinogens and in the DNA repair mechanisms. To study this relationship, the data set comprised demographic and life-style (age, gender, smoke and alcohol), and genetic data (the individual genotype of 11 polymorphic genes), with the information on 124 HNSCC patients and 231 healthy controls. Results with both algorithms are presented and analyzed.

Categories and Subject Descriptors

J.3 [**Computer applications**]: Life and medical sciences; I.2.6 [**Artificial Intelligence**]: Learning—*Learning classifier systems*

General Terms

Algorithms

Keywords

Learning classifier systems, XCS, decision trees, genetic data, oral cancer

*This work has been carried out in the framework of the BIOPATTERN European Network of Excellence.

GECCO'05, June 25–29, 2005, Washington, DC, USA.

1. INTRODUCTION

Personalized medicine is by many considered one of the most fascinating and difficult challenges to current medical science. Very often, two persons which the statistics would classify as equal (same gender, age, lifestyle) have different reactions to drugs, or have different chances to develop certain diseases. The key to understand these variations has probably been found with the possibilty to decode DNA. DNA has clearly a chief part in regulating the chemical and biological responses of the human body; it can however be very difficult to establish a clear gene-effect relationship, as the biochemical responses are hard to trace through the human body. Moreover, genes often interact with each other, and some singularly "detrimental" or ineffective alleles can become beneficial when found together. These issues suggest the use of machine learning algorithms which can extract complex patterns from the observed data, and present them to the physicians in a human-readable form, amenable to further investigation.

In this work we consider the development of head and neck squamous cell carcinoma (HNSCC). This kind of cancer is mainly associated with smoking and alcohol drinking, but genetic polymorphism of enzymes involved in the metabolism of tobacco carcinogens and in the DNA repair mechanisms can influence the risk factor. The subjects were thus described with a combination of individual demographic and lifestyle data (gender, age, smoking and drinking habits) and genetic data (the individual genotype at 11 polymorphic genes potentially relevant to this disease) — along with a single value, which stated if they had cancer or not when the database was compiled.

We developed an XCS classifier system tailored to work with the different types of values found in this data set (boolean, integer, real and gene-class). This kind of classifier system was chosen for its ability to build very general accurate rules [7], whose interpretation is immediate. We then extended it with a ruleset reduction algorithm, in order to obtain a small set of mixed clinical and genetic rules that could suggest to physicians which genes increase or reduce oral cancer risk, and the direction to follow for more focused genetic research.

Preliminar results on this problem appeared in [2]. Here we completed and extended the tests on XCS, and compared our approach with decision trees, one of the most common methodologies in rule-based learning, with respect to descriptive power, predictive accuracy and clarity of results.

2. PROBLEM DESCRIPTION

The data set we analyzed was designed to explore the influence of genotype on the chance to develop head and neck squamous cell carcinoma (HNSCC). It is already well-known that this kind of cancer is associated with smoking and alcohol-drinking habits, it is more commong among males and its incidence increases with age. The individual risk however could be modified by genetic factors; therefore genotype information, regarding eleven genes involved with carcinogen-metabolizing (`CCND1`, `NQO1`, `EPHX1`, `CYP2A6`, `CYP2D6`, `CYP2E1`, `NAT1`, `NAT2`, `GSTP1`) and DNA repair systems (`OGG1`, `XPD`) was provided by molecular testing.

Nine of these genes have two allelic variants; let's call them a_1 and a_2. Since the DNA contains two copies of each gene, there exist three possible combinations: a_1a_1, a_2a_2 (the homozygotes) and a_1a_2 (the heterozygote — order does not matter). The homozygotes where represented with values 0 and 2, while the heterozygote with 1. Due to dominance, the heterozygote is equivalent to one of the homozygotes; however, for many of the considered genes this dominant effect is not known. So class 1 is either equivalent to class 0, or to class 2. The remaining two genes (`NAT1` and `NAT2`) have 4 allelic variants, which result in 9 combinations; they were sorted by their activity level, and put on an integer scale from 0 to 8.

The full data consists of 355 records, with 124 positive elements (HNSCC patients) and 231 negative (controls). Each record reports the person's gender, age, total smoke and alcohol consumption, gene values, and a boolean target value which specifies whether he had cancer when the database was compiled or not. The data was collected in different periods between 1997 and 2003; this has led to many missing data among the genotypic information of patients. Actually only 122 elements have complete genotypic description; the remaining 233 have missing values ranging from 1 to 9, with the average being 3.58. As an overall figure, of the $11 \times 355 = 3905$ genotype values, just 3070 are present: 21% of the genotype information is missing.

3. XCS

In [12] and then in [13], Wilson proposes XCS as an evolution of Holland's Learning Classifier Systems (LCS) [4], a machine learning technique which combines reinforcement learning, evolutionary computing and other heuristics to produce adaptive systems. Similarly to its ancestors, an XCS maintains and evolves a population of classifiers (rules) through a genetic algorithm. These rules are used to match environmental inputs and choose subsequent actions. Environment's reward to the actions is then used to modify the classifiers in a reinforcement learning process.

XCS introduces a measure of classifiers' fitness based on their accuracy, i.e. the reliability of their prediction of the expected payoff, and applies the GA only on the action set, the subset of matching classifiers which suggest the chosen action. This gives the system a strong tendency to develop accurate and general rules to cover problem space and allow the system's "knowledge" to be clearly seen. In the following we provide a brief description of XCS. For full details see [3].

3.1 System description

The core component of XCS is a set of classifiers, that is condition-action-prediction rules, where the **condition** specifies a pattern over the input states provided by the environment, the **action** is the action proposed (e.g. a classification), and the **prediction** is the payoff expected by the system in response to the action. Additionally each classifier has associated an estimate of the **error** made in payoff predictions, and a **fitness** value.

XCS implements a reinforcement learning process: at every step the system is presented an individual from the data set and it examines its set of classifiers to select those matching the input situation. These classifiers form the *match set*. Then for each possible action the system uses the fitness–weighted average prediction of the corresponding classifiers to estimate environmental reward. At this point, the XCS can choose the best action looking for the highest predicted reward. However, during learning, the action is usually selected alternating the previous criterion with random choice, useful to better explore the problem space. The actual reward returned by the environment is then used to update the classifiers in the *action set*, i.e. the subset of the *match set* corresponding to the selected action. A genetic algorithm is also executed on this set to discover new interesting classifiers.

To reduce the number of rules developed, XCS implements various techniques, such as the use of macroclassifiers, the subsumption and the deletion mechanisms. In fact the system uses a population of macroclassifiers, i.e. normal classifiers with a **numerosity** parameter, representing the number of their instances (microclassifiers). This helps in keeping track of the most useful rules and improves computational performance at no cost.

Subsumption is used to help generalization: when the GA creates a new classifier with a condition logically subsumed by his parent (i.e. matching a subset of the inputs matched by the parent's) it is not added to the population, but the parent's numerosity is incremented. A similar check is also occasionally done among all the classifiers in the current action set.

Finally the deletion mechanism keeps the number of microclassifiers under a fixed bound. The classifier to be removed is chosen with a roulette wheel selection biased towards low–fitness individuals and assuring approximately equal number of classifiers in each action set.

As already stated this process leads to the evolution of more and more general rules. For each classifier we can define a measure of generality following [16], ranging from 0 (most specific) to 1 (most general). A possible termination criterion is to stop evolution when the average generality value of the population gets stable.

4. ADAPTATION TO THE PROBLEM

In facing the problem of HNSCC development prediction from clinical and genetic data, we looked for a method which could provide a meaningful insight of its classification process, instead of focusing only on accuracy. In this regard, XCS showed many advantages over other well-established classification systems (for experimental comparison between XCS and other machine learning algorhitms, see for instance [1]). As seen in Wilson's works on Wisconsin Breast Cancer data [16] and Holmes' ones on epidemiologic surveillance data [5] (using EpiCS, a similar classifier system), the use of explicit rules to match the input data allows an easy visualization of the criteria the system employs in each classification and a comparison with physicians' previous knowledge.

As we have seen above, the data set is characterized by

the massive presence of missing data, especially in the genotype part. In these cases, essentially every classification technique is expected to experience a degradation of performance. However XCS allows at least their seamless management: an individual with missing data is matched only by those classifiers which have a wildcard on that value. The rationale underlying this choice is to avoid taking decisions based on data we do not have. This is different from Holmes' approach in [6], where missing values are matched by every classifier — thus producing a kind of average value for that data.

4.1 Data type integration

Another key aspect which lead us to choose XCS was the easiness of integration of different kind of data. In fact, the type of the information contained in the data set varies from binary (i.e. sex), to continuous-valued (i.e. age, indicators of smoking and alcohol-drinking habits), and to a special class data for the genotype. Whilst the original formulation of XCS is targeted to binary input, the shift to other data types, such as real or integer ones, has already been proved to be very easy (see respectively [16, 14]).

For the integer and real data types, our implementation is based on those proposed in the cited literature. For the nine genotypic values with two allelic variants we needed instead an ad-hoc treatment. As discussed in Sec. 2, in these values class 1 is either equivalent to class 0 or to class 2 — so it is meaningless to have a classifier isolating that single class. The possible patterns for classifiers are then the following: `00` matching class 0, `22` matching class 2, `01` (matching 0 and 1), `12` (matching 1 and 2) and `##` (matching all values, nulls included as for the other data types). This is actually equivalent to modeling all the non-empty subsets of the $\{0, 1, 2\}$ set, without the options `11` and `02`.

4.2 Ruleset reduction

During learning XCS tends to evolve an accurate and complete mapping of condition-action-prediction rules matching the data. Consequently, in particular on a very sparse data set as in our study, the final number of rules is quite high. Similar problems, which break the knowledge visibility property, were experienced in other studies on "real" data sets [16, 15]. These works suggest to let the system evolve many steps after reaching the maximum performance, and then to extract a small subset of rules which reach the same performance level. This is the function of the *Compact Ruleset Algorithm* (CRA), first proposed by Wilson [15], which we implemented with a small modification: since we do not reach 100% accuracy, CRA stops adding rules when the accuracy obtained equals the value reached with all rules.

5. RESULTS

We had two aims in testing the system: evaluating its ability to correctly classify unseen data after training and checking if it could find interesting rules. We applied a ten-fold cross-validation and repeated the experiment ten times (each time with a different folding), in order to obtain results independent from the particular folding. Each experiment was allowed to run for $500,000$ steps, as a few tests showed that the generality value reached stability by this point. The used parameters were chosen following [3]. Experiments were run with several population sizes, ranging from 6400 to 200 microclassifiers. Final results are summarized in Table 1.

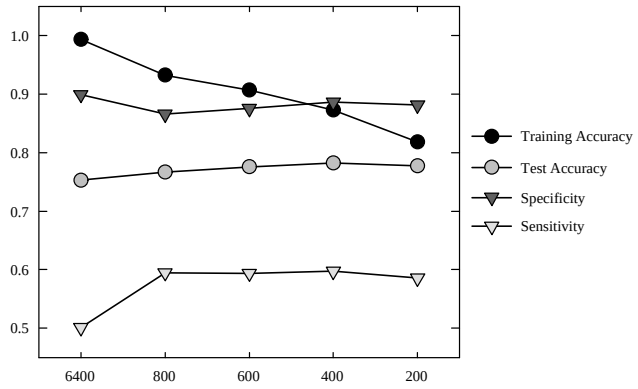

Figure 1: XCS performances with varying population sizes.

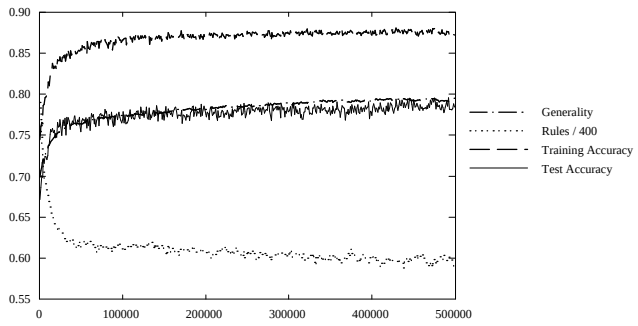

Figure 2: Plot of average evolution in the experiments with a population of 400 microclassifiers.

In the experiment with 6400 classifiers, the accuracy on the training set reached almost optimal value, while it decreased in the experiments with lower population sizes. However the accuracy on the test set was at least comparable, and even showed a slightly increasing trend with smaller populations (see Fig. 1). This suggests that the high accuracy of the 6400 test is due to overfitting, and lower population sizes are preferable. In particular, XCS performances appear stable for populations in the range from 200 to 800. In fact, ANOVA on the before-CRA test performance of the 200–800 experiments failed to reject the null hypothesis ($p = 0.515$), so it cannot be concluded that they are actually statistically different. On the other side, the unpaired t-test between the 6400 experiment and all the others showed a significant difference ($p = 0.0067$). The evolution of the system for a population of size 400 is plotted in Fig. 2.

The CRA successfully extracted a small subset of the original rules which maintained the maximum performance on the training set. It is interesting however to compare the test set accuracy before and after CRA. While for smaller populations the performance is stable, for bigger population there is a significant worsening (see table 1; the reported p-values are for a 2-tailed paired t-test). It could be profitable to design CRA as a pruning algorithm — that is, allowing to lose some accuracy on the training set, in order to perform better on the test set.

Nevertheless the small sets of rules extracted made it fea-

Table 1: Summary of the ten 10-fold cross validation experiments. Specificity and sensitivity are relative to the test set. Last column shows p values for the paired 2-tailed t-test on the effect of CRA over test accuracy.

Max rules	CRA	Final rules	Accuracy		Specificity	Sensitivity	CRA effect (p values)
			Training	Test			
6400	Before	1659 ± 115	$99 \pm 1\%$	$75 \pm 2\%$	$90 \pm 2\%$	$50 \pm 5\%$	0.0034
	After	47 ± 14	$99 \pm 1\%$	$72 \pm 3\%$	$77 \pm 2\%$	$65 \pm 3\%$	
800	Before	413 ± 25	$93 \pm 1\%$	$77 \pm 1\%$	$87 \pm 2\%$	$59 \pm 3\%$	0.0009
	After	49 ± 21	$93 \pm 1\%$	$74 \pm 2\%$	$82 \pm 1\%$	$61 \pm 5\%$	
600	Before	333 ± 22	$91 \pm 2\%$	$78 \pm 2\%$	$88 \pm 2\%$	$59 \pm 3\%$	0.0014
	After	34 ± 11	$91 \pm 2\%$	$75 \pm 2\%$	$83 \pm 3\%$	$62 \pm 3\%$	
400	Before	236 ± 19	$87 \pm 2\%$	$78 \pm 2\%$	$89 \pm 3\%$	$60 \pm 2\%$	0.2048
	After	16 ± 9	$87 \pm 2\%$	$79 \pm 1\%$	$89 \pm 2\%$	$62 \pm 3\%$	
200	Before	119 ± 17	$82 \pm 4\%$	$78 \pm 2\%$	$88 \pm 5\%$	$59 \pm 5\%$	0.9988
	After	9 ± 5	$82 \pm 4\%$	$78 \pm 2\%$	$90 \pm 2\%$	$56 \pm 5\%$	
See5		Not applicable	$79 \pm 2\%$	$69 \pm 2\%$	$76 \pm 2\%$	$57 \pm 4\%$	Not applicable

Table 2: Examples of rules extracted by the system, with their correct/matched ratio, fitness and error.

Classifier	Ratio	Fitness	Error
IF $age \leq 41$ THEN $cancer = false$	37/37	0.200	0
IF $smoke \geq 20$ AND $EPHX1 \in \{1,2\}$ AND $GSTP1 \in \{0,1\}$ THEN $cancer = true$	50/56	0.347	160

sible to manually look for possibly interesting rules. As an example we provide in Table 2 two of such rules in human readable form. The first rule is common knowledge rediscovered by the system. Instead the second one has been judged interesting by physicians: in fact previous studies already reported an increased lung cancer risk associated to GSTP1 in combination with EPHX1 polymorphisms [11], so it will be interesting to investigate on the role of these genes in relation to HNSCC risk.

6. COMPARISON WITH DECISION TREES

In order to evaluate our approach, we compared it to a classical machine learning tool for classification and prediction: decision trees [8]. Decision trees were since they are a well-known machine learning method which complies with our requirements about interpretability, treatment of different data types, and robustness to missing data.

A decision tree is a classifier in the form of a tree structure, where each leaf node indicates the value of a target class and each internal node specifies a test to be carried out on a single attribute, with one branch and sub-tree for each possible outcome of the test. The classification of an instance is performed by starting at the root of the tree and moving through it until a leaf node is reached, which provides the classification of the instance.

Among the variety of algorithms for decision trees induction from data, probably the most known and used are ID3 and its enhanced version C4.5 [9]. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. The algorithm uses a greedy search, that is, it picks the best attribute and never looks back to reconsider earlier choices. The central focus of the decision tree growing algorithm is selecting which attribute to test at each node in the tree. The goal is to select the attribute that is most useful for classifying examples. A good quantitative measure of the worth of an attribute is a statistical property called information gain that measures how well a given attribute separates the training examples according to their target classification. This measure is used to select among the candidate attributes at each step while growing the tree.

Table 3: Decision tree obtained from the entire dataset, along with the correct/matched ratio for each branch.

```
packyears <= 0.04875: false (135.9/158.9)
packyears > 0.04875:
:...age > 0.78: false (12/12)
    age <= 0.78:
    :...gstp1 <= 0: true (63.3/102.1)
        gstp1 > 0:
        :...nat2 <= 3: false (30/43.6)
            nat2 > 3: true (24.2/38.5)
```

6.1 Decision trees results

Decision tree induction on our dataset was performed using the See5 software [10]. After some testing, we found out that the default parameters (pruning CF = 25%, minimum case per branch = 4) worked well for this dataset; boosting was not employed, since it did not appear to improve performance. We applied a ten-fold cross-validation and repeated it ten times, as in the experiments with XCS (that is, with 10 different foldings). In this case, results' variability is due only to the random folding in the cross-validation procedure, since the decision tree induction algorithm is deterministic.

The results are reported in Table 1, where the accuracy, sensitivity, and specificity obtained with See5 are compared with those obtained with XCS. The accuracy value is close to the one obtainable predicting always the most frequent class (65%), but sensitivity and specificity ensure decision trees are doing something more clever. Finally, the decision tree obtained with the execution of See5 on the entire dataset is reported in Table 3.

The figures in Table 1 show a clear performance advantage of XCS over See5, both on training and test sets. This gives a quite good level of confidence on extracted rules, suggesting XCS managed to convey in them useful knowledge.

However, sets of rules obtained with XCS are slightly less readable and interpretable than decision trees. Moreover, XCS results show a quite high variability. In fact, classification accuracy does not change much between runs, but the actual rulesets appear quite different. Since interpretability is our main concern, this constitutes a remarkable problem: there is no evident way to get a single "final" set of rules. In this respect an appealing See5 characteristic is that it extracts a single decision tree from a given dataset.

7. CONCLUSIONS AND FUTURE WORK

In this work we applied an XCS system to the analysis of a mixed clinical and genetic data set regarding the risk of developing HNSCC. The long-term goal is to identify the genes actually involved in the susceptibility to oral cancer, and highlight possible interactions between them. XCS has confirmed its flexibility in adapting to different data types and seamless handling of missing values. The rules extracted from the first experiments suggest that the system can produce interesting results. Moreover, they are easily converted in human-readable form, and can be immediately evaluated by physicians.

Classification accuracy was higher than that obtained using a standard algorithm, such as decision tree induction. The reached performance value on test cannot be considered "high" in an absolute sense; however, given the particular nature of the input data, it is not completely clear how better this value could become. For instance, this data set is noisy not only on some input variables (smoke and alcohol habits), but also on the target: more than other diseases, cancer cannot be deterministically predicted. Regarding the first issue, it would be useful to perform some tests on the effects of noise in XCS. Concerning the target variable, a possible direction is prediction of a risk factor instead of a raw class, as in [5].

Another interesting aspect to investigate is the ruleset reduction algorithm: CRA is mainly focused on maintaining the training performance achieved, while a more pruning-like strategy could be beneficial for generalization. CRA should moreover include as a chief goal to regularize the algorithm output, in order to produce more stable results. Results stability could also be achieved as a post-processing step; for instance, it could be possible to find similar rules recurring among different executions. This would require a measure of similarity between rules, and a clustering algorithm able to group them together.

8. ACKNOWLEDGEMENTS

We would like to thank the following people for providing the data set and supporting us during the analysis: A. Abbondandolo, R. Barale, S. Bonatti, F. Canzian, G. Casartelli, G. Margarino, P. Mereu.

9. ADDITIONAL AUTHORS

Additional authors: Alessio Micheli (Dipartimento di Informatica, Università di Pisa, email: `micheli@di.unipi.it`), Anna Maria Rossi (Dipartimento di Scienze dell'Uomo & dell'Ambiente, Università di Pisa, email: `a.m.rossi@geog.unipi.it`) and Antonina Starita (Dipartimento di Informatica, Università di Pisa, email: `starita@di.unipi.it`).

10. REFERENCES

[1] A. Bagnall and G. Cawley. Learning classifier systems for data mining: A comparison of XCS with other classifiers for the Forest Cover dataset. In *Proceedings of the IEEE/INNS International Joint Conference on Artificial Neural Networks (IJCNN-2003)*, volume 3, pages 1802–1807. IEEE Press, 2003.

[2] F. Baronti, V. Maggini, A. Micheli, A. Passaro, A. M. Rossi, and A. Starita. A preliminary investigation on connecting genotype to oral cancer development through XCS. In *Proceedings of WIRN 2004*. [in print], 2004.

[3] M. V. Butz and S. W. Wilson. An algorithmic description of XCS. In P. L. Lanzi and et al., editors, *IWLCS 2000*, volume 1996 of *LNAI*, pages 253–272. Springer-Verlag, 2001.

[4] J. H. Holland. Adaptation. In R. Rosen and F. M. Snell, editors, *Progress in theoretical biology, 4*. New York: Plenum, 1976.

[5] J. H. Holmes. Learning classifier systems applied to knowledge discovery in clinical research databases. In Lanzi et al., editor, *Learning Classifier Systems. From Foundations to Applications*, volume 1813 of *LNAI*, pages 243–261. Springer-Verlag, 2000.

[6] J. H. Holmes and W. B. Bilker. The effect of missing data on learning classifier system learning rate and classification performance. In Lanzi et al., editor, *IWLCS 2002*, volume 2661 of *LNAI*, pages 46–60. Springer-Verlag, 2002.

[7] T. Kovacs. XCS classifier system reliably evolves accurate, complete, and minimal representations for boolean functions. Technical Report CSRP-97-19, University of Birmingham, June 1997.

[8] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81 – 106, 1986.

[9] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

[10] Rulequest Research. See5/C5.0. `http://www.rulequest.com/`.

[11] J. To-Figueras, M. Gene, J. Gomez-Catalan, E. Pique, N. Borrego, and J. Corbella. Lung cancer susceptibility in relation to combined polymorphisms of microsomal epoxide hydrolase and glutathione s-transferase p1. *Cancer Letters*, 173(2):155–162, 2001.

[12] S. W. Wilson. Classifier fitness based on accuracy. *Evolutionary Computation*, 3(2), 1995.

[13] S. W. Wilson. Generalization in the XCS classifier system. In J. R. Koza and et al., editors, *Genetic Programming 1998: Proceedings of the Third Annual Conference*, pages 665–674, University of Wisconsin, USA, 22-25 1998. Morgan Kaufmann.

[14] S. W. Wilson. Get real! XCS with continuous-valued inputs. In Lanzi et al., editor, *Learning Classifier Systems. From Foundations to Applications*, volume 1813 of *LNAI*, pages 209–219. Springer-Verlag, 2000.

[15] S. W. Wilson. Compact rulesets from XCSI. In P. L. Lanzi and et al., editors, *IWLCS 2001*, volume 2321, pages 197–210. Springer-Verlag, 2001.

[16] S. W. Wilson. Mining oblique data with XCS. In P. L. Lanzi and et al., editors, *IWLCS 2000*, volume 1996 of *LNAI*, pages 158–174. Springer-Verlag, 2001.