The Stability of Long Action Chains in XCS

BARRY, A. M.(✉)

Faculty of Computer Studies and Mathematics,

University of the West of England,

Coldharbour Lane, Frenchay,

Bristol, UK.

BS16 1QY

Alwyn.Barry@uwe.ac.uk
(++44) [0]117 344 3135

XCS [1][2] represents a new form of Learning Classifier System [3] that uses accuracy as a means of guiding fitness for selection within a Genetic Algorithm. The combination of accuracy-based selection and a dynamic niche-based deletion mechanism achieve a long sought-after goal – the reliable production, maintenance, and proliferation of the sub-population of optimally general accurate classifiers that map the problem domain [4]. Wilson [2] and Lanzi [5][6] have demonstrated the applicability of XCS to the identification of the optimal action-chain leading to the optimum trade-off between reward distance and magnitude. However, Lanzi [6] also demonstrated that XCS has difficulty in finding an optimal solution to the long action-chain environment Woods-14 [7]. Whilst these findings have shed some light on the ability of XCS to form long action-chains, they have not provided a systematic and, above all, controlled investigation of the limits of XCS learning within multiple-step environments. In this investigation a set of confounding variables in such problems are identified. These are controlled using carefully constructed FSW environments [8][9] of increasing length. Whilst investigations demonstrate that XCS is able to establish the optimal sub-population *[O]* [4] when generalisation is not used, it is shown that the introduction of generalisation introduces low bounds on the length of action-chains that can be identified and chosen between to find the optimal pathway. Where these bounds are reached a form of over-generalisation caused by the formation of *dominant classifiers* can occur. This form is further investigated and the *Domination Hypothesis* introduced to explain its formation and preservation.

Keywords: XCS, Action-Chain, Generalisation

1. Introduction

XCS [1][2] represents a new form of Learning Classifier System [3] that uses accuracy as a means of guiding fitness for selection within a Genetic Algorithm. The combination of accuracy-based selection and a dynamic niche-based deletion mechanism achieve a long sought-after goal – the reliable production, maintenance, and proliferation of the sub-population of optimally general accurate classifiers that map the problem domain. Much has now been written on XCS and its operation, and the interested reader is referred to [1][2][4][14] and [15] for further details. This paper assumes a familiarity with the operation of XCS.

Although a number of significant results have been achieved in regard to the ability of XCS within single-step environments, research into the performance of XCS within multiple-step environments has been more limited. Wilson [1][2] provided a proof-of-concept demonstration of the operation of XCS within the Woods2 environment. Lanzi [5] identified that within certain Woods-like environments XCS was unable to identify optimum generalisations. This was attributed to two major factors: an inequality in exploration of all states in the environment allowing over-general classifiers to appear accurate, and an input encoding that meant that certain generalisations were not explored as often as others. Lanzi [6] sought to apply these lessons to more complex Woods-based environments and discovered that XCS was additionally unable to establish a solution to the long chain Woods-14 problem [7]. This was due in part to the number of possible alternatives to explore in each state that prevented XCS from attributing equal exploration time to later states within the chain. Further work [10][11][12][13] has examined the use of memory bits to overcome the problem of perceptual aliasing within multiple-step environments.

Whilst these investigations have generated useful, and at times dramatic, results and shed considerable light upon the operation of XCS within multiple-step environments, they have not investigated the limits of XCS learning within these environments in a systematic manner. To establish some limits on the capabilities of XCS the problems of action chain length, exploration complexity and input encoding complexity must be disentangled. This paper presents investigations that

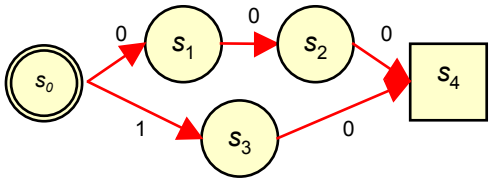
seek to establish some limits of XCS within the area of action chain length. Test environments are constructed to ensure that the three problems identified by Lanzi are controlled in such a way that a single issue can be focused upon. It could be argued that this requires potentially artificial environments that may never be found within real test situations. It is recognised that the problem of exploration complexity is intrinsically tied to this area of investigation and is worthy of separate study. The problem is here referred to as "*action chain learning*" rather than "*rule chain learning*" because XCS chains action sets rather than individual classifiers, although in some respects the ideas of "*rule chaining*" and "*action chaining*" can be synonymous.

2. A Highly Controllable Environment

In order to investigate this area empirically a suitable test environment is required. The Woods environment is a useful general learning test, but is not easily scaled with fine control in either length or complexity. Therefore, the Woods environments are set aside in favour of a Finite State World (FSW) environment similar to that proposed by Grefenstette [8] and used extensively by Riolo [16, 17, 18]. A FSW is an environment consisting of *nodes* and directed *edges* joining the nodes. Each node represents a distinct environmental *state* and is labelled with a unique state identifier. Each node also maintains a message that the environment passes to the XCS when at that state. Each directed edge represents a possible *transition* path from one node to another and is labelled with the action(s) that will cause movement across the edge in the stated direction to a destination node. An edge may lead back to the same node. Each node has exactly one label and message, and each message is unique within a Markovian FSW and normally equivalent to the node's label. At least one node must be identified as a *start* state, signifying that the XCS will be operating in that state when each new learning trial begins. If more than one start state is provided the actual state from which a trial is started is selected arbitrarily from the available start states. Additionally, one or more nodes are identified as *terminal* states. Transition to any one of these states represents the end of a learning trial and each will have an associated reward value representing an environmental reward that is passed to XCS upon transition into such a state. Terminal states do not have any transitions emanating

from them - upon arrival the trial is ended, the next iteration will represent a new trial and the environment will reset to a [selected] start state.

FSW are more precise than Woods environments - each state has a distinct label so that aliasing problems do not occur, and configurations that would not be possible within a Woods environment can be created, as demonstrated below:



Whereas within the Woods environments it is not possible to create long action chains without aliasing problems (unless more sensory stimuli were introduced, thereby changing the base problem), within FSW it is possible to extend the chain of states as far as desired. It is therefore is simple to extend environments whilst controlling the complexity generated by the number of states, their stimuli, and their interconnection. FSW are thus ideal for the controlled tests required within this work.

3. Confounding Variables

In investigating some limits on the length of action chains that XCS can learn, it has to be recognised that there are many inter-related factors that can influence learning.

a) Exploration Complexity

Consider a FSW environment with a chain of n states where each non-terminal state s_i has two edges, one to the next state s_{i+1} and one back to itself. In such an environment the probability of reaching state s_{i+1} from s_i is 0.5 in explore mode. For successive states the probability remains the same. Thus, the probability of moving from any non-terminal state s_i to another state on this chain s_{i+m} will be 0.5^m. Therefore, as the length of a state chain increases the ability of explore mode within XCS to explore the chain will dramatically decrease when there remains only one start state on the chain. Clearly the more possible pathways there are to choose, the more problematic exploration becomes. Within XCS this problem is overcome for the exploration of the optimal action chain once that action chain

has been explored sufficiently to dominate the system prediction of the chain because exploitation can then utilise this route and continue to learn the optimum prediction. However, the discovery of this optional route remains a major hurdle. An aspect of this problem was identified by Lanzi [6] when studying the Woods14 problem [7], and Lanzi noted that it could be solved by employing a biased exploration strategy. The very existence of this problem indicates one of the areas that must be controlled in order to carry out this investigation.

b) Environment 'Shape'

Although intrinsically related to the issue of exploration complexity, the issue of environment 'shape' is worthy of separate mention. Certain environments will, by nature of their degree of connectivity, include areas that are more difficult to reach. Lanzi [5] noted that the Maze6 environment was difficult for XCS to find optimum solutions within because exploration rates for some areas were higher due to the shape of the environment. Furthermore, if an environment provides opportunities for looping back to earlier states (or even to the same state) the frequency in which a reward state is encountered will diminish, leading to longer periods between external reinforcement. There is potentially much more work to be done in this area to understand it fully that is beyond the scope of this research effort.

c) Input encoding

The fundamental power of XCS is its ability to generalise whilst learning. The application of the Generalisation Hypothesis [2] and the Optimality Hypothesis [4] to multiple-step XCS learning has not been investigated, but confirmation of this ability is surely a key objective for XCS research.

The prediction of classifiers within the XCS population in a multiple-step environment will be updated either as a result of payoff from the environment or as a result of payoff generated by moving the Animat controlled by the XCS into a position where a different classifier can operate and receive an environmental payoff. Where payoff is not received from the environment, the stable prediction of the classifier will be dependent upon the stable prediction of the classifiers that operate in the following step. The payoff received by the action set in the previous

iteration is calculated from equation 1, where *S*i is the set of System Predictions from the action sets formed during matching in the current iteration. The update for the prediction within a classifier is calculated (ignoring the initial section of the MAM technique for the present) using equation 2:

$$r_{i-1} \leftarrow \gamma.\max(S_i) \qquad \text{[eq. 1]}$$

$$p_{i-1} \leftarrow p_{i-2} + \beta\,(r_{i-1} - p_{i-2}) \qquad \text{[eq. 2]}$$

The discount factor γ will reduce the payoff to the preceding classifiers so that when moving further back in time from the classifier that received the environmental payoff the stable prediction of the classifiers will decrease by a power of γ each time. In an ideal situation the stable prediction of a classifier *t* steps away from the environmental reward *R* should therefore be:

$$\gamma^{t} R \qquad \text{[eq. 3]}$$

Unfortunately it is possible that as the payoff diminishes down the action sets that are invoked as progress is made through a chain of states in even a simple single-chain FSW, the generalisation mechanism of XCS will generalise over the classifiers covering the early states. The ability of XCS to generalise may be dependent upon the amenability of the input messages attributed to each state. Lanzi [6] has already noted that an over-redundant encoding can affect the ability of XCS to operate effectively in the presence of generalisation pressure due to the inadequate or uneven exploration of some of the message bits. It could be the case that potential generalisations over similar actions in separate states may be aided or hindered by the coding of the messages from these states.

d) Parameterisation

The rate at which XCS learns its strength values is fundamentally controlled by the *learning rate* parameter β. This adjusts the rate at which the current error, prediction, and fitness estimates are adjusted and reflects the responsiveness of these values to changing environmental input. In multiple-step environments the rate of learning is also affected by the *discount factor* parameter γ. This parameter reflects the rate of decline in payoff down the sequence of action sets leading to a rewarding state. A high value of γ will result in a lower differentiation in payoff between the states. Whilst this may allow longer action chains to survive, the

closer payoff values may prevent the generalisation capabilities of XCS rapidly finding the accurate yet general classifiers that can be proliferated through the action of the G.A. and subsumption. Lower values of γ may help this generalisation process, but possibly at the cost of reduced length chains.

In addition to these parameters affecting the learning rate, other parameters such as the choice of explore-exploit strategy, the frequency of the genetic algorithm, the amount of effector covering, the time until a classifier is regarded as experienced will all affect the learning of action chains to one degree or another. None of these parameterisation issues has been investigated, and this remains a fertile area for future work.

4. Experimental Hypotheses

Within XCS the discount factor γ reduces the payoff to the classifiers within preceding action sets. This payoff reduction has two purposes. Firstly it reflects the increasing degree of uncertainty regarding the role of the preceding action sets in leading to the final reward. Perhaps more fundamentally, it allows XCS to take account within selection between competing pathways of the distance to the reward as well as the ultimate reward magnitude. The side-effect of the discount within XCS, however, is that as the payoff decreases by a power of the discount factor on each step (see equation 3) the payoffs received by action sets become increasingly similar. The generalisation hypothesis claims that XCS will be able to identify classifiers of the optimum generality to map the state $\times$ action $\times$ payoff landscape of a problem. However, as the action chain length increases and the payoff to early action sets decreases the generalisation capability of XCS may simply generalise over the very similar predictions within these early action sets. The effect of this is likely to be that XCS can no longer select the optimum route to a reward. This reasoning leads to hypothesis 1:

Hypothesis 1
There will be a point in the lengthening of a single action chain to a stable fixed point reward when the payoff to the action sets covering the initial states will be sufficiently similar to cause incorrect

generalisations, thereby preventing XCS identifying a correct state × action × payoff mapping over those states.

In addition to the potential problems with generalisation, there is no information on the performance of XCS without the burden of generalisation in very long action chains. It may be hypothesised that XCS will be able to identify all the classifiers required to map the state × action × payoff landscape in the simple though long corridor FSW. However, there is uncertainty about whether XCS will be able to establish the very small payoff predictions in the early states accurately enough to continue to select the optimal route given the relatively noisy update generated by the feedback from competing action sets. Given that the distinction between the stable prediction for action sets representing the optimum and the nearest sub-optimal route will be very small, it can be hypothesised that XCS will be unable to distinguish between the optimal and sub-optimal routes once prediction values become small.

Hypothesis 4.2
As the payoff reduces to fractions of unity XCS without generalisation will become unable to reliably select the optimal path within a simple two-choice corridor FSW environment.

5. Choice of a Test Environment

The test environment chosen is a FSW representation of a so-called "corridor environment". The environment is pictured in figure 1, and the message produced from each state of this environment is the binary coding of the state number.

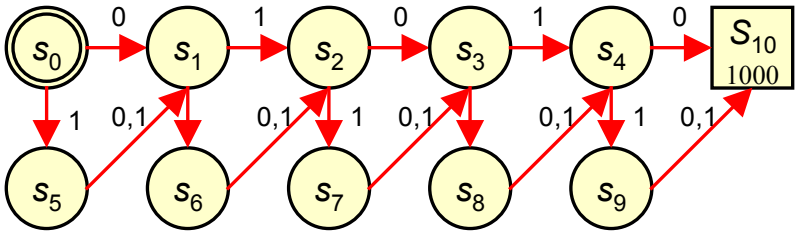


Figure 1 - A Corridor Finite State World

This environment has the following features:

- It can be trivially extended by small or large increments as longer test action chains are required.
- It includes a choice of route at each state so that the ability of XCS to decide the optimal route as the action chain increases can be determined.
- The sub-optimal route does not prevent progress towards the reward state.
- The optimal route is always re-joined to limit the penalty of a sub-optimal choice.
- The stable payoff received for a sub-optimal choice will always be equivalent to the γ discount of the payoff received for the optimal choice.
- The alternation of actions prevents generalisation from prematurely producing very general classifiers that cover much of the optimal path to reward.
- The small number of separate actions limits exploration complexity.

The problem of exploration complexity is removed, as far as possible, by the limit to an optimal and sub-optimal choice in each state combined with the immediate re-joining of the optimal route. The problem of exploration shape is controlled by the use of a simple state chain. The problem of input encoding is controlled by making the input from each state the binary code of the integer number of the state for all experimental work apart from those experiments explicitly designed to investigate changes in input encoding. The problem of parameterisation will be solved by utilising the set of parameter values given in Table 1, derived from those used by Wilson [1][2]. These parameter settings were chosen to allow comparison with previous XCS work, although the mutation probability is higher than Wilson used in the Woods-2 experiments but the same as that used within his multiplexer work. The population size was selected to provide sufficient space for each fully specific classifier to achieve a maximum numerosity of 20. It has been noticed in the course of experimental work that XCS can increase the numerosity of non-optimal classifiers to 6 or [more rarely] 8 copies. Therefore, to ensure dominance the 'rule of thumb' of providing a minimum of 12 classifiers per member of [O] was adopted throughout this work. Providing 20 spaces per required classifier allows for the additional classifier space required for the normal XCS exploration and the fact that generalisation may allow fewer classifiers to be used to represent the state space. It is acknowledged that tuning of parameters may succeed in producing different and possibly better results than those presented in

the following sections. An exploration of the parameter space is a valid area of investigation in itself, but beyond the scope of this investigation.

N (population size)	$20 \times (\text{states} - 1) \times 2$
$\mathbf{P}_i$ (initial population size)	0
γ (discount rate)	0.71
β (learning rate)	0.2
θ (GA Experience)	25
ε_0 (minimum error)	0.01
α (fall-off rate)	0.1
$\mathbf{X}$ (crossover probability)	0.8
μ (mutation probability)	0.04
p_r (covering multiplier)	0.5
$\mathbf{P(\#)}$ (generality proportion)	0.33
p_i (initial prediction)	10.0
ε_i (initial error)	0.0
f_r (fitness reduction)	Not Used
m (accuracy multiplier)	0.1
s (Subsumption threshold)	20
f_i (initial fitness)	0.01
Exploration trials per run	5000
Maximum iterations per trial	chain length $\times$ 10

Table 1 - Parameter settings for action chain length experiments

6. Experimental Method

The investigation of these hypotheses has been divided into three stages. All stages will initially base their tests on the FSW within figure 1 expanded to represent optimal action chain lengths of 5 (as pictured in figure 1), 10, 15, 20, 25, and 30. The XCS implementation used within these experiments is XCSC [15], a publicly available implementation that has been shown to be equivalent to the performance of the XCS implementations used in [2] and [4]. It differs from the more recent description of XCS [14] in using Wilson's type 1 deletion technique [1] rather than Kovac's type 3 technique [4], which is unproven in multiple-step environments, and in using action-set subsumption of the children of the GA operation rather than over all members of the action set.

The first stage will seek to provide base-line results to demonstrate that XCS is capable of learning the stable payoff for each of the actions in each state of the

environment. In this stage a population of classifiers with no generalisation sufficient to cover all the state × action pairs in the environment will be introduced. Without activating the induction mechanisms, XCS will be run to seek to identify the stable predictions. Wilson's error measure [1] is insufficient to track the error over the chain of actions, since it ignores the magnitude of error relative to the actual payoff received. Instead, a measure termed System Relative Error is used [9][15]. The Relative Error of an action set is the absolute difference between the payoff and the current action set prediction as a proportion of the larger of these two values (i.e. the relative magnitude of the rate of convergence). The System Relative Error is the average of the Relative Error measures calculated from the action sets that lead to an external reward, measured only during exploitation episodes. To help identify the extent of the Relative Error contributing to this average, the minimum and maximum Relative Error measures in an episode are also recorded. For each test XCS will be run ten times and the results presented will be the averages of these runs unless otherwise stated. To capture the degree of coverage of the problem environment a reporting technique is introduced from the field of Data Mining, previously applied to XCS in [19].

The second stage increases the difficulty of the experimental task. Each environment is presented without an initial population. With generality turned off, the induction mechanisms must establish a population of specific classifiers with payoff predictions to map the state × action × payoff space. These experiments will be used to assess the validity of hypothesis 2. The final stage requires XCS to identify and establish optimally general accurate classifiers to map the state × action × payoff space. Each environment is presented to XCS to identify whether hypothesis 1 is supported.

7. Baseline Results

Each of the test environments was presented to an XCS implementation [15] in turn, capturing the output of 10 runs in each environment. After the runs for each test environment were completed the performance was captured in a chart representing the number of iterations to the reward state, the system relative error alongside the minimum and maximum relative error in the action chain, and the population size. The final measure is irrelevant in this stage since the population is

pre-installed and fixed. The coverage tables for each run were also captured and averaged. The resulting coverage table for the environment was pictured using a line graph with a line for the predictions for the optimal action in each state and a line for the sub-optimal action in each state. A line graph was adopted to emphasise the payoff relationship between the data points. For the 25 and 30 length environments the line graphs were also re-plotted using a logarithmic prediction scale to reveal the differences in the predictions of the classifiers covering the early states of the environment.

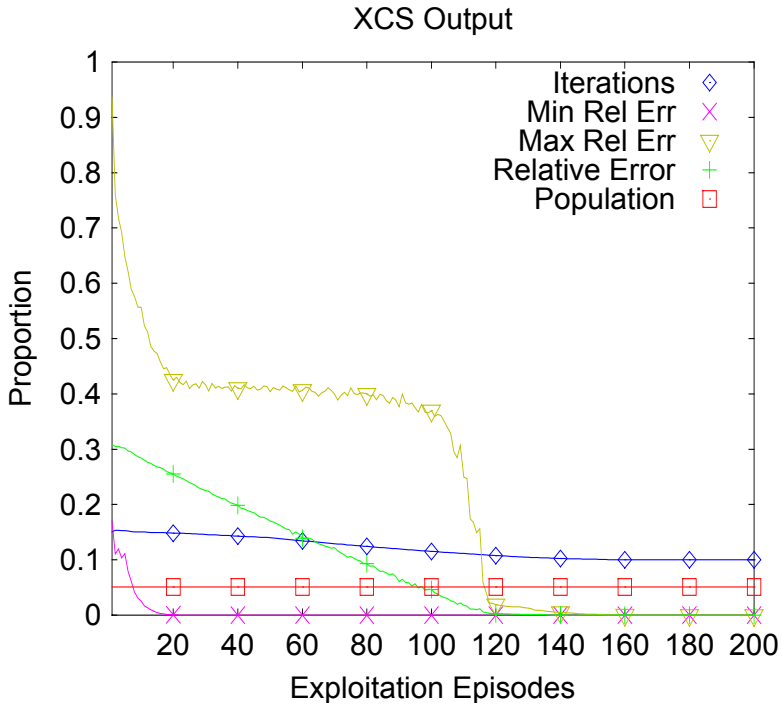


Figure 2 - The reduction of System Relative Error within action chain length 30 in a corridor FSW environment with one non-optimal action per state, a pre-loaded population and no induction.

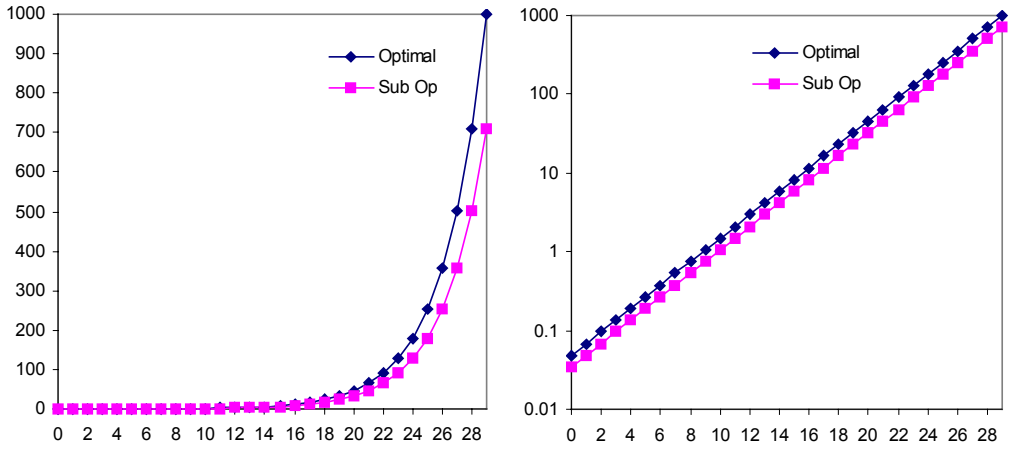


Figure 3 - The convergence of payoff prediction over the optimal and sub-optimal action chains within a length 30 corridor FSW environment with one optimal and one sub-optimal action per state, a pre-loaded population and no induction (*N*=2000, Iterations per Episode=300, Condition Size=8)

In all cases the system relative error rapidly drops to zero within 150 exploitation episodes indicating rapid convergence even within long action chains. Figure 2 illustrates the learning performance for the length 30 environment (60 states) and figure 3 shows the prediction and logarithmic prediction graphs for this environment by way of illustration. As would be expected in a pre-loaded population without induction, there was no error evident in the final prediction even for the earliest states despite the fact that the optimal payoff at this state would be 0.034 and the sub-optimal 0.024.

8. Providing induction in long action chain learning

The next series of experiments removed the population initialisation, starting XCS off with no initial classifiers. All the classifier induction capabilities of XCS were enabled, but the generalisation probability was set to 0.0 so that only fully specific classifiers would be generated. These experiments were specifically designed to investigate hypothesis 2, although they also provide interesting time-to-convergence comparative results with the previous experiments. The results for the length 30 run are given within Figure 4. Other runs are not shown due to space constraints.

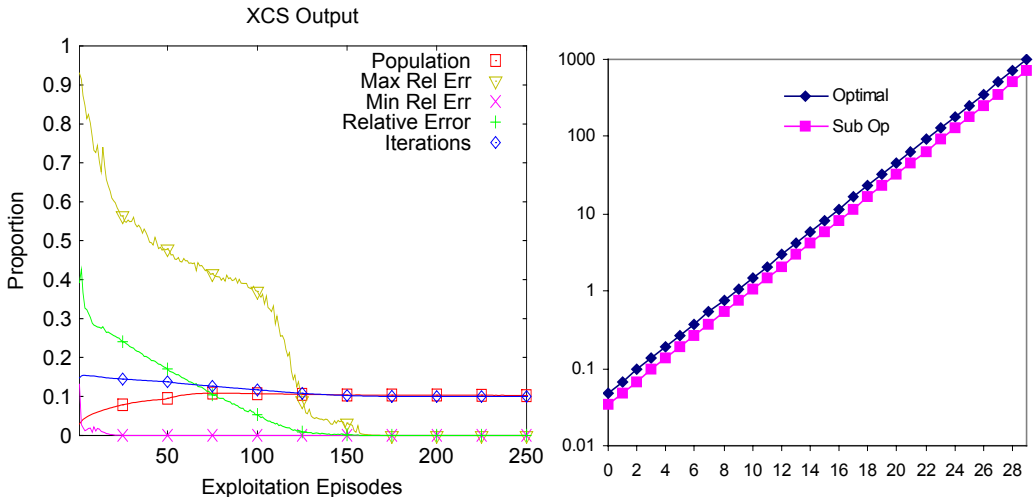


Figure 4 - The convergence of payoff prediction in the presence of classifier induction within action chain length 30 in a corridor FSW environment with one non-optimal action per state.

The results demonstrate that the time to system relative error reduction to zero is very similar to that for the situation where the initial population is already provided. Similarly, the time taken for XCS to be able to accurately select the optimal route is almost unchanged. As the action chain length grows to 25 and 30 steps the time taken for the system relative error to reduce to zero increases compared to the non-induction test. It is likely that this is due to the additional time taken to discover all the required classifiers, marginally lengthening the amount of feedback required to focus upon the very small predictions in early states.

The payoff prediction plots within figure 4 indicate that XCS was able to accurately predict the payoff for all classifiers right up to 30 states, and was able to select the optimal route by 150 exploitation episodes. This finding was counter to hypothesis 2, which suggested that XCS would be unable to select over the very small payoff predictions in these early states. It was initially thought that the averaging contained within the standard XCS report of iterations (it is the moving average of the previous 50 episodes) may be hiding occasional sub-optimal selections. The 30-action chain test was thus repeated with this averaging removed. Each of the 10 test results were then individually checked, since further test averaging could also hide some results. A typical run is shown in figure 5.

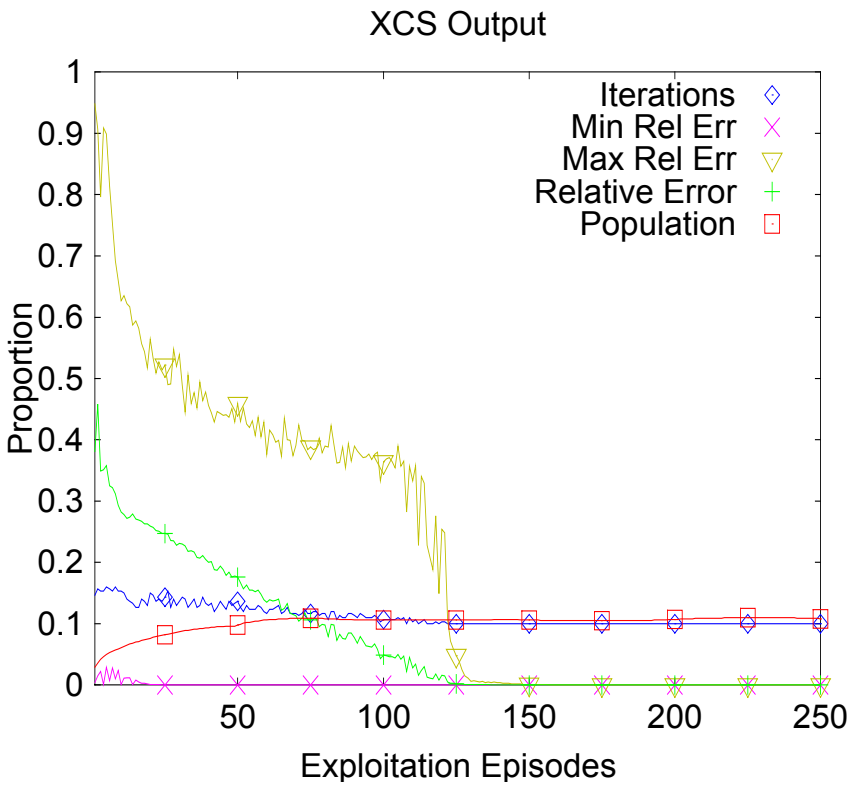


Figure 5 - A single test using a 30 length action chain FSW with induction, no generality, and iteration averaging removed.

It is clear from figure 5 that the average, in this case, does present a true picture. Clearly XCS was able to identify the optimal pathway, even through the very small payoff prediction states. In an attempt to test the hypothesis on a more extreme case, a 40 action chain environment was constructed by extending the FSW environment in the same way that the other environment were created from the five state environment pictured in figure 1. The result of a typical run from this test is shown in figure 6.

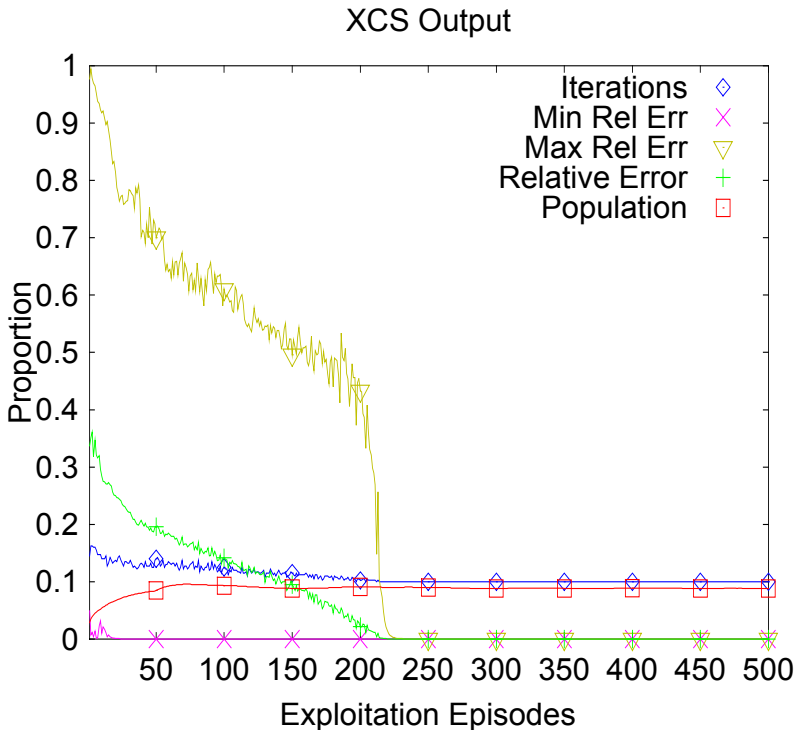


Figure 6 - A single test using a 40 length action chain FSW with induction, no generality, and iteration averaging removed.

Condition	Action	Prediction	Error	Accuracy	Fitness	Num	AS	Exp
00000000	0	0.001582	0.0000	1.0000	1.0000	24	24	7333
00000000	1	0.001123	0.0000	1.0000	1.0000	24	24	2667
00000001	0	0.001582	0.0000	1.0000	1.0000	33	33	2643
00000001	1	0.002227	0.0000	1.0000	1.0000	23	23	7357
00000010	0	0.003137	0.0000	1.0000	1.0000	25	25	7353
00000010	1	0.002227	0.0000	1.0000	1.0000	32	32	2647
00000011	0	0.003137	0.0000	1.0000	1.0000	24	24	2609
00000011	1	0.004419	0.0000	1.0000	1.0000	20	20	7391
00000100	0	0.006224	0.0000	1.0000	1.0000	21	21	7359
00000100	1	0.004419	0.0000	1.0000	1.0000	25	25	2641
00000101	0	0.006224	0.0000	1.0000	1.0000	27	27	2618
00000101	1	0.008766	0.0000	1.0000	1.0000	30	30	7382
00000110	0	0.012346	0.0000	1.0000	1.0000	20	20	7411
00000110	1	0.008766	0.0000	1.0000	1.0000	29	29	2589
00000111	0	0.012346	0.0000	1.0000	1.0000	26	26	2598
00000111	1	0.017389	0.0000	1.0000	1.0000	20	20	7402
00001000	0	0.024491	0.0000	1.0000	1.0000	23	23	7430
00001000	1	0.017389	0.0000	1.0000	1.0000	27	27	2570
00001001	0	0.024491	0.0000	1.0000	1.0000	26	26	2633
00001001	1	0.034495	0.0000	1.0000	1.0000	21	21	7367

Table 2 - Classifiers covering the states s_0 to s_9 of a 40 action chain length FSW.

Even though the optimal prediction payoff in state s_0 is now 0.00158 and the sub-optimal prediction in the same state is 0.001123, it is clear from figure that XCS is able to identify these predictions sufficiently accurately to select correctly

between them in a close to optimal number of learning cycles. The classifiers covering the first ten states of this environment from the same run are given in table 2, illustrating that all prediction values are accurately represented. Clearly this trend could not be infinitely extended, but combined with the ability to modify γit is clear that a non-generalising XCS can rapidly learn optimal routes in this kind of environment for long action chains. The hypothesised disruption of the payoff prediction caused by continued learning with XCS was not demonstrated, although it may be seen within environments with noisy feedback - an area for further investigation. Hypothesis 2 is therefore not upheld.

9. Learning with Generalisation

This series of investigations re-used the action chain FSW environment used in the previous tests. Parameterisation was kept the same as that within section 7 apart from the setting of the generality parameters to 0.33.

9.1 The Length 5 FSW

The initial experiment was carried out with the length five FSW pictured in figure 1. The results of this test are pictured in figure 7. These results show that under the action of generalisation XCS is able to identify the optimal action route for action chains of length five within this FSW environment by 350 exploitation episodes. This is just under six times the time required for the fully-specific induction test; the search space has undergone a 7.5 times increase. The rise and then decline of the population curve demonstrates that XCS is able to identify the correct generalisations. An analysis of the population revealed that [O] was present. Unfortunately an analysis of the numerosity of the members of [O] revealed that other classifiers within the accurate sub-population were not as dominated by the members of [O] as would normally be expected.

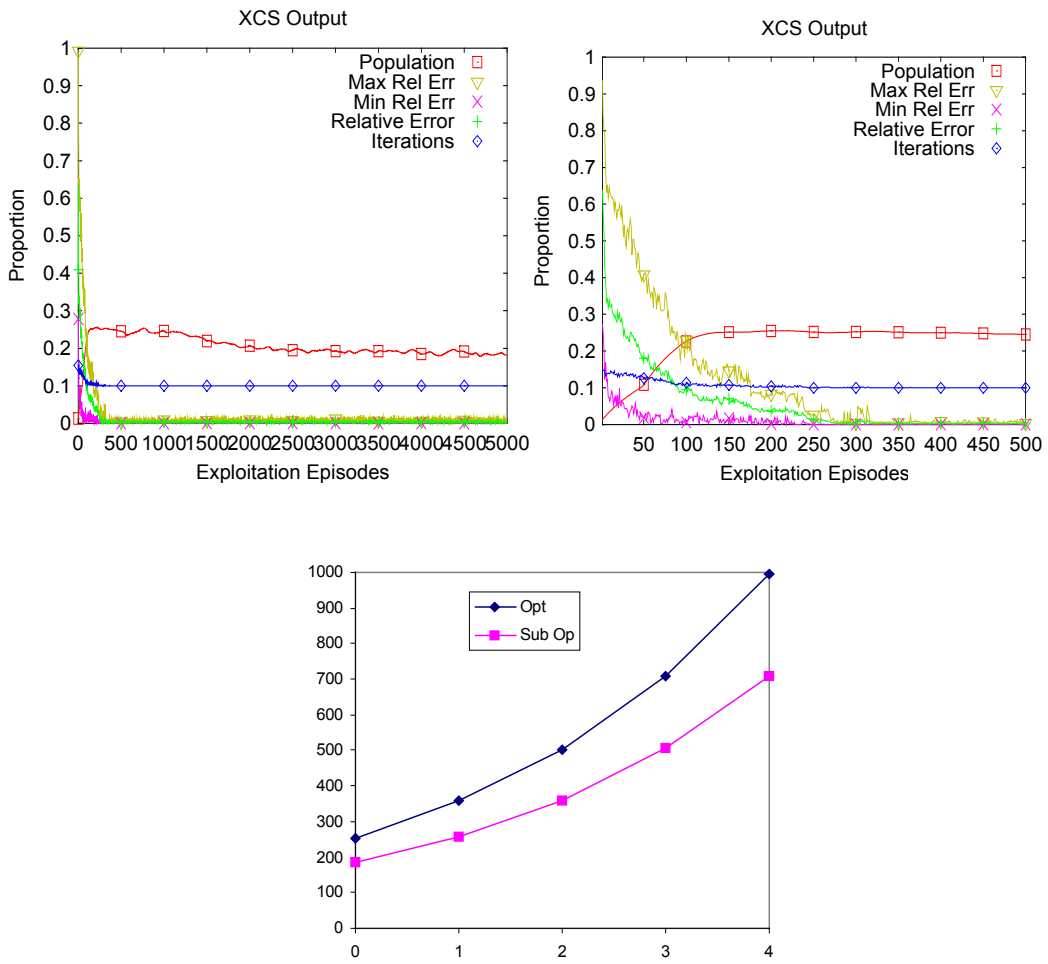


Figure 7 - The convergence of payoff prediction in the presence of generality pressure and classifier induction within a length 5 corridor FSW.

An analysis of the accurate sub-population revealed that there were over-specific classifiers within the accurate sub-population that were maintaining a higher than expected numerosity. These classifiers would normally be subsumed on creation within the G.A., but classifiers whose actions are mutated will not be subsumed because they no longer belong to the parent action set. Although these classifiers should be removed by normal deletion dynamics, they appear to remain within this environment, though less dominant nearer the reward state. This remained the case in an expanded run of the test that ran for 30000 exploitation episodes.

To test this hypothesis the mutation rate was lowered from 0.04 to 0.01, as used in Wilson's Woods-2 experiments [1][2]. Over ten runs for both the 0.04 mutation rate and the 0.01 mutation rate the dominance of the optimally general classifiers was captured as a percentage of the action-set size for each action set. An F-Test

of the action set dominance for the optimal route between the 0.01 mutation rate and the original 0.04 mutation rate revealed that the variance for each set was not equal (F=0.4262, F-crit=0.3146). A one-tailed Wilcoxen test was therefore applied to test the hypothesis that the introduction of the 0.01 mutation rate would improve the focus of the action sets as measured in the dominance statistic. This test revealed a significant difference (T=3, T-crit=3 at 0.005) between the original and new dominance rates. Given that the average dominance with mutation rate 0.04 was 60.29 and the average dominance with mutation rate 0.01 was 69.17, it was concluded that the lower mutation rate improves domination of the action set for the optimal route. The same tests were applied to the sub-optimal route, revealing that the variances are not equal (F=0.7211, F-crit=0.3146 at the 0.05 level) and there is a significant difference between the two sets of results (one-tailed Wilcoxon test: T=0, T-crit=3 at the 0.005 level). The average dominance with mutation rate 0.04 was 60.91 and the average dominance with mutation rate 0.01 was 70.44. Therefore it was concluded that the lower mutation rate also improves domination of the action set for the sub-optimal route.

9.2 The Length 10 FSW

Having confirmed that XCS is capable of identifying [O] within the simplest of the test environments, the problem complexity was expanded to the length 10 environment. Initially the 0.01 mutation rate was applied to this environment, having produced better results in the previous test. The performance of XCS within this environment and the coverage graph achieved is pictured in figure 8.

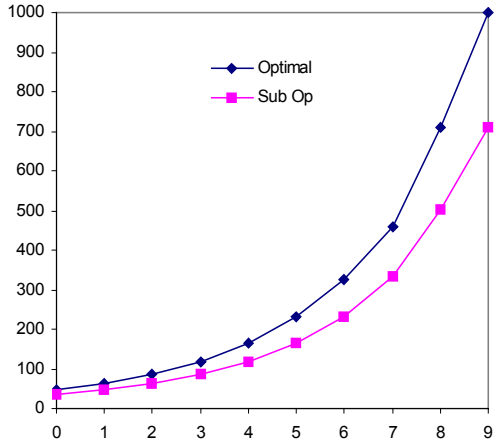


Figure 8 - The average performance of ten runs of XCS within a length 10 corridor FSW with system relative error reducing to zero over 15000 episodes.

A statistical comparison of the dominance rates for ten runs using 0.01 and 0.04 mutation rates revealed that there was no significant difference for the optimal route ($t=1.35$, $t_{crit}=1.729$ at 0.05 level), although for the sub-optimal route the 0.01 mutation rate continued to provide improved results ($t=6.478$, $t_{crit}=1.729$ at 0.05 level). However, the 0.04 mutation rate runs were able to reduce the maximum relative error curve much earlier (3000 exploitation iterations compared to 11000 exploitation iterations).

Although difficult to spot, the 'iterations' plot in figure 8 indicates occasional non-optimal path choice. If investigating the hypothesis on the location of disruption lying within the early states is correct (hypothesis 1), then any sub-optimal route choice should take place within these early states. To further investigate this hypothesis the frequency of non-optimal route choice within the 0.04 mutation rate run was recorded and a typical run from the 10 runs chosen. The incorrect action choices within last 2000 exploitation episodes were extracted (the System Relative Error had reduced by 3000 exploitation episodes). The frequency of sub-optimal action choice per state was plotted in a histogram, shown in figure 9. This indicates that all of the incorrect decisions were made from state s_0 (moving to state s_{10}), state s_1 (moving to state s_{11}), state s_2 (moving to state s_{12}) and state s_3 (moving to state s_{13}), with generally decreasing frequency as process is made from the early states. Since the dominance results and the coverage graph demonstrate that the payoff prediction of these early states has been discovered and is being maintained by the action sets, the results in figure 9 lend strong support to the hypothesis that the early states are much more influenced by the additional classifiers added for exploration of the problem space. If this finding is correct, the problem should become much worse as the action chain is lengthened, threatening the ability of XCS to produce and maintain [O] over long action chains, as hypothesis 1 predicts.

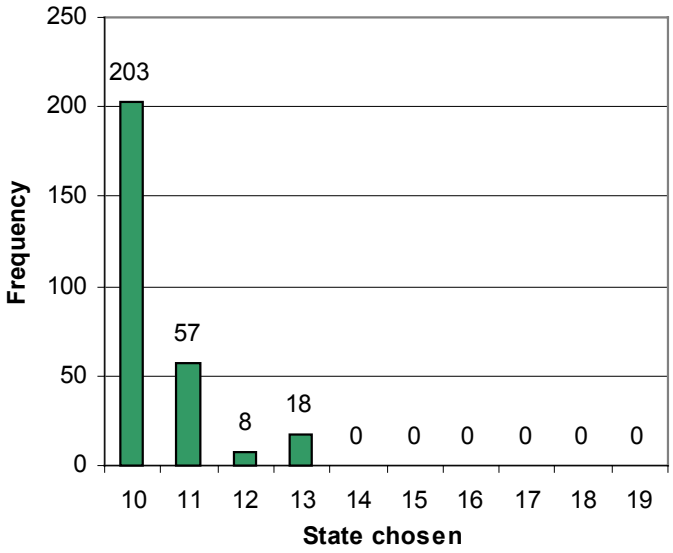


Figure 9 - The rate of choice of non-optimal route within the last 2000 exploitation episodes of a typical run of XCS within a length 10 corridor FSW with the mutation rate set at 0.04.

9.3 The Length 15 FSW

The experiments within the length 15 test environment indicate that this analysis is indeed correct. Figure 10 pictures the performance results for the 0.04 mutation rate run and the averaged coverage graph.

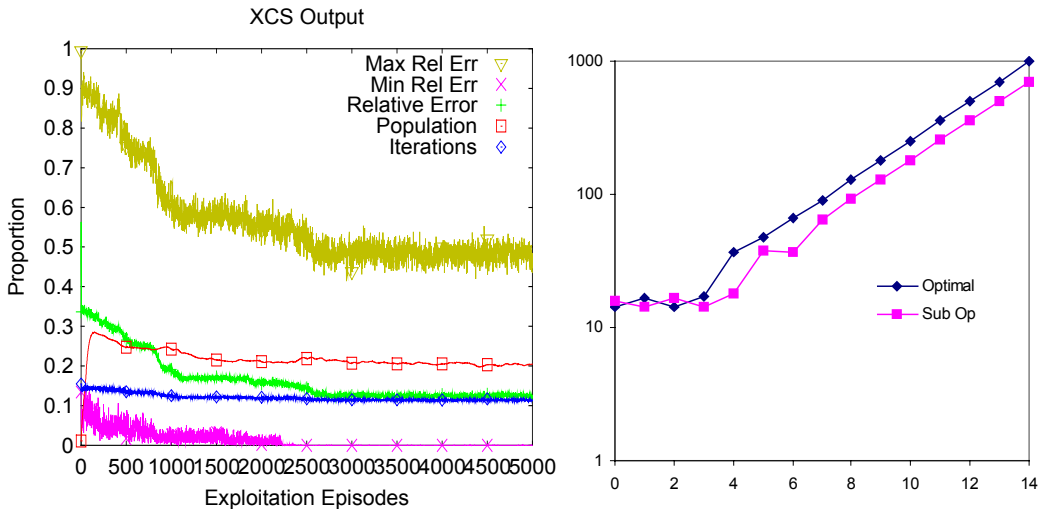


Figure 10 - The averaged performance of ten runs of XCS in the length 15 FSW, at mutation rate 0.04 and the coverage graph with logarithmic prediction scale (y axis) for the action sets. The average iteration count in the last 2000 steps was 17.15 rather than the optimal 15.

In these runs, all ten populations found and were able to establish all members of [O] apart from those covering states s_0 to s_3. These four states were represented by a single classifier for each action. As was predicted and figure 10 illustrates, the additional classifiers in the population within the early action sets disrupt the payoff prediction so that the payoff differences within the first four states cannot be differentiated by XCS. If XCS cannot differentiate the payoffs, it will generalise over these states. Thus, hypothesis 1 is shown to hold, and it is possible to tentatively suggest that an action chain length of 11 steps appears to represent the limits of reliable generalisation over the action chain for this environment and parameterisation. However, it was noticeable that the generalisation occurred over bits 0 and 1 and therefore it may be that the convenience of the input encoding encouraged this limit point.

It had been expected from previous results that XCS with a lower mutation rate would produce better results. In figure 10 the population curve remains high, indicating the presence of additional accurate but more specific classifiers within the population. If a lower mutation rate was able to focus this population further, as was the case previously, then the lower amount of competition within the action sets may allow XCS to distinguish between more of the early states. When the experiment was repeated with a mutation rate of 0.01 the maximum relative error curve did not reduce below the 0.8 level at any time in the run. From an examination of the populations produced it was found that within the 0.01 mutation rate experiment only three of the ten runs produced the expected coverage results. Indeed, the coverage tables exhibited action sets with seemingly impossibly high numerosity, and predictions lay between 30 and 50 for most action sets. These values are depicted within figure 11, averaged from the seven poorly performing runs.

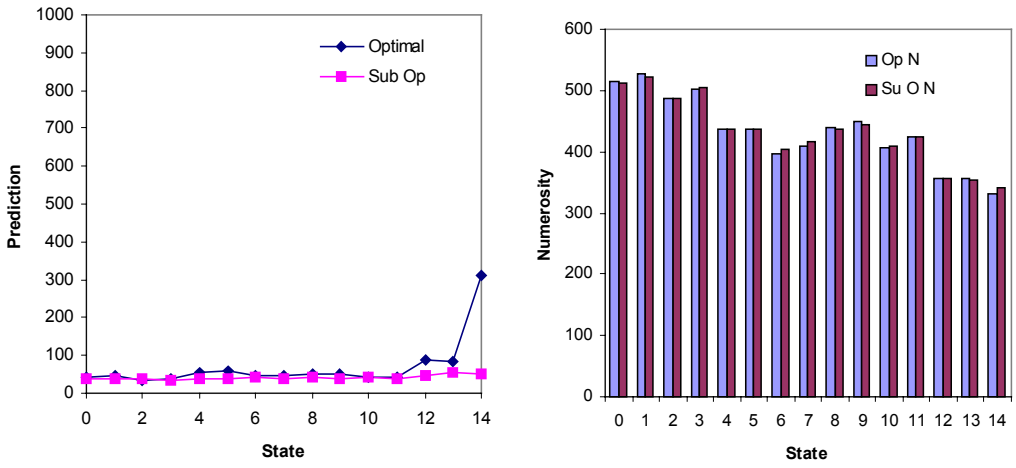


Figure 11 - The coverage graph and the histogram of average numerosity for the action sets within the seven non-optimal runs of XCS at mutation rate 0.01 within the length 15 FSW.

The reason for this difference can be readily located through an examination of the coverage tables and the populations of the runs. An inspection of the poorly performing populations revealed that a fully general classifier in each of the two action classes had established themselves and gathered a very large numerosity. Since these classifiers appeared in every action set, they increased the numerosity of the action set producing the abnormally high numerosity results pictured in the histogram in figure 11. Intriguingly, these classifiers, where present, were all apparently of low accuracy and numerosity within the 0.04 mutation rate XCS.

9.4 The Dominance of Fully General Classifiers

A detailed inspection of the populations revealed that in each population case one of the two fully general classifiers continued to be considered of high accuracy even though they cannot accurately reflect the payoff for their recommended action in all environmental states. The key to explaining the phenomena therefore lies in an explanation of how these fully general classifiers can be considered to be accurate. It was hypothesised that the continued preference for the fully general classifiers would arise from the discovery of the fully general classifiers whilst the population is in its initial stages without a good convergence on accurate classifiers. Since all action sets apart from the one leading to the reward state receive payoff as the discounted maximum system prediction of the next match set, in the early states the predictions will be low and even when discounted will

remain similar. If the action chain is long enough it is possible that the generalists have sufficient time to become accurate over a large proportion of the action chain and therefore gain a larger relative accuracy (fitness) than other competing classifiers. Once this is the case, the general classifiers could utilise the fact that they will obtain many more G.A. opportunities to accumulate numerosity. Once a sufficiently high numerosity is established the fully general classifier would exert a large influence over the action set, keeping the prediction within each action set close to that of the fully general classifier. The inaccurate prediction would become the payoff to earlier classifiers, allowing them to accurately reflect an incorrect payoff and promoting the false accuracy of the over-general. This would in turn enhance the ability of the fully general classifier to proliferate. If the reward input is infrequent in relation to the internal payoff, and if the initial states cannot be disambiguated due to their distance from the reward state, a breeding ground for the fully general classifier would exist. Once fully general classifiers are established, true members of [O] would be considered inaccurate at their true prediction and would be unable to compete to drive out the fully general classifiers. This is a vital hypothesis that identifies a genuine limiting factor on the ability of XCS to find and establish [O] within a long action chain environment, and will be more formally expressed as the *Domination Hypothesis*:

> *Fully general classifiers covering each action set will be considered inaccurate in all environments where the path to a stable environmental reward is greater than length 1 except when each of the following holds: 1) the classifiers acquire a high relative accuracy before stable environmental feedback can be passed through all the states to establish the true stable payoff for each action in each state; 2) the classifiers participate frequently in the Genetic Algorithm so that their numerosity becomes high relative to that of other competing classifiers; 3) the time [in iterations] between stable environmental reward is more than the time required to remove the influence of that reward from the recency-weighted prediction calculation of the fully-general classifiers. In these circumstances the fully-general classifiers will dominate the prediction calculation of the action-set, provide payoff to their participation in other action sets that is close to their own prediction, and so increase their relative accuracy and thus their*

participation frequency within the Genetic Algorithm. In this state the true optimal sub-population is considered inaccurate and cannot be established by XCS.

Before this hypothesis is investigated further the question of why this phenomenon occurred within the 0.01 mutation rate environment and not within the 0.04 environment must be considered. A possible explanation lies in the mutation rates themselves. A higher mutation rate, even though only marginally higher, had a profound effect on the operation of XCS in the length 5 and length 10 FSW environments. It is therefore hypothesised that the difference in mutation rate allows sufficient early exploration within the 0.04 rate runs to prevent the early establishment of the fully general classifiers. In contrast, within the 0.01 rate runs the lower exploration rate that previously encouraged population focus now limits competition with the fully general classifier too much. If this hypothesis is correct, it would suggest that a profitable area of further work would be to examine the introduction of a dynamic control of the mutation rate within XCS.

To investigate the Domination Hypothesis further the 0.04 mutation rate experiment was re-run. The version of XCS used in these experiments used a simplified form of action-set subsumption, discussed in [15], to that identified in [14]. This form does not regularly compact the action set down to its minimal most general form, and so leaves the removal of over-specific classifiers to the G.A. This modification allows the ability of the G.A. to establish the dominant classifiers to be examined. It was hypothesised that if the fully general classifiers were obtaining full accuracy as suggested then the introduction of any higher subsumption pressure should serve to further establish these classifiers and increase the likelihood of fully general classifier domination of the population. A population-wide subsumption mechanism was thus introduced which, if a new classifier produced by the G.A. is not subsumed by its parents or any member of the G.A., looks for a subsuming classifier in the population as a whole. This is not as severe a subsumption mechanism as the action-set subsumption in [14] but does increase the subsumption pressure.

The results of this experiment are shown in figure 12. It was noticeable when comparing figures 10 and 12 that the XCS performance has decreased even though the population subsumption should have encouraged the formation of [O] through decreased competition in the population. An examination of the populations revealed that three of the ten runs did now contain dominant fully general classifiers that had prevented the formation of [O] and that these were the cause of the drop in performance. Thus the introduction of population subsumption appears to have aided the dominance of the fully general classifiers.

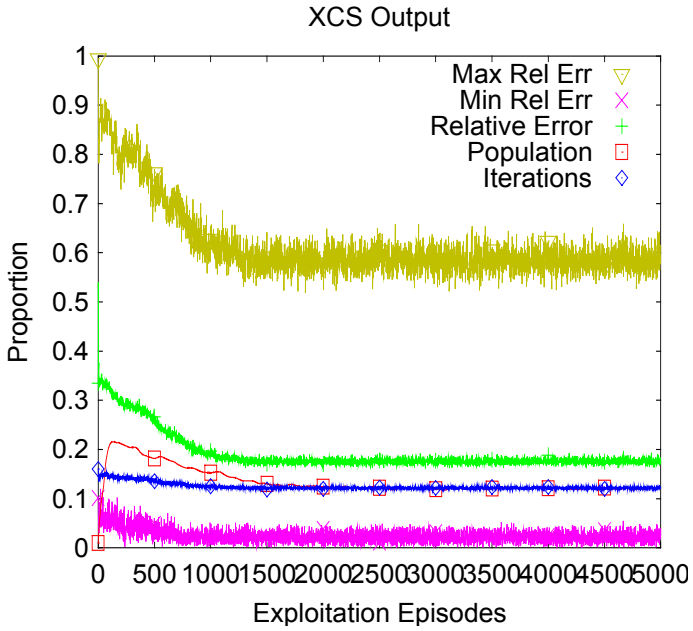


Figure 12 - The average performance of ten runs of XCS in the length 15 FSW with mutation rate 0.04 and population-wide subsumption in the G.A.

Further experiments using the population subsumption version of XCS were performed that reduced the population size limit from 1200 micro-classifiers to 800 and then 600 micro-classifiers. It was thought that by limiting the population size pressure would be exerted on the fully-general classifiers that would limit their formation. In fact, this was an incorrect assumption - the reduction in population size only served to reduce the space for exploration and therefore increase the likelihood that fully general classifiers would dominate the populations.

A further test of the hypothesis involves extending the action chain once more. If the hypothesis on dominant classifier formation is correct, the increased length of

the action chain will make environmental reward less frequent and allow the fully general classifiers to establish themselves more easily. These tests will also allow further verification of the tentative hypothesis on the limits of action chain length before the inability to accurately identify payoff prediction is seen. The length 20 FSW environment was therefore re-introduced to XCS and, following the results of the length 15 experiments, the mutation rate was set at 0.04. Ten runs within the environment were conducted, and the performance was captured and averaged. The average performance of XCS in this environment is shown in figure 13.

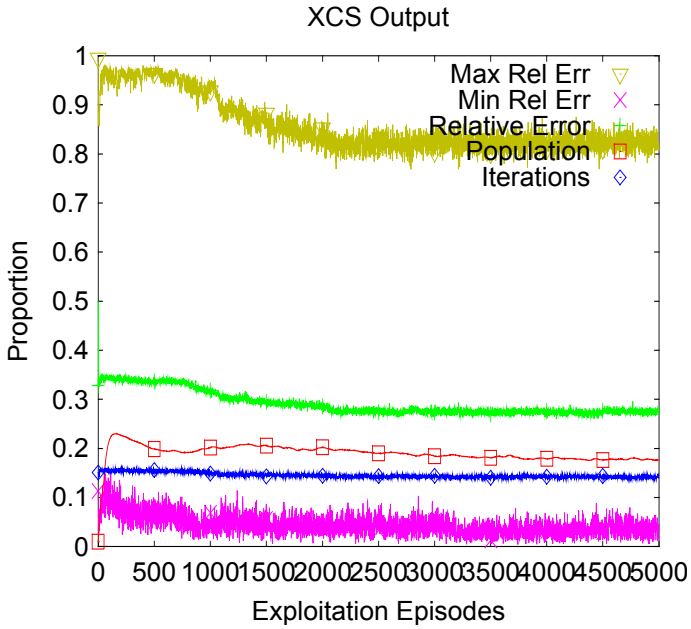


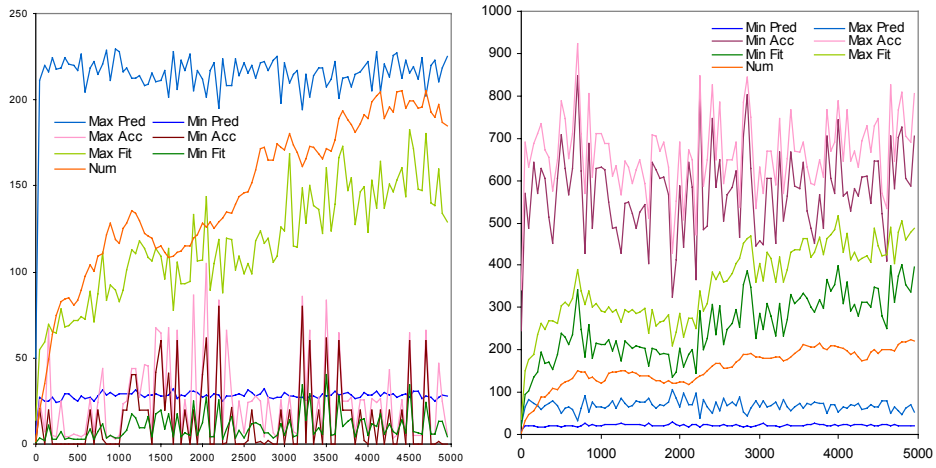
Figure 13 - The average performance of ten runs of XCS in the length 20 FSW.

In figure 13 the system relative error remains high and does not appear to be able to be further reduced. The number of iterations required to reach the reward state is much higher than the optimal. An analysis of the coverage table for each population revealed that only five of the ten runs produced adequate coverage. Examining the populations of those runs that did not produce the expected coverage revealed that each contained dominant full-generality classifiers of high numerosity. These results confirm the hypothesis that the longer the action chain to a regular environmental reward, the more likely a dominant full generalist is to appear.

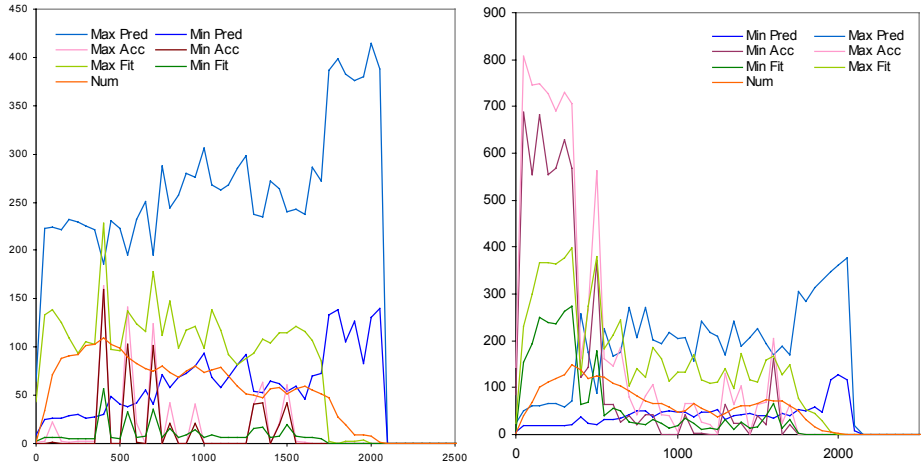
In order to investigate the formation of the dominant fully general classifiers further the length 15 FSW with 0.01 mutation was re-run with additional reports

added to capture the fitness, accuracy, prediction, and numerosity of the fully general classifiers in the population averaged over each episode whilst they exist. These details from twenty runs were output to a file and the results corresponding to typical good performance and poor performance populations were extracted and plotted. Figure 14 identifies the performance of these classifiers.

a) Two fully general classifiers dominating the population



b) Two high numerosity fully general classifiers removed from the population



c) Two fully general classifiers rapidly removed from the population

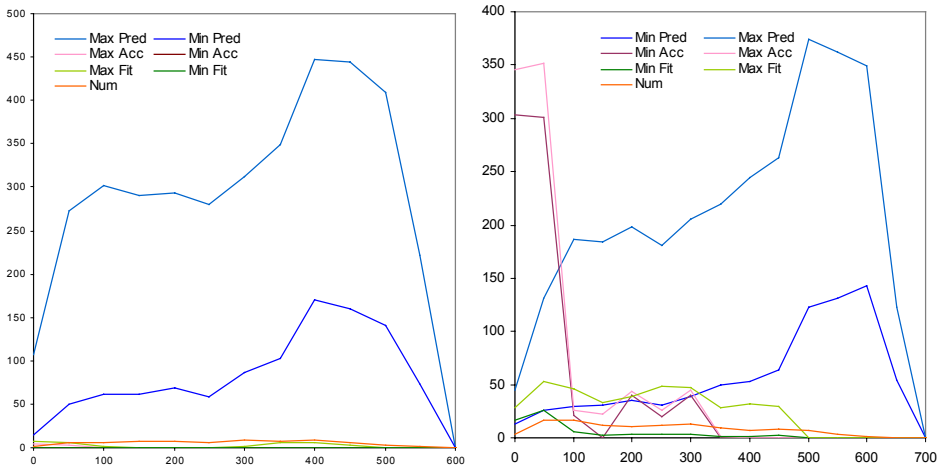


Figure 14 - The prediction, accuracy, fitness and numerosity traces of three typical life histories of the two fully general classifiers that may appear within the Length 15 FSW. Each measure averaged over the preceding 50 exploitation episodes.

It was immediately noticeable that some populations did contain fully general classifiers that accumulated large numerosity values and yet were able to remove these classifiers to establish more accurate classifiers (see figure 14b). Other good populations were able to remove the fully general classifiers without allowing them to achieve significant numerosity (figure 14c). The poor performing populations retained their high numerosity classifiers (figure 14a). Although figure 14 provides plots of typical runs, it should be noted that when each of the populations in the run was plotted in this manner, those falling into any one of these three categories of life history produced similar plots of prediction, accuracy, and fitness. This would suggest that there is a factor underlying the phenomena that generates these typical trends.

An analysis of the graphs in figure 14 suggests that the early appearance of the fully general classifiers is an important factor, since in all cases the fully general classifiers did appear very early in the exploration (in exploration episode 28, 37, and 23 respectively) and within similar population sizes (95, 88, and 86 respectively). However it is also clear that this is not in itself a sufficient condition, as the hypothesis identifies. The key aspect of the graphs that is worthy of further investigation is the fact that all the while that the accuracy of the classifiers is kept higher than zero the numerosity of the classifier increases. The actual level of accuracy appears to be fairly irrelevant - it would appear from an

analysis of the populations that the accuracy of these classifiers was higher than its competitors, and thus it is the *relative* accuracy that is the important feature. (It should also be noted that since the accuracy level never reaches 1.0 the increase in numerosity of the fully general classifiers must be through the action of the G.A. rather than the subsumption mechanism). It is the sudden reduction of the accuracy of classifier 2 in figure 14b that signals the start of the demise of that classifier. A comparison of classifier 1 in 14a and classifier 1 in 14b is instructive. Whilst they both display a low accuracy, classifier 1 in 14a is able to continuously regain sufficient accuracy to give it the fitness necessary to compete within the G.A. and replicate itself. Classifier 1 in 14b is unable to sustain its accuracy and on each period of zero accuracy it loses numerosity. Once numerosity is lost the ability of an inaccurate classifier to dominate the action sets is reduced. This in turn causes the action sets to start to gain their true payoff prediction, giving the fully general classifier a lower relative accuracy. As the accuracy continues to reduce, so its ability to compete in the G.A. is reduced, causing the classifier to be gradually driven out of the population. The flatter prediction curve of classifier 1 in 14a when compared with classifier 1 in 14b indicates that it has been able to exert sufficient influence to push down the payoff predictions in the action sets so that it can maintain a high relative accuracy. Interestingly, classifier 2, representing the sub-optimal pathway, is more able to control the range of the prediction. This is actually an artefact of the action classifier 2 represents. This action never leads directly to an environment reward state and therefore all system predictions in the action sets the classifier occurs within are fed from other action sets. The control of prediction is thus a simpler task, giving higher accuracy and fitness. Although further investigation clearly needs to be carried out to further explore the Domination Hypothesis, these results do lend considerable support to the Hypothesis.

9.5 The Length 20 FSW

Although the performance in the Length 20 FSW experiment pictured in figure 13 demonstrated a high System Relative Error, there is a large element of this averaged result that is attributable to the five populations that developed dominant fully general classifiers and thus were unable to develop a useful state × action × payoff map. Once the performance of XCS with these poor populations is

removed the true performance of XCS in this length 20 environment can be seen (figure 15). Interestingly the results illustrate that XCS is able to learn the payoff predictions within the last 11 states but produces general classifiers to cover the earlier states. This finding is in agreement with the earlier tentative hypothesis that with the particular parameterisation used and within this form of environment XCS is able to represent accurately with optimal generalisation up to 11 actions in the action chain. However this result can to some extent once again be explained by message coding convenience. The classifiers representing states s_0-s_7 and state s_8 are shown in table 3. Two classifiers cover states s_0 to s_7, which can each be conveniently represented by generalising the first three condition positions. State s_8 is much more difficult to generalise alongside the other states and would be unlikely to appear within any other generalisation. Thus, even if the environment were extended to length 21 it may be that the representation of s_8 and s_9 would remain more accurate than that of states s_0 to s_7 purely because of the difficulty of including s_8 and s_9 in a generalisation over the earlier states. Clearly this aspect warrants further investigation.

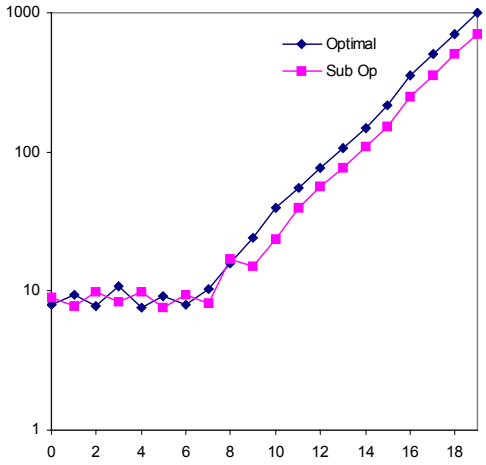


Figure 15 - The average system prediction in each action set of the five good performance runs of XCS in the length 20 FSW environment.

Classifier	Pred.	Error	Acc.	Fitness	N	AS	Exp.
#000###→0	5.795	0.0007	0.9208	1.0000	43	54.8	28581
#000###→1	8.520	0.0030	0.8875	1.0000	37	46.8	42037
###10#0→1	13.092	0.0082	0.6207	1.0000	13	57.1	7395
###100#→0	14.474	0.0055	0.2645	1.0000	6	51. 5	3619

Table 3 - The accurate classifiers covering states s_0 to s_8 within the length 20 FSW.

Considering the increasingly poor performance of XCS within the length 15 and 20 FSW environments, further experiments with length 25 or 30 FSW environments were considered unnecessary and the experimental investigation was terminated at this point.

10. Summary of Results

Baseline investigations using a preloaded population of classifiers within the chosen progressive two-action corridor FSW environment of increasing length demonstrated that XCS was able to rapidly identify the payoff predictions for classifiers in chains of up to length 30. The length 30 limit was chosen as a hypothesised maximum size for use within later experimental work, and does not reflect a true maximum length for XCS without induction mechanisms.

In testing the validity of Hypothesis 2 within the same test environments, XCS was run without generalisation but with an initially empty population and induction mechanisms enabled. Contrary to expectations, XCS was able to establish and proliferate the optimal population and identify the correct payoff predictions in all of the test environments. The results indicated that XCS was able to complete this task within approximately the same time as that required to establish the payoff predictions when supplied with an initial population. This may, however, be an artefact of this environment, which introduces very little exploration complexity. A further experiment extended the environmental length to a minimum of 40 actions to reward. In this test XCS continued to correctly identify all payoff predictions and utilise them to select the optimal route even though the payoffs for the two actions from the first state were very small. Thus, the hypothesis that the act of establishing the correct payoff predictions could itself cause a breakdown in the ability of XCS to differentiate between very small differences in the payoff predictions for early states in a long chain environment was not substantiated.

Investigation of Hypothesis 1 involved the application of generalisation in the learning of the optimal sub-population [O]. In application to the length 5 FSW XCS was able to establish and proliferate [O] in a slightly sub-linear time when compared to the increase in coding complexity. However, the dominance of [O]

was higher when using a mutation rate of 0.01 than when the mutation rate was 0.04. Investigation with the length 10 FSW showed that XCS was less able to establish a high dominance of [O] using either mutation rate, although the 0.01 mutation rate still demonstrated a measure of superiority. Furthermore, XCS was unable to consistently select the optimal route to the reward state. Further analysis revealed that XCS was selecting the incorrect route in the earliest states. Once the length 15 FSW was introduced this error became more pronounced, with an average of two sub-optimal actions chosen in each episode. The coverage table revealed that the predictions for alternative actions in the first four states were confused and overlapping, making it impossible to select deterministically the correct pathway. Comparison between mutation rates of 0.01 and 0.04 revealed that the higher mutation rate was now preferable. Further experiments with the length 20 FSW revealed that the early state payoff prediction confusion was continued although it was noticed that XCS was consistently able to identify correct payoff predictions for the last 11 states. It was recognised, however, that the 11-state threshold could be an artefact of the encoding used and cannot be used as a definitive limit for the parameterisation used.

Within both the length 15 and length 20 FSW environments it was noted that an increasing number of runs where [O] was not fully identified or proliferated occurred. This was identified as due to the emergence of strong full-generality classifiers. Although such classifiers should by definition be inaccurate and therefore be eliminated by XCS, they were able to establish a huge dominance of all action-sets. The *Domination Hypothesis* was proposed to explain this phenomena and a first verification of the hypothesis was presented by tracking the life history of fully general classifiers in the circumstances where they were established and where they were removed. It was noted that these classifiers need to be established early in the XCS operation and that they dominate when they are able to establish an early control of the system prediction so that they can continue to hold a non-zero accuracy. Once any period of zero accuracy is established they will rapidly lose fitness and thereafter numerosity. The dominance hypothesis represents an important phenomenon within XCS that is worthy of further study.

11. Discussion

Previous work within multiple-step environments with XCS have not investigated the limits that may apply to the length of action chains that can be learnt by XCS. Previous investigations with the CFS-C LCS implementation by Riolo [16,17,18] demonstrated limits in the formation of long rule-chains by the LCS. Using a single action corridor FSW of twelve states and a pre-set initial population of classifiers that provided the appropriate rule chain, Riolo [16] demonstrated that the rule-chaining mechanism proposed in [21] would allow the classifiers in the chain to converge to the same prediction value. He further demonstrated, using an environment with two length 10 single action chains and with a start-state having two actions to enable choice between the two chains, that a seeded population could learn to choose the optimal route. The work presented in this chapter used seeded populations only for base-line results and demonstrated that even in environments presenting a more complex choice XCS was able to select the optimal path within 120 trials, somewhat faster than the 170 trials of the traditional LCS within the simpler environment. In fact, with non-seeded populations and no generalisation XCS was able to learn the correct pathway within 140 trials. Clearly the lack of comparative parameterisation (and implementation details, such as the explore/exploit regime) makes simple performance comparisons impossible, and it would be naïve to claim from these results that XCS provides a faster or more effective learning environment than a traditional LCS.

Riolo [18] examined the ability of the LCS to establish rule chains under the action of the induction mechanisms - and in particular, using the Triggered Chaining Operator to establish rule-chains. This work was performed using the GREF-1 FSW environment - a length 4, four action environment using 16 states to provide four pathway with multiple links between the pathways. Whilst the performance improved with the introduction of the TCO, it remained weak. Riolo identified that although rule-chains were established, the LCS failed to maintain the rule-chains. Not only did the lower strength of the earlier rules prevent their duplication, thus keeping them safe from later deletion, but there were also parasitic rules that caused payment to earlier classifiers to reduce and so threaten their existence. Adding a "support" component to the bids of classifiers in the

rule-chain, a niche deletion mechanism to limit competitors within each match set, and a form of Create Effector Operator to introduce classifiers in poorly performing match sets helped to reduce parasites and increase the speed of formation and the maintenance of rule chains. These measures brought performance up to 90% of the optimal performance in 12,000 trials. Riolo concluded that the population needed to be treated more like an "*ecology of rules, with niches (states or situations) that support species (co-bidding rules) competing for limited resources (classifier-list space, message-list space, strength)*". Whilst the checks-and-balances approach of CFS-C and other traditional LCS approaches sought to achieve this balance, XCS is able to meet these requirements fully. The test environments used in the earlier experiments within this chapter do not match the GREF-1 environment for complexity, but the results with XCS presented earlier suggest that XCS will be able to establish and maintain action-chains for the GREF-1 environment with ease. To test this hypothesis, the GREF-1 environment was re-created and figure 14 presents the average of ten runs in this environment using a population limit of 800, no initial population, and all induction operators turned on. The other parameterisation is the same as that used for the experiments within section 9. The output was generated with an additional Performance curve to allow some degree of comparison with Riolo's results.

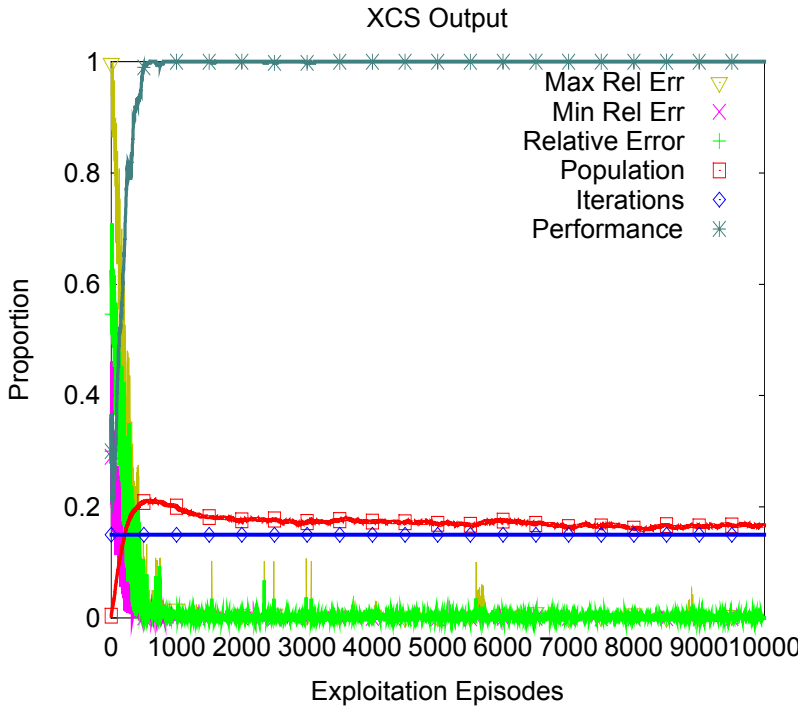


Figure 16 - The average performance of XCS in the GREF-1 environment.

It can be seen that XCS achieves optimal performance by 1000 exploitation trials (2000 trials) whereas even the best run presented in [18] achieved just over 90% performance by the equivalent of 10,000 trials. An examination of the coverage tables identified that all runs established a high dominance optimal sub-population identifying all action chains to the reward. In addition XCS had established the optimal generalisations, something that Riolo's work did not seek to achieve. Unfortunately, although these results suggest the superiority of XCS, any direct comparison is foolhardy given the differences in operation and parameterisation between these LCS implementations. These findings must therefore be expressed as *indicating* that XCS is able to learn a solution and achieve a better on-line performance within the GREF1 environment than Riolo's CFS-C.

Lanzi [6] noted the problems XCS faced when seeking to solve the Woods-14 test environment, and these were discussed in section 3. Whilst the Woods-14 environment requires an action chain of length 18 to reach the reward state, the structure of the environment itself makes a direct application of the results of this chapter to that environment problematic. Within the Woods-14 environment many of the successive steps require a different action to be undertaken, preventing generalisation over states and potentially encouraging the more accurate representation of the payoff prediction over the early states. As Lanzi notes, the increased number of actions available bring their own penalty, making it difficult for XCS to explore the environment sufficiently to reach the reward state and begin to feedback the predictions.

It is interesting to note the potential relevance of Lanzi's *specify* operator [5] to the results in this chapter. Although *specify* was introduced to tackle the problem of over-generalisation due to exploration inequality, it could be the case that a carefully application of the *specify* operator could counter the development of full-generality classifiers. *Specify* introduces new more specific classifiers in order to tackle over-generalisation due to uneven exploration. The introduction of more general classifiers within a long action chain would exert additional pressure on the fully general classifiers and provide more opportunity for mixing within the G.A. This would encourage the formulation of more optimal classifiers and may

lead to a break in the strangle-hold of the fully general classifiers. Unfortunately, *specify* would tackle the problem by tackling the symptoms rather than the cause. This is a matter for further work beyond the scope of this research.

12. Conclusions and Further Work

This chapter has sought to identify some limits on the length of action chain learning within XCS. Using a progressive two-action corridor FSW-based environment with a single start state and single terminal state it has been shown that whilst XCS can reliably and rapidly establish the optimal population and the correct payoff prediction mapping for optimal action lengths of up to (and possibly beyond) 40 actions in these environments where no generalisation is used, the introduction of generalisation introduces inadequate payoff prediction in the early states once the predictions are sufficiently close to generalise over without reduction in accuracy. This can be as early as 11 action-steps from the terminal states, although the precision of this limit may be dependent upon the generalisation convenience of the encoding. Whilst it is recognised that the limits identified are dependant upon the parameterisation used, the exploration complexity of the test environment, and the exploration strategy adopted, it is clear that XCS (at least in the version utilised here) faces clear and restrictive limits upon its ability to establish the correct payoff prediction mapping over the states of a test environment in the face of generalisation pressure.

Further investigation is required to establish the nature of these limits with alternative parameterisation, particularly over the γ parameter to see if a higher value, as used within other Temporal Difference learning work, could extend the length of action chain over which accurate payoff-prediction maps can be established. It is likely that some upper value of γ will be reached at which point generalisation will combine early states due to their prediction similarity. Further work is also required to establish limits in other forms of environment. Although the environment used was designed to limit the complexities of exploration, an environment that provides a reward of zero for non-optimal actions may provide a clearer distinction between the predictions and thus allow XCS to identify longer chains of optimal actions. Alternatively, an environment with a sub-optimal action

that leads back to the same state may cause unequal exploration of the environment, thus preventing progress towards the terminal state and further hindering the establishment of action chains. Finally, the complexity of exploration was deliberately controlled, and further work is required to establish the limits of action chain length in the face of increasing numbers of alternative routes, and the ability of XCS to choose between alternative reward magnitudes as the number of alternative action routes increases.

Whilst the problem of over-generalisation over the early states has been revealed by this work, no particular solution has been proposed. This was beyond the scope of the work, but further work [15] has shown that new techniques can be introduced to XCS that would allow XCS to be applied within longer rule chains. Unfortunately all of these techniques introduce substantial new features to XCS and seek to avoid or postpone rather than tackle the cause of the over-generalisation. In order to tackle the cause, the true cause has to be identified. XCS has been criticised for its dependence upon fixed rather than relative measures of error [15]. As the payoff values become small towards the start of a long action chain, so the more meaningless the error calculation and minimum error cut-off parameters become to these early classifiers. Fluctuations in later reward will be smoothed and not detectable in the earlier classifiers. At the same time a large variation in payoff because of the introduction of a new classifier through the G.A. could cause early classifiers to become inaccurate when the same introduction later in the action chain would have negligible effect on the existing classifiers at that point in the action chain. These problems cannot be readily solved by a simple modification to a parameter. For example, reducing the minimum error parameter may help the earlier action sets distinguish between accurate and inaccurate classifiers but would also make it much more difficult for the later action sets to find accurate classifiers. It would be better, instead, to re-examine the calculation of error as a fixed proportion of the magnitude of the reward and look for methods of calculating error as the difference between the predicted and actual payoff as a proportion of the predicted payoff. This is a fertile area for further study.

The Domination Hypothesis was introduced to explain the appearance of fully general classifiers dominating the population in the longer FSW under generalisation pressure. Although a preliminary investigation of the hypothesis was presented, there is much scope for further investigation of the causes of these classifiers and potential solutions to the problem of dominant fully general classifiers. A naive solution that prevents the use of fully general classifiers (since, by definition, they can never be accurate) is not adequate. Fully general classifiers can act as a form of advanced Covering operator, distributing actions to newly discovered environmental niches. It may be that preventing the formation of fully general classifiers whilst encouraging action sharing covering would enable XCS to establish more accurate payoff predictions within multiple step environments similar to the test environments used within this research work.

References

[1] Wilson, S.W. (1995), Classifier Fitness Based on Accuracy, *Evolutionary Computation*, 3(2), 149-175.

[2] Wilson, S.W. (1998) , Generalisation in the XCS classifier system, in Koza, J. R., Banzhaf, W., Chellapilla, K., Deb, K., Dorigo, M., Fogel, D. B., Garzon, M.H., Goldberg, D.E., Iba, H., Riolo, R. (eds.), *Genetic Programming 1998: Proceedings of the 3rd Annual Genetic Programming Conference (GP98)*, Morgan Kaufmann, San Francisco, CA.

[3] Lanzi, P-L, Riolo, R. (2000), A Roadmap to the Last Decade of Learning Classifier System Research, in Lanzi, P. L., Stolzmann, W., Wilson, S. W. (eds.), *Learning Classifier Systems, From Foundations to Applications*, 33-62, Springer-Verlag.

[4] Kovacs, T. (1996) , *Evolving Optimal Populations with XCS Classifier Systems*, Master's Thesis, School of Computer Science, University of Birmingham.

[5] Lanzi, P-L. (1997) ,A Study of the Generalisation Capabilities of XCS, in Back, T. (ed.), *Proceedings of the 7th International Conference on Genetic Algorithms (ICGA97)*, 418-425, Morgan Kaufmann, San Francisco, July 1997

[6] Lanzi, P-L. (1997)), *A Model of the Environment to Avoid Local Learning (An Analysis of the Generalisation Mechanism of XCS)*, Technical Report 97.46, Politecnico di Milano, Department of Electronic Engineering and Information Sciences, 1997.

[7] Cliff, D., Ross. S. (1994) , Adding Temporary Memory to ZCS, *Adaptive Behaviour*, 3(2), 101-150.

[8] Greffenstette, J. (1987) , Multilevel Credit Assignment in a Genetic Learning System, in Grefenstette, J. J. (ed.), *Proceedings of the Second International Conference on Genetic Algorithms (ICGA87)*, Cambridge MA, July 1987, Lawrence Erlbaum Associates, 202-207.

[9] Barry, A.M. (1999) , Aliasing in XCS and the Consecutive State Problem 1 - Problems, in Banzhaf, W., Daida, J., Eiben, A. E., Garzon, M. H., Honavar, V., Jakiela, M., Smith, R. E. (eds.), *Proceedings of the First Genetic and Evolutionary Computation Conference (GECCO99)*, Morgan Kaufmann, San Francisco, CA, 19-26.

[10] Lanzi, P-L. (1998) , Adding Memory to XCS, in Proceedings of the IEEE Conference on Evolutionary Computation (ICEC98), IEEE Press.

[11] Lanzi, P-L. (1998) An analysis of the memory mechanism of XCSM, in Koza, J. R., Banzhauf, W., Chellapilla, K., Deb, K., Dorigo, M., Fogel, D. B., Garzon, M. H., Goldberg, D. E., Iba, H., Riolo, R. L., (eds.), *Genetic Programming 1998*: Proceedings of the Third Annual Conference, Morgan Kaufmann: San Francisco, CA, 643-651.

[12] Lanzi, P-L. (1998) *Reinforcement Learning by Learning Classifier Systems*, PhD Thesis, Politecnico di Milano, 1998

[13] Lanzi, P-L., Wilson, S.W. (1999), *Optimal classifier system performance in non-Markovian environments*, Technical Report 99.36, Dipartimento di Elettronica e Informazione - Politecnico do Milano, 1999.

[14] Butz, M., Wilson, S.W. (2000), *An Algorithmic description of XCS*, Technical Report 200017, IlliGAL, University of Illinois.

[15] Barry, A.M. (2000) , *XCS Performance and Population Structure in Multi-Step Environments*, Ph.D. Thesis, Queens University Belfast, September 2000.

[16] Riolo, R. (1987) , Bucket Brigade Performance: I. Long Sequences of Classifiers, in Grefenstette, J.J. (ed.), *Proceedings of the Second International Conference on Genetic Algorithms (ICGA87)*, Cambridge, MA, July 1987, Lawrence Erlbaum Associates.

[17] Riolo, R. (1988) , *Empirical Studies of Default Hierarchies and Sequences of Rules in Learning Classifier Systems*, PhD Thesis, University of Michigan.

[18] Riolo, R. (1989) , The Emergence of Coupled Sequences of Classifiers, in Schaffer, J.D. (ed.), *Proceedings of the Third International Conference on Genetic Algorithms (ICGA89)*, 256-264, George Mason University, June 1989, Morgan Kaufmann.

[19] Saxon, S., Barry, A.M. (2000) , XCS and the Monk's Problems, in Lanzi, P. L., Stolzmann, W., Wilson, S. W. (eds.), *Learning Classifier Systems, From Foundations to Applications*, Springer-Verlag.

[20] Wilson, S.W. (1994), ZCS: A zeroth level classifier system, *Evolutionary Computation*, 2(1), 1-18.

[21] Holland, J. H. (1986) , Escaping Brittleness: The possibilities of General-purpose Learning Algorithms Applied to Parallel Rule-Based Systems, in Mitchell, T.M., Michalski, R. S., and Carbonell, J.G. (eds.), *Machine Learning, An Artificial Intelligence Approach*, Vol. II, ch. 20, 593-623, Morgan Kaufmann.