Adaptation and Genomic Evolution in EcoSim

Marwa Khater and Robin Gras

School of Computer Science, University of Windsor
ON, Canada
{khater,rgras}@uwindsor.ca

Abstract. Artificial life evolutionary systems facilitate addressing lots of fundamental questions in evolutionary genetics. Behavioral adaptation requires long term evolution with continuous emergence of new traits, governed by natural selection. We model organism's genomes coding for their behavioral model and represented by fuzzy cognitive maps (FCM), in an individual-based evolutionary ecosystem simulation (EcoSim). Our system allows the emergence of new traits and disappearing of others, throughout a course of evolution. In this paper we show how continuous adaptation to a changing environment affects genomic structure and genetic diversity. We adopted the notion of Shannon entropy as a measure of genetic diversity. We emphasized the difference in genetic diversity between EcoSim and its neutral model (a partially randomized version of EcoSim). In addition, we studied the effect that genetic diversity has on species fitness and we showed how they correlate with each other. We used Random Forest to build a classifier to further validate our findings, along with some meaningful rule extraction.

Keywords: artificial life modeling, individual-based modeling, genetic diversity, entropy, fitness.

1 Introduction

Charles Darwin's theory of adaptation through natural selection came to be widely seen as the primary explanation of the process of evolution and forms the basis of modern evolutionary theory. Darwin's principle of natural selection relies on a number of propositions. The individuals in a population are not identical but vary in certain traits. This variation, at least partly, is heritable. Individuals vary in the number and the quality of their descendants, depending on the interactions of the individual's trait and its environment. Populations with these characteristics may become more adapted to their environment over generations. The key to adaptation by natural selection is the effect of a multitude of small but cumulative changes. While most of these changes are random, the majority of those that are preserved are not damaging to the fitness of individuals. Instead these variations may turn out to be somehow beneficial to the reproductive success of their carrier. From a genetic perspective, the combination of mutation and natural selection, enforce the emergence of new traits and disappearance of others. These continuous genetic changes help preserve genetic diversity.

T. Ziemke, C. Balkenius, and J. Hallam (Eds.): SAB 2012, LNAI 7426, pp. 219–229, 2012.

Genetic diversity serves as a way for populations to adapt to changing environments. With more variations, it is more likely that some individuals in a population will possess variations of alleles that are suited for their current environment. Those individuals are more likely to survive to produce offspring bearing those alleles. These alleles will propagate through the population over many generations because of the success of these individuals. In summary, genetic diversity strengthens a population by increasing the likelihood that at least some of the individuals will be able to survive major disturbances. Many biological studies showed that a decrease in population genetic diversity can be associated with a decline in population fitness [1] [2] [3]. Because overall population diversity seems to affect both short-term individual fitness and long-term population adaptive capacity, there is a need to develop an empirical quantitative understanding of the relationship between population genetic diversity and population viability.

Like in many disciplines; simulation modeling played a great role in studying evolutionary processes. Many biological studies that require data of hundreds of years can be obtained by simulation modeling that produces results in a matter of a few hours or days depending on the computational cost of each system. In this paper we show how individuals in EcoSim [4], an evolutionary predator-prey ecosystem simulation, follow the Darwinian evolutionary process through natural selection. We show how genetic evolution and diversity governs the adaptation process. We show that EcoSim's individuals adapt to their changing environment by comparing their behavior with a neutral model - a partially randomized version of EcoSim. We use the Shannon entropy, which is a measure of unpredictability and disorder from Information theory, as a measure of genetic diversity and present the difference in entropy between EcoSim and the neutral model to emphasize the adaptive characteristics of EcoSim. Furthermore, we investigate the relationship between genetic diversity and species fitness and present the correlations found between these two measures in EcoSim. The rest of the paper is organized as follows: A brief description of EcoSim and its neutral model is presented in Section 2 and 3 respectively. Section 4 depicts the details of the entropy as a genetic diversity measure followed by a comparison between EcoSim and its neutral model, in terms of entropy as a genetic diversity, in 5. The correlation results between entropy and fitness are presented in Section 6, followed by building a classifier for inference in Section 7. A summed up conclusion is presented in Section 8.

2 The EcoSim Model

In order to investigate several open theoretical ecology questions we have designed the individual-based evolving predator-prey ecosystem simulation platform EcoSim, introduced by Gras et al. [4] [1]. Our objective is to study how individual and local events can affect high level mechanisms such as community formation, speciation and evolution. EcoSim uses Fuzzy Cognitive Map as a behavior model [5] which allows a combination of compactness with a very low

[1] `http://sites.google.com/site/ecosimgroup/research/ecosystem-simulation`

computational requirement while having the capacity to represent complex high level notions. The complex adaptive agents (or individuals) of this simulation are either prey or predators which act in a dynamic 2D environment of 1000 x 1000 cells. Each individual possesses several physical characteristics including age, minimum age for breeding, speed, vision distance, levels of energy, and the amount of energy transmitted to the offspring. Preys consume grass, and predators predate on prey individuals. Grass distribution is dynamic, as it diffuses in the world and disappears when consumed by preys. An individual consumes some energy each time it performs an action such as evasion, search for food, eating or breeding. Each individual performs one action during a time step based on its perception of the environment. Fuzzy Cognitive Map (FCM) [5] is used to model the individual's behavior and to compute the next action to be performed. The individual's FCM is coded in its genome and therefore subjected to evolution. A typical run lasts tens of thousands of time steps, during which more than a billion of agents are born and several thousands of species are generated, allowing evolutionary processes to take place and new behaviors to emerge to adapt to a constantly changing environment. Our simulation embodies species as a set of individuals sharing similar genomes [6]. Indeed, every member of a species has a genome that is within a threshold distance away from the species genome - an average of the FCMs of its members. To model the process of speciation, EcoSim allows splitting of a species into two sister species. The splitting mechanism produces two clusters of individuals with high intra-cluster similarity and strong inter-cluster dissimilarity. It is worth noting that the speciation mechanism is only a labeling process: the information about species membership is not used for any purpose during the simulation but only for post-processing analysis of the results.

Formally an FCM is a graph which contains a set of nodes, each node being a concept, and a set of edges, each edge representing the influence of one concept on another. In each FCM, three kinds of concepts are defined: sensitive (such as distance to foe or food, amount of energy, etc), internal (fear, hunger, curiosity, satisfaction, etc) and motor (evasion, socialization, exploration, breeding, etc.). We use a FCM to model an agent's behavior (structure of the graph) and to compute the next action of the agent (i.e. through the dynamics of the map). The FCM serves as a genome for each individual. The genome length is maximum 390 sites, where each site corresponds to an edge between two concepts of the FCM. The FCM allows the formation of new edges and disappearing of others through the evolutionary process. During a breeding event the FCMs of two parents are combined and transmitted to their offspring after the possible addition of some mutations which is similar to the genetic process of recombination. The behavior model of each individual is therefore unique.

3 The Neutral Model

In order to study the effect of adaptation on evolution, we built a neutral shadow [7] of EcoSim. All selection processes and behaviors in the neutral shadow for the predator/prey are random, which eliminates natural selection from this model.

In terms of the behavioral model of this version, all the actions such as eating, hunting (for predators), socializing, searching for food and escaping (for prey) are removed. The only two actions any individual can take are reproduction and movement. Unlike in the EcoSim, in the neutral model there is no necessity for the individuals to have genetic similarity to reproduce. Instead, in the neutral model the reproduction action is done by randomly choosing any 2 individuals in the world. The statistics of genetic operations (mutation rates and crossover) are the same as EcoSim. In EcoSim, individuals choose to reproduce according to their internal state, suitable environmental conditions and behavior model but not in the neutral model. To preserve population dynamics in neutral model similar to that of EcoSim, the Lotka-Volterra computational model [8] is used. This model controls the number of births and deaths at each time step. In addition, death of individuals and pairs of parents for reproduction are randomly selected. In this way a similarity in population sizes between the neutral shadow and EcoSim is preserved. Finally, the movements in the neutral model are random, but the distribution of distances is kept the same as in EcoSim.

The crucial property of EcoSim neutral shadow is that its evolutionary dynamics are identical to EcoSim except that neither the presence not the frequency of a genotype can be explained by its adaptive significance. This is because all selection in the neutral model is random, so no genotype has any dominance over any other. In other words, although gene states are subject to the same variation as in EcoSim, they have no evolutionary fitness consequences or effects. In addition, changes in the environment have no effect on individuals in the neutral model. Consequently, the process of natural selection is considered to be eliminated in this neutral model.

4 Entropy as a Measure of Genetic Diversity

Depending on the specific problem or representation being used, ranging from biological domain to genetic programming, numerous diversity measures and methods exist. Foe example, Sherwin [9] has shown the efficancy of Shannon entropy in measuring diversity in ecological community and genetics. He has also highlighted the advantages of using entropy based genetic diversity measures, and surveyed these diversity measures. A close relationship between biological concepts of Darwinian fitness and information-theoretic measures such as Shannon entropy or mutual information, was found [10]. Shannon Information theory [11] defines uncertainty (entropy) as the number of bits needed to fully specify a situation, given a set of probabilities. These probabilities can be estimated by simply counting the abundance of each genotype (site) in the population. The per-site entropy of an ensemble of sequences X, in which genotype s_i occurs with probability p_i is calculated as

$$H(X) = -\Sigma p_i \log_2(p_i) \tag{1}$$

where the sum goes over all different genotypes i in X. Next, the entropy content of the whole sequence (genome) is approximated by summing the per-site

entropy over all sites in the sequence. This is only an approximation because it ignores interactions between sites (i.e. epistasis). We do not have a fixed set of alleles but they are discreet values that change over time in the simulation. The lower the entropy, the less diverse are the genomes of a population and vice versa. There is a limit in the desired values of entropy in EcoSim. When it approaches its maximum(corresponding to an uniform distribution of all genotypes) it indicates a completely uniform population close to randomness. On the other hand very low entropy (close to 0) means that there is too much similarity between individual genomes, and means that individulas need to diverge more in order to adapt to a dynamic environment. A good balance between learning from the environment (low genetic diversity) and increasing the diversity (high genetic diversity) should be met in order to ensure the well being of species.

5 Evolution in EcoSim verses Neutral Model

The FCM of each individual plays the role of its genome and has a maximum size of 390 sites. Every site is a real discreet number which measures the level of influence from one concept to another. Initially all prey and predator individuals are given the same values for their genome respectively. Time step after another, as more individuals are created, changes in the FCM occur due to the formation of new edges, removal of existing ones and changes in the weight associated to edges. We neglect the first couple of thousand of time steps in our calculations to overcome any misleading results due to the initial similarity between individual genomes. In each time step we have a value of entropy of all existing preys species, along with the entropy of the entire population of prey. We also calculated the fitness for every species as the average fitness of its individuals. We define fitness of an individual as the age of death of the individual plus the sum of the age of death of its entire direct offspring. Accordingly, the fitness value mirrors the individual's capability to survive longer and produce high number of strong adaptive offspring.

The information contained within a genome determines how the organism behaves in its current environment. Thus, this information determines the capability of the organism to reproduce and transmit its genome. The environment changes from one place to another and from one time step to the next. Individuals that evolve in different parts of the world have different information about the environment they evolve in stored in their genome. Furthermore, as we model a predator-prey system, we also have co-evolution. The strategies (behavior) of each kind of individula(predator/prey) are continuously changing as they try to adapt to the other kind. The more the individuals try to learn the more the environment changes and the more there is still something different to learn. This fact drives the individuals to keep learning and continuously try to come up with survival strategies that helps them adapt to their changing environment. This is the reason behind the fluctuations we see in the EcoSim entropy curves see Fig.1. On the other hand the neutral model shows much more steadiness in the entropy values. Under highly random conditions and when natural selection

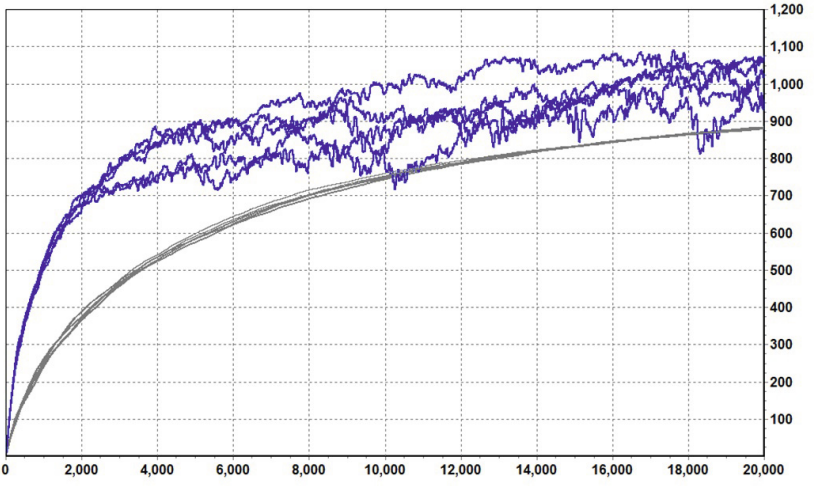

Fig. 1. Global Entropy for 10 different runs of the simulation. Top 5 curves are for EcoSim and lower 5 for Neutral Model.

is eliminated, the genomic structure shows neither learning nor adaptation to the surrounding environment. These results show that entropy changes through the course of evolution. The EcoSim simulation gave us the chance to acquire data for thousands of generations and to study the performance of entropy as a genetic diversity measure.

6 Measuring Correlation between Entropy and Fitness

In order to further emphasize the importance of genetic diversity to adaptation and thus the well being of individuals, we were encouraged to study the effect that genetic diversity has on fitness. EcoSim gives us the chance to study the relation between species genetic diversity and species fitness without the limits in environmental conditions and time scales found in biological studies [2] [3] [12], but in highly variable environments and across evolutionary time. There are many factors affecting genetic diversity and fitness, and the correlation between them. At every time step we calculate the entropy and the fitness for all existing species. In order to investigate their possible correlations, we first begin by calculating the Spearman's cross correlation [13], between entropy and fitness of all prey species. A perfect Spearman correlation of +1 or -1 is attained when each of the variables is a perfect monotone function of the other; a value close to zero means that there is no correlation.

In our evolutionary ecosystem the effect of entropy on fitness is not immediate. A time shift between the variation in entropy and its effect on fitness is therefore expected. Also, because we did not determine which attribute is the cause of the other we calculate the correlation in both shift directions. We computed the Spearman correlation coefficient, between these two time series for every possible

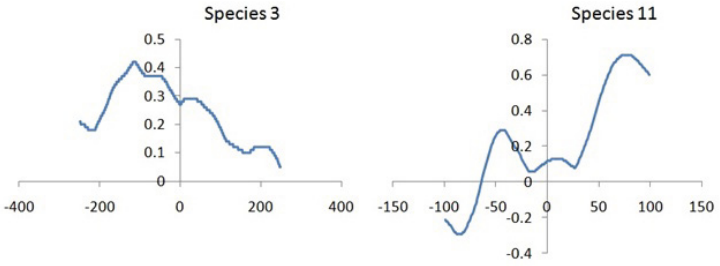

Fig. 2. Different species correlation values between entropy and fitness. x-axis represents the different time shifts. Y-axis represents the correlation values.

shift between -s and +s time steps. Thus, we correlated the entropy at time t with fitness at time t + S where S ranges from -s to +s. We've presented the cross-correlation charts for some prey species in Fig.2. The x-axis in these charts represents the different shifts for the time series. The y-axis represents the cross-correlation value at the corresponding shift. From the figure we see that not only different species have different cross-correlation values, but also the same species correlated differently based on the time shift.

It should be noted that the dynamic environment, co-evolution and changing parameters with time, all affect species behavior. Thus, correlation values for the same species might vary through the course of evolution. This presents a feasible problem to study in our model but not in biological experiments. This fact encouraged us to add a time frame to the two series and measure correlation within the specific time frame. Consequently, we split these time series into sliding windows. Within each window we calculated all possible correlations with different shifts ±s. Then we chose the highest correlation value (whether positive or negative) and assigned it to the corresponding species instance of the time series.

In order to examine the possible correlation values between species entropy and fitness at every time step we usde data collected from 5 different runs of the EcoSim simulation, each running with the same initial conditions. Each replicate ran for 16,000 time steps and generated 110,000 instances (an instance corresponding to one given species at give one time step) in average. For each instance we calculated the Spearman's cross correlation between entropy and fitness for the corresponding window. We assigned three different classes to the correlation values. Correlations with values between -0.5 and 0.5 are class WEAK CORR which shows the situation where there is either no or weak correlation. Correlation values above 0.5 are high positive (HIGHP) and correlation values below -0.5 are high negative (HIGHN) respectively. Different shift values and sliding windows ranges have been examined and previously presented in [14]. We choose±25 as a shift value based on the analysis of which shift leads to the highest correlations and a sliding windows of 200. In an average of 5 different runs of the simulation 26.8% of instances had HIGHP correlation, 38.4% of instances had HIGHN correlation and 34.7% of instances had WEEK correlation.

Although there are many factors that might affect fitness besides entropy, we managed to find strong correlation between entropy and fitness for all prey species. We observed high values for both negative and positive correlations. These results support the claim of the great influence the genetic diversity has on the well being of species. High positive correlation values mean that an increase in the genetic diversity, results in an increase in species fitness. However, there are many ways to interpret these results. A newly forming species with a small population would gradually tend to increase its genetic diversity and will therefore correlate positively with fitness. Also, since individuals in EcoSim adapt to a constantly changing environment these adaptations could be mirrored in the increase of individuals' genome similarity (and thus a decrease in entropy), as new behaviors diffuse in the population. Conversely negative correlations imply that a species decreases diversity in order to reach stability by learning from its environment and adapting.

7 Building a Random Forest Classifier for Inference and Rule Learning

Our motivation to validate these results and further investigate the reason behind these correlation values encouraged us to build a classifier. The purpose of building this classifier is first to see if some specific species properties can predict the current evolutionary behavior of the species, that is if it is learning from the environment or increasing its diversity to be able to react to a future change in the environment. It can also help to understand what factors and conditions affect the evolutionary of behavior. The Random Forest [15] technique includes an ensemble of decision trees and incorporates feature selection and interactions within the learning process. It is nonparametric, efficient, and has high prediction accuracy for many types of data including high dimensional ones. We chose features from both individual's internal and physical concepts, such as average energy level, reproduction rate, population size, speed of individuals, spatial dispersal and others, to predict the class correlation value between genetic diversity and fitness. All together we chose 15 features that best described internal and physical properties of any species and verified if they could predict the class correlation variable. To increase the quality of the classifier we used feature selection [16] in order to extract the most important features from the above list. This step provided more semantics about which features most influence the value of correlation. The best chosen features were population size, entropy, fitness, spatial dispersal, average age of the individuals and number of failed reproductions. We used the Random Forest classifier implemented in the weka environment [17]. We split instances for every run into two sets: train and test and used 10 fold cross validation. The average of 5 classifiers testing accuracies representing 5 different simulation runs was 96.7%. The high classification accuracy validates our use of entropy as a measure of genetic diversity and its high correlation with fitness. It also shows that there exist specific conditions of the species that lead to a positive or negative correlation between fitness and genetic diversity.

The model generated by the Random Forest can be challenging to interpret. To by-pass this limitation we use the JRip rule learner [18] to extract more semantics from the prediction model and gain more insight about the conditions affecting correlation between genetic diversity and fitness. JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is an optimized version of IREP. Different IF THEN rules are learned from JRip to predict the three correlation classes. In 5 different runs 19 rules were discovered in average with average accuracy of 76% using 10 fold cross validation see Table 1. We were mainly interested in studying the rules that predict the HIGHP and HIGHN classes and we present some of these rules having the highest number of instances.

Table 1. JRip rule learner accuracies and number of produced rules for five different runs of the simulation

Run	Train accuracy	Test accuracy	No. of rules
Run 1	76%	75.6%	24
Run 2	71.7%	72%	23
Run 3	75%	75.8%	24
Run 4	79%	78.8%	7
Run 5	75.5%	76.1%	18
Average	75.4%	76%	19

- IF number of individuals is low, AND fitness is low, AND entropy is low, AND failed reproduction is high THEN correlation is HIGHP.
- IF number of individulas is low, AND age is high, AND fitness is low, AND entropy is low THEN correlation is HIGHP.
- IF fitness is low, AND age is medium, AND spatial dispersal is low THEN correlation is HIGHP.
- IF number of individuals is high, AND age is high, AND entropy is high, AND spatial dispersal is high THEN correlation is HIGHN.
- IF spatial dispersal is high, AND number of individulas is high, AND age is medium, AND entropy is medium, AND fitness is high THEN correlation is HIGHN.
- IF failed reproduction is low, AND entropy is high, AND number of individuals is high THEN correlation is HIGHN.

The other discovered rules were also similar. In general, we found that a low number of individuals associated with a low entropy, low fitness and low spatial dispersal led to a high positive correlation between entropy and fitness. Small

species tended to increase their genetic diversity in order to increase their fitness. On the other hand, a high number of individuals associated with a high entropy, high fitness and high average age led to high negative correlations between entropy and fitness. Large species in terms of population size tended to move towards lower genetic diversity as individuals learned common survival strategies that tended to increase their fitness.

8 Conclusion

We showed how the evolutionary process implemented in EcoSim affects the behavioral model of the individuals as they adapt to a changing environment. To emphasize the capability of EcoSim to model evolutionary behavioral adaptation we compared it to a partially random version focusing on genetic diversity. We showed how entropy used to measure genetic diversity, behaves differently in both systems. The fluctuation in entropy curves for EcoSim showed how individuals try to learn and adapt to their environment. On the other hand the neutral model showed more steadiness in the curves due to more randomness and elimination of natural selection process. Furthermore, we presented high correlation values between species fitness and genetic diversity which strongly indicates how genetic diversity affects the well being of the species. A validation step was performed with the use of machine learning techniques. A random forest classifier was built to predict the correlation values based on internal and physical properties of species used as features. The rules discovered from the rule learner, which seem to be biologically pertinent, gave us more understanding about the conditions affecting the values of correlation between genetic diversity and fitness.

Acknowledgments. This work is supported by the NSERC grant ORGPIN 341854, the CRC grant 950-2-3617 and the CFI grant 203617 and is made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca).

References

1. Reed, D., Frankham, R.: Correlation between fitness and genetic diversity. Conserv. Biology 17, 230–237 (2003)
2. Markert, J., Champlin, D., Gutjahr-Gobell, R., Grear, J., Kuhn, A., McGreevy, T., Roth, A., Bagley, M., Nacci, D.: Population genetic diversity and fitness in multiple environments. BMC Evolutionary Biology 10 (2010) 1471–2148–10–205
3. Vandewoestijne, S., Schtickzelle, N., Baguette, M.: Positive correlation between genetic diversity and fitness in a large, well-connected metapopulation. BMC Biology 6 (2008) 1741–7007–6–46
4. Gras, R., Devaurs, D., Wozniak, A., Aspinall, A.: An individual-based evolving predator-prey ecosystem simulation using fuzzy cognitive map as behavior model. Artificial Life 15(4), 423–463 (2009)
5. Kosko, B.: Fuzzy cognitive maps. Int. Jornal of Man-Machine Studies, 65–75 (1986)

6. Aspinall, A., Gras, R.: K-Means Clustering as a Speciation Mechanism within an Individual-Based Evolving Predator-Prey Ecosystem Simulation. In: An, A., Lingras, P., Petty, S., Huang, R. (eds.) AMT 2010. LNCS, vol. 6335, pp. 318–329. Springer, Heidelberg (2010)
7. Bedau, M.A., Snyder, E., Packard, N.H.: A classification of longterm evolutionary dynamics. In: Proc. of Art. Life VI, pp. 228–237. MIT Press (1998)
8. Volterra, V.: Variations and fluctuations of the number of individulas in animal species living together. Animal Ecology 3, 409–448 (1931)
9. Sherwin, W.B.: Entropy and information approaches to genetic diversity and its expression: Genomic geography. Entropy 12, 1765–1798 (2010)
10. Bergstrom, C., Lachmann, M.: Shannon information and biological fitness. In: IEEE Information Theory Workshop, pp. 50–54 (2004)
11. Shannon, C.: A mathematical theory of communication. Bell Systems Technical Journal, 379–423 (1948)
12. Oostermeijer, J., van Eijck, M., den Nijs, J.: Offspring fitness in relation to population size and genetic variation in the rare perennial plant species gentiana pneumonanthe (gentianceae). Oecologia 97, 289–296 (1994)
13. Siegel, S.: Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York (1956)
14. Khater, M., Salehi, E., Gras, R.: Correlation between Genetic Diversity and Fitness in a Predator-Prey Ecosystem Simulation. In: Wang, D., Reynolds, M. (eds.) AI 2011. LNCS, vol. 7106, pp. 422–431. Springer, Heidelberg (2011)
15. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
16. Yang, Q., Salehi, E., Gras, R.: Using Feature Selection Approaches to Find the Dependent Features. In: Rutkowski, L., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2010. LNCS (LNAI), vol. 6113, pp. 487–494. Springer, Heidelberg (2010)
17. Witten, I., Frank, E.: Data Mining- Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, USA (2000)
18. Cohen, W.: Fast effective rule induction. In: 12th International Conference on Machine Learning, pp. 115–123 (1995)