Investigating the Effect of Spatial Distribution and Spatiotemporal Information on Speciation using Individual-Based Ecosystem Simulation

Morteza Mashayekhi, Robin Gras

***Abstract*— In this paper, we investigate the impact of species' spatial and spatiotemporal distribution information on speciation, using an individual-based ecosystem simulation (Ecosim). For this purpose, using machine learning techniques, we try to predict if one species will split in near future. Because of the imbalanced nature of our dataset we use smote algorithm to make a relatively balanced dataset to avoid dismissing the minor class samples. Experimental results show very good predictions for the test set generated from the same run as the learning set. It also shows good results on test sets generated from different runs of Ecosim. We also observe superior results when we use, for the learning set, a run with more species compare to a run with less species. Finally we can conclude that spatial and spatiotemporal information are very effective in predicting speciation.**

***Index Terms*— smote, spatial distribution, spatiotemporal information, speciation, speciation prediction**

I. Introduction

THERE are more than twenty definitions for species concept in literature [1] however the most commonly used by most biologists is a group of organisms that are able to exchange genes within themselves but are reproductively isolated from other such groups. It means that there is no gene flow between two of such community [2]. They have separate ancestor-descendent tree of life with different tendencies and evolutionary path.

Speciation is the division of one single species into two or more genetically distinct ones. This process extends through time and leads to a hierarchal tree of historical relationship between species. Two steps are entailed in speciation: [3] a new population should be established which could be in the same habitat or completely separated of the main population depending on speciation mechanism; then a reproductive isolation should occur, due to different habitats, physical barrier, etc., to reduce or prevent gene flow between organisms of the different species. The spatial distribution of individuals in one species can act as an isolator and be a leading phenomenon for speciation [3], [4], [5]. For example, in [6] it has been proved that there is a linear relationship between genetic and geographic distance. It means that increasing the physical distance between individuals increases the probability of speciation.

Because speciation is a continuous ongoing process, the current spatial distribution of a species is not necessarily a reliable index of the species' historical distribution during its life time. Losos et al. mentioned three evidences in [7] showing that the present spatial distribution of a species is greatly different from the one at its creation time. Therefore, observing species during its whole life time is also important to understand and eventually predict speciation.

Observing and studying species in nature to extract their spatial distribution information is a highly difficult and time consuming process. For this reason using computer science techniques to simulate such a system is a good alternative solution. One special type of such simulator is individual-based simulation [8] in which individual specificities affecting the overall system are modeled. In this paper we use Ecosim[1] [9], developed by Gras et al., which is an individual-based evolving predator-prey ecosystem simulation. In this program two organism types, prey and predator, are simulated in a torus like world which is a 1000×1000 matrix of cells. Every cell can contain some amount of grass and meat which are food for prey and predator respectively. Each individual, based on its type, is able to perform some actions. For example prey can move, eat grass, escape from predator, mate with other prey if they are genetically similar enough and produce an offspring with a modified combination of its parent's genome, etc. Predators can move, eat meat, hunt prey, mate with other predator, etc. All individuals act based on their behavioral model implemented by a fuzzy cognitive map (FCM). A FCM is a weighted graph, each node being a sensitive(such as the distance to food, to a friend or to a sexual partner), internal(such as fear, hunger or satisfaction) or action (like escape, eat or reproduce) concept in our case, and each edge is a level of influence between two concepts. The FCM is coded inside the genome of each individual and therefore

Manuscript received February 22, 2012. This work was supported by the NSERC grant ORGPIN 341854, the CRC grant 950-2-3617 and the CFI grant 203617.

M. Mashayekhi is a PhD student in the School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4 Canada (phone: 519-253-3000 Ext. 3003; e-mail: mashaye@uwindsor.ca).

R. Gras, is an associate professor in the School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4 Canada (e-mail: rgras@uwindsor.ca).

[1] More information about Ecosim can be find at http://sites.google.com/site/Ecosimgroup/research/ecosystem-simulation

DOI: 10.5176_2010-2283_2.1.135

subject to the evolutionary process. As a consequence, every individual as a unique behavioral model inherits from its parents. In this simulation species concept is also represented as a set of individuals having a high level of genetic similarity. Depending on the evolution of the system, speciation or extinction may happen for every species at any time step. Speciation is done by a 2-means clustering algorithm presented in [10]. Initially the system starts with one species of prey and one species of predator and due to the evolution of the individual's population, when the maximum genetic distance among two individuals of a species becomes greater than a predefined threshold, two new species emerged. Information about all individuals and the world is saved for each time step. Several studies have been done using Ecosim. For example in [11] Devaurs et al. have shown that the behavior of the simulation is realistic by comparing the species abundance pattern in the simulation environment with real communities of species. Also, the chaotic behavior of the system with multi-fractal properties has been proven in [12], [13] as it also has been observed for real ecosystems.

Although there are many factors involved in speciation, in this paper we want to answer to the questions such as how spatial and spatiotemporal patterns influence speciation? Which metrics are important and in what extent? For answering such questions, we have applied machine learning techniques on the data generated by Ecosim to evaluate if spatial distribution and spatiotemporal information of species can predict splitting of species. If we could predict speciation by using this information, it means that they have impact on speciation. We are also interested to extract predictive rules on speciation based on spatial and spatiotemporal information that could help to understand this complex phenomenon.

Subsequent to this introduction, we explain the dataset preparation phase in section II. In section III the learning algorithm and evaluation metrics are described. Section IV discusses experimental results and shows the speciation prediction to see if one species split in next 100 time steps. Finally section V is the conclusion.

II. Preparing Data

The information about all the objects in the world i.e. species, individuals and food in each time step is stored separately. Therefore we have a huge amount of information and for this paper we just extract spatial distribution and spatiotemporal information for every species.

A. *Spatial Distribution Information*

In individual-based simulation, we have access to all the information for each individual. So it is possible to specify the location of each individual at any time step in a 3-dimensional vector with two spatial and one temporal dimension. In Ecosim, the world is a torus which can be easily implemented by a rectangular array and allows the individuals to pass across one boundary and enter the opposite boundary. Based on the circular condition of the world, applying traditional statistics is not possible, so we use circular statistics to calculate a circular center of the spatial distribution of a species as defined in [14]. Afterward we calculate the average and the standard deviation of the Euclidian distance of all the individuals to the center of their species.

B. *Spatiotemporal Metrics*

In this part we use some spatiotemporal metrics described in [15] and adapt them to EcoSim concepts in order to have some historical information about species spatial distribution. These metrics are used to characterize the complex spatiotemporal dynamics of ecological mosaics or categorical maps. This characterization is based on analysis of space-time cubes of data with two spatial dimensions *x*, *y* and one time dimension *t* and we call it 3D world. This cube includes successive spatial information of the environment sampled at uniform time intervals. Each spatial image in 3D world is a grid of cells or pixels like the cells in the world in Ecosim. By adding temporal dimension, each spatial pixel becomes a 3-dimensional voxel having two spatial and a temporal dimension with $t=1$. Persistent entities, like prey in our simulation, occupied 3-dimensional forms consisting of several adjacent voxels in space and time dimensions are called blob. In the 3D world we may have different kind of blob types. For example in the world of Ecosim each species is considered as a blob type. Moreover, each voxel in the 3D world may belong to different blob types because each cell in EcoSim can contain multiple individuals from different species. In addition, each blob type may be composed of multiple separated blobs in 3D world. For example one species blob type may be consisting of four subpopulation blobs in the 3D world like the dotted pattern blobs in Fig. 1.

There are two 3D metrics categories for analyzing blobs: composition and configuration metrics. Volume, surface area, shape complexity and fractal dimension are examples of composition metrics. A blob volume is the number of voxels it occupies. Surface area is the number of voxels in a blob with faces not shared by adjacent voxels of the same blob type. For calculating adjacency, we used 6-voxel vonNeuman neighbors (Fig. 1). Shape complexity is a ratio between blob volume and volume of its bounding box. For example if the dotted line cube volume is 4 in Fig. 1 then the shape complexity of the wavy format blob type would be 0.5.

Fractal dimension is used to quantitatively describe how one object occupies its volume. We used count boxing method [16] to calculate fractional dimension for each species. Moreover, we calculate some other composition metrics. For instance, space-time density is the ratio of blob type volume and the 3D world volume and population density is the number of individual per voxel. Blob number is the number of isolated blobs in a specific blob type.

Contagion and *STC* (spatiotemporal complexity) are two configuration metrics. Contagion is calculated based on (1) to measure dispersion of a blob type. This metric is based on voxel adjacencies and probability of finding a voxel of one blob type next to voxels in other blob types. Lower value of contagion shows many small blobs and higher value indicates few large blobs.

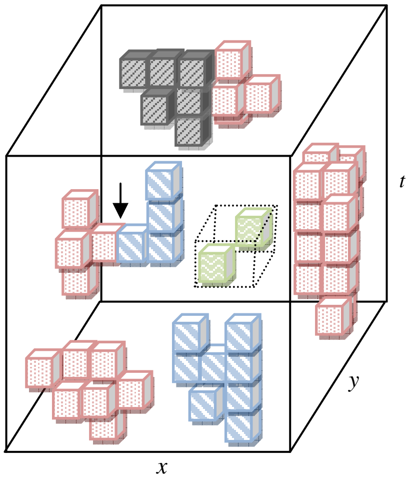

Fig 1. A Simple example of 4 blob types in a 3D world. Arrow shows 2 adjacent voxels with one shared face. The dashed cube is the bounding box of the green (wavy format) blob type.

$$RC = 1 - \frac{EE}{EE_{\max}} \qquad (1)$$

Where, RC is contagion and $EE_{\max} = b \times \ln(b)$ and b is number of blob types. Also

$$EE = -\sum_{i=1}^{b}\sum_{j=1}^{b} p_{ij} \ln(p_{ij}) \quad ; \quad p_{ij} = \frac{n_{ij}}{n_i} \qquad (2)$$

Where n_{ij} is number of adjacencies between voxels of blob type j and voxels of blob type i and n_i is the sum of all adjacencies for all species.

STC is a measure to describe how one blob type occupies the three dimensional space. *STC* is calculated by counting the number of voxels occupied by blob type i in a three-dimensional window of dimension $n \times n \times n$ where n is much smaller than the 3D world size ($n=5$ in our case). This window moves successively in the space-time cube and measures the different occupation levels from 0 to n^3 and then *STC* is calculated by (3). p_k is the relative frequency of occupation levels. *STC* is able to differentiate various patterns like uniform blob shapes (for example a column), random and complex pattern. *STC* value is lower for uniform or ordered blob shapes and is higher for complex shapes.

$$STC = \frac{\sum_{i=0}^{n^3} p_k \ln(p_k)}{\ln(n^3+1)} \quad ; \quad (0<STC<1) \qquad (3)$$

In total we have three spatial and eleven spatiotemporal metrics. Calculating these metrics for every species has been done for five independent runs of Ecosim for 10000 time steps of the simulation. The size of time dimension t in the 3D world for calculating spatiotemporal metrics is assumed to be 50. By increasing this size we would have more precise information about species history but it also increases the computational complexity of the defined metrics. So the size of the window to calculate spatiotemporal metrics is assumed to be 1000×1000×50.

III. Learning Algorithm and Evaluation Criteria

We would like to evaluate the capacity of the mentioned metrics in section II to predict speciation or species splitting events using the result of 10000 time steps of five different runs of Ecosim. Then we applied the following steps:

1) In each time step, we calculated the spatial information for each species and also 3D spatiotemporal metrics considering the information of the fifty previous time steps for each species to construct the blob types and to compute configuration and composition metrics.

2) Afterward we made one learning and one testing set. There are two classes in this dataset, Positive (Pos.) and Negative (Neg.), which specify if the speciation event will happen in next 100 times.

Repeating theses steps for all runs lead to five learning sets and five testing sets from five different runs. The main problem in all these datasets is that about 90 percent of samples belong to Neg. class and only about 10 percent of them are in Pos. class. It means that just 10 percent of species split in next 100 time steps. We have therefore an imbalanced dataset problem. There are two main approaches to address the imbalanced learning set problem [17]. One of them is to assign distinct costs to misclassified samples and try to minimize the overall cost on the training set. The second one is re-sampling, either by under-sampling major class or over-sampling minor class. In this research we examined different algorithms and finally we found out the smote algorithm [18] surpasses other algorithms in our case. For each sample of minority class, smote generates synthetic samples by selecting some of the nearest neighbors and generates new samples along the line segments connecting k minority class nearest neighbors. So we apply the smote algorithm on all learning sets. To guaranty that our learned models have the capacity to accurately predict the initial data, we only use the smote algorithm for the learning sets keeping the testing sets with the initial imbalanced properties of the whole dataset. Then we apply the C4.5 [19] algorithm to build decision tree based on attributes mentioned in section II for all learning sets. The interest of using such approach is that the obtained trees can be used for both speciation events prediction and rules extraction. These rules can effectively specify the most important factors in speciation according to spatial and spatiotemporal information. Then we evaluate the classifier performance on the test sets. To investigate the impact of different learning set on speciation event prediction, we repeat the above procedure for the four other datasets. Finally the last step is to assess the obtained results.

The performance of a machine learning algorithm is typically evaluated by overall accuracy. However it is not applicable for an imbalance dataset where only 10 percent of species split. For example, when there are 95% negative and 5% positive samples in a given dataset the accuracy of one classifier that detects 100% of Neg. class and 0% of Pos. class will be 95%. In this case the learning algorithm mostly learns the major class (Neg. class) while the minor class is highly important because it shows the correct prediction of samples with a speciation event. Consequently, simple overall accuracy is not a good measure to evaluate our classifiers performance.

For evaluating these classifier performances we use two metrics [20]; Recall and area under ROC curve (AUC) in addition to the overall accuracy based on confusion matrix for a 2-class classification problem. Recall shows the percentage of the given class correctly classified. ROC curve is used to show the classifier performance based on the Recall and false positive rate. Area under ROC curve (AUC) is a useful metric to measure how classifier performances are close to optimal.

IV. Experimental Results and discussion

In this section, we discuss the results of our experiments and also investigate the effect of the different attributes we used for speciation prediction.

A. *Classification with Spatial and Spatiotemporal Features*

As mentioned in section II, we have five different runs of Ecosim to generate one learning set and five testing sets. Afterward we applied the smote algorithm on the learning sets to make them balanced. However we do not change the class distribution of the testing sets. Finally, we applied the C4.5 algorithm using 10-folded cross validation to build the decision tree.

Tables I gives the distribution of learning sets for the 5 different runs before and after applying the smote algorithm. In all test sets we have in average approximately 10% Pos. class instances and the rest are in Neg. class. The three experiments Run1, Run2 and Run4, have almost the same number of species and they lead to very similar results. To simplify the presentation of the results we therefore chose to present in detail only the results for the Run3, Run4 and Run5 representing respectively situation with small, medium and large number of species. When we did the experiment before applying smote, we reached a high value for overall accuracy (above 90%) but very low Recall for minor class (less than 0.3). This happens because the classifier tends to learn the samples from the majority class and almost ignoring the ones from the minority class.

Table I
Learning Set Distribution Percentage and Number of Species per run

Learning Set	Pos. class percentage	Pos. Class Percentage after Smote	Number of Species
Run1	9.5%	48.6%	218
Run2	10%	48.6%	195
Run3	9.5%	48.3%	115
Run4	9.5%	48.4%	238
Run5	9.5%	43.4%	438

As mentioned in section III, because of the imbalance nature of our dataset, we examine true positive rate or Recall, AUC and overall accuracy to compare and evaluate the obtained results. In all the results, the oversampling method highly improved the Recall and the AUC values especially for minor class. As expected, we observed that we always have better prediction for the test sets coming from the same run as the learning set. For example in Fig. 2, Test5 and learning set Run5 are from the same run. It shows that the classifier comes out very good result for Test5 in compares to other test sets. Although the results for the test sets from the other runs (Test1, Test2, Test3 and Test4 in Fig. 2) is not as good as Test5, it shows that the classifier have learned some general rules of Ecosim speciation event. Some similar results have been obtained with Run3 too (Fig. 3).

By observing the results we noticed three different cases:

1) As mentioned in Table I, number of species in Run5 is 438. It means that for Run5 we can expect to have more valuable information in that dataset compared with other datasets like Run3 with 115 species. It is effectively confirm by our results; when we use Run5 as learning set we have better predictions for all the testing sets as it appears clearly in Fig. 2.
2) On the other hand, the worst results is when we used Run3 learning set and use it to predict the class for testing sets samples (Fig. 3). We can see also that the results are much more variables than with the other learning sets, confirming the lack of pertinence of the learned model.
3) Run4, Run1 and Run2 have an intermediate situation between case 1 and 2 when we have around 190 to 250 species.

Therefore, we found out that if we use a run with more species to make a classifier it has better generalization ability than a classifier that has been trained with a learning set from a run with lower species. It also means that some general rules about speciation exist in our system, as having more examples of speciation in one run help to predict speciation in another run with different conditions.

This is a strong result that comforts the choice of an individual-based system for understanding the speciation process. Our results also show a good capacity to predict speciation using spatiotemporal information. Even for the worst TP rate i.e. 22%, it indicates that the predictor effectively capture some important properties of speciation. It is even clearer if we consider that the average TP rate is 86% for prediction in the same run and 51% for prediction on other runs. Moreover, we observed that to obtain the best prediction results, we should have almost an evenness distribution of our two classes; Pos. and Neg., in the datasets (Table I).

B. *Comparing the Effect of Spatial Distribution and Spatiotemporal Information on Prediction*

To answer our initial questions we investigate the effect of the different attributes we used for speciation prediction. For example it is interesting to know which information; spatial or spatiotemporal metrics (information from 2D world or 3D world respectively); is more effective in the prediction. This will be helpful to extract some biological rules involved in speciation event.

For this purpose, we repeat the procedure described in section III two more times with different combination of attributes; first with only spatial distribution information and second with only spatiotemporal metrics. Fig. 4, Fig. 5 and Fig. 6 show the results summary.

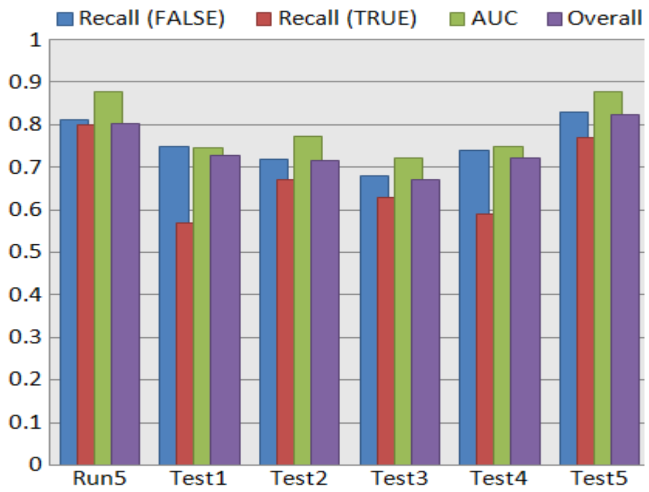

Fig. 2. Run5 is learning set and others are testing sets

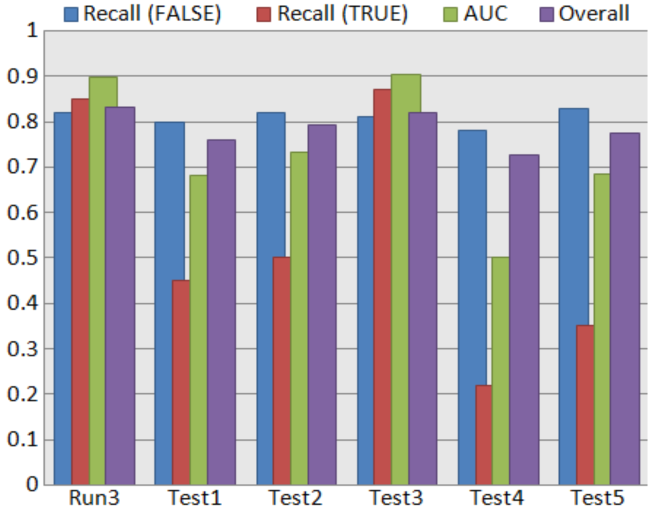

Fig. 3. Run3 is learning set and the others are testing set

In these figures, for Pos. class we represent the average of overall accuracy, Recall and area under ROC curve for all the learning and testing sets together (the column showed by All label), testing sets from the same run, learning sets, testing sets from other runs and finally all testing sets for all five runs. For figure legends, ST means the result for the dataset with only spatiotemporal metrics, S means only spatial information and ST+S means all the attributes.

By considering these figures, it is clear that the best results are when all the attributes are used in the learning process i.e. ST+S dataset. The most important results are for Testing (others) as they show the generic prediction capacity of our models, but the results for Testing (same run) are also important as they show that some specific property of each run have been captured which can be useful to characterize a specific run. Even though S has only three attributes while ST contains eleven metrics, when comparing them, the latter has only better result for overall accuracy on Testing (others). It means that spatial information has good capacity to learn generic rules and indicates that just using three spatial attributes leads to a relatively good prediction. Therefore spatial distribution information of individual in the world of Ecosim is very effective in predicting speciation. However it does not mean that dataset ST is not helpful in prediction, spatiotemporal information seems to be able to find specific properties of each run as depicted in Fig. 4, Fig. 5 and Fig. 6 where ST has better results for testing sets from the same run.

Moreover, adding spatiotemporal information to the spatial ones increases the quality of the prediction significantly. If we build the classifier based on datasets S and S+ST before oversampling; Recall or TP rate is very low for minor class (about 0.05 to 0.08) in S while that of S+ST is around 0.20 to 0.3 with approximately the same overall accuracy. It also improves AUC about 15% on average. Therefore it shows that by adding spatiotemporal metrics, the classifier is able to predict more minor class samples in presence of unbiased dataset. On the other hand, for biased datasets we observed 5% improvement for both overall accuracy and AUC for dataset S+ST on average for all runs in compare to that of S. However, if we consider the Testing (same run), it improves AUC, overall accuracy and Recall for 10%, 8% and 10% respectively.

These results show that spatial information of individuals in the world has great effect in speciation event prediction and spatiotemporal metrics can improve it. We also observed this fact in the rules extracted from the classifiers. For example for most of the predictors, spatial standard deviation is the decision tree root showing its importance for speciation prediction. However, more in deep analysis of the set of rules generated still need to be done.

V. Conclusion

In this research, we wanted to study how effective spatial and spatiotemporal information are in speciation prediction in an artificial ecosystem. We used 14 measures to extract this information and applying oversampling technique to build classifiers.

We obtained very good results for the test set coming from the same run as the learning set. Reasonably good results have been also obtained for the test sets from different runs showing that classifier can extract general rules about speciation that exist in our system. For all datasets; S, ST, S+ST, we also observed that the classifier performance goes up when the number of species contained in its learning set increases. It means that giving more examples of speciation events, even if they come from the same run, make the predictor more generic, which in turn means that some generic traits exist in our simulation that characterize the speciation events. This is highly important for the potentiality of our approach to discover some information useful for real prediction. Finally, we noticed that spatial information of individuals in Ecosim has tremendous effect on speciation prediction, as it has also been observed in real ecosystems, while spatiotemporal information can improve it in some extent.

For future work, we will study more in detail the results of speciation prediction and extract some important rules involved in speciation. It is also possible to work on other information of species like their genome or mating factors, to give better prediction for speciation.

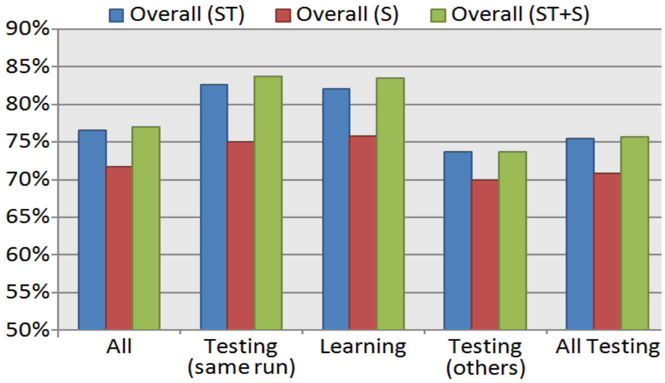

Fig. 4. Comparing overall accuracy for three combinations of attributes

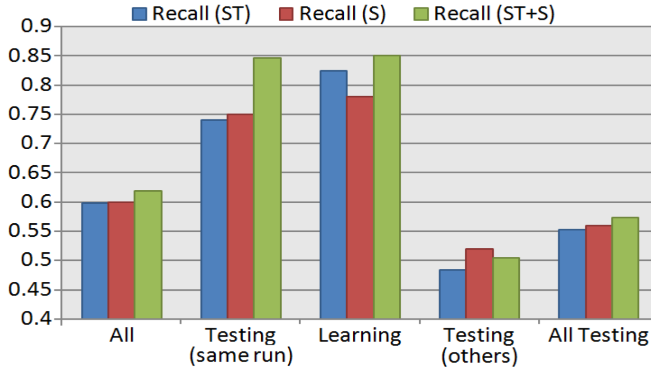

Fig. 5. Comparing Recall (Pos. Class) for three combinations of attributes

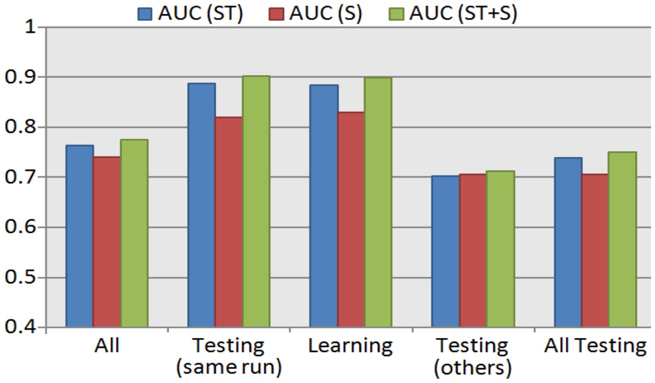

Fig. 6. Comparing AUC for three combinations of attributes

ACKNOWLEDGMENT

This work is supported by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca).

REFERENCES

[1] K.D., Queriroz, "Species concepts and species delimitation," *Syst. Biol. J.* vol. 56, no. 6, pp. 879–886, Aug. 2007.

[2] B. D. Mishler and M. J. Donoghue, "Species concepts: a case for pluralism source," *Systematic Zoology*, vol. 31, no. 4, pp. 491-503, Dec. 1982.

[3] E. Mayr, "Ecological factors in speciation source," *Evolution*, vol. 1, no. 4, pp. 263-288, Dec. 1947.

[4] S. Gavrilets, "Perspective: models of speciation: what have we learned in 40 years?" *Evolution*, vol. 57, no. 10, pp. 2197–2215, Oct. 2003.

[5] E. Mayr, *Systematic and the Origin of Species from the Viewpoint of a Zoologist*, NY, Columbia University, 1942, ch. 3.

[6] S. Ramachandran, et al., "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa," *PNAS* vol. 102, no. 44, pp. 15942-15947, Nov. 2005.

[7] J. B. Losos and R. E. Glor, "Phylogenetic comparative methods and the geography of speciation," *TRENDS in Ecology and Evolution*, vol. 18, no. 5, May 2003.

[8] V. Grimm, S.F. Railback, *Individual-Based Modeling and Ecology*, Princeton University Press, Princeton, 2005.

[9] R. Gras, D. Devaurs, A. Wozniak, A. Aspinall, "An individual-based evolving predator-prey ecosystem simulation using fuzzy cognitive map as behavior model," *Artificial Life*, vol. 15, no. 4, pp. 423-463, July 2009.

[10] A. Aspinall, R. Gras, "K-mean clustering as a speciation mechanism within an individual-based evolving predator-prey ecosystem simulation," in Proc. *6th Int. Conf. Active Media Technology*, Heidelberg, 2010, pp.318-329.

[11] D. Devaurs, R. Gras, "Species abundance patterns in an ecosystem simulation studied through Fishers logseries, Simulation Modeling Practice and Theory," *Simulation Modelling Practice and Theory*, vol. 18, no. 1, pp. 100-123, Jan. 2010.

[12] A. Golestani, R. Gras, "Regularity analysis of an individual-based acosystem simulation," *chaos: Interdisciplinary J. of Nonlinear Science*, vol. 20, no. 20, pp. 1-13, Oct. 2010.

[13] A. Golestani, R. Gras, "Multifractal phenomena in ecosim, a large scale individual-based ecosystem simulation," *Int. Conf. Artificial Intelligence*, Las Vegas, 2011, pp.991-999.

[14] S. M. Ibne, R. Gras, "Computation of population spatial distribution in individual-based ecosystem simulation," IEEE ALIFE 2011, Paris, in press.

[15] L. Parrott, R. Proulx, and, X. Thibert-Plante, "Three-dimensional metrics for the analysis of spatiotemporal data in ecology," *Ecological Informatics*, vol. 3, no. 6, pp.343-353, Dec. 2008.

[16] K Foroutan-pour, P Dutilleul, D.L Smith, "Advances in the implementation of the box-counting method of fractal dimension estimation", *Applied Math. and Comput.,*vol. 105, no. 2, Nov., 1999.

[17] H. Haibo, E.A.Garcia, "Learning from imbalanced data", *IEEE Trans. Knowledge and Data Engineering*, vol.21, no. 9, pp.1263-1284, Sept. 2009.

[18] N. V. Chawla, K. W. Bowyer,L. O. Hall, W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Artificial Intelligence*, vol.16, no. 1, pp.853-867, Jan. 2002.

[19] J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., 1993.

[20] N. V. Chawla, *Data Mining for Imbalanced Datasets: An Overview*, Data Mining and Knowledge Discovery Handbook, Springer, 2010.

Morteza Mashayekhi received the BSc degree in computer engineering from University of Isfahan, Isfahan, Iran in 2000, the MSc degree from Isfahan University of Technology, Isfahan, Iran in 2003. He was a faculty member in Karaj Azad University, Tehran, Iran from 2003 to 2010. He has published some articles in test and verification of digital circuits using evolutionary algorithms. Currently, he is a PhD student in Modeling & Simulation of Complex Biological Systems Lab, School of Computer Science, University of Windsor. His research interests are artificial Life, modeling & simulation of biological systems and machine learning.

Robin Gras received his BSc and his MSc in computer science at the university of Rennes, France. He completed his PhD in computer science applied to bioinformatics at INRIA of Rennes in 1997. From 2000 to 2004 he was senior scientist in Swiss Institute of Bioinformatics Geneva Switzerland after being post-doctorate from 1998 to 2000 in the same institute and lecturer, in 1998, at the University of Rennes, France. From 2000 to 2002 he was also consultant for GeneProt Inc. concerning the automation of protein identification and characterization process. He has been funded by NSERC, SSHRC, GeneProt Inc. (Switzerland), CNRS (France), INRIA (France). He also received CFI and ORF infrastructure grants. Now, he is an associate professor and Canada Research Chair in school of computer science and cross-appointed by the Biological Department. His domains of research are: artificial life, theoretical biology, ecosystem simulation, predator-prey model, bioinformatics, combinatorial optimization, machine learning.