

Mining Ecological Data with Cellular Automata

Alexander Campbell, Binh Pham, and Yu-Chu Tian

Centre for Information Technology Innovation
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
ab.campbell@qut.edu.au

Abstract. This paper introduces a Cellular Automata (CA) approach to spatiotemporal data mining (STDM). The recently increasing interest in using Genetic Algorithms and other evolutionary techniques to identify CA model parameters has been mainly focused on performing artificial computational tasks such as density classification. This work investigates the potential to extend this research to spatial and spatiotemporal data mining tasks and presents some preliminary experimental results. The purpose is twofold: to motivate and explore an evolutionary CA approach to STDM, and to highlight the suitability of evolutionary CA models to problems that are ostensibly more difficult than, for example, density classification. The problem of predicting wading-bird nest site locations in ecological data is used throughout to illustrate the concepts, and provides the framework for experimental analysis.

1 Introduction

Space-time dynamics are ubiquitous in both real-world and artificial systems. The recently emerging field of spatiotemporal data mining (STDM) opens up a number of possibilities for the integration of previously distinct research areas. Cellular Automata (CA) techniques, in particular the use of Genetic Algorithms and other evolutionary approaches to learn the transition rules from given patterns, show great potential to be combined with data mining techniques in order to tackle problems of previously prohibitive complexity. They can also give insight to improve on previous approaches.

In order to illustrate our approach, we use a classic problem in spatial data mining - the prediction of nest site locations in a wading bird ecology - and examine the merits and difficulties of applying an evolutionary CA approach. The data consists of nest site locations and environment variables (e.g. vegetation type, water depth). Given a set of spatial patterns (the training data), the aim is to construct a CA neighbourhood and transition rule such that they will predict the locations (or significantly aid prediction) of nest sites on unseen data sets consisting of the environment variables only.

Genetic Algorithms have been used to discover CA rules that will perform computational tasks such as density classification (complete cellular grid to become one of two states depending on which had the higher initial density) [1],

synchronisation (all cells to flash on and off in synchrony) and most recently period three and quasiperiod three behaviours [2]. They have also been applied to identifying the neighbourhood and transition rules that will recreate spatiotemporal patterns generated by CA [3]. The closest in nature to our work is that of Richards et al. [4] who also look at extracting cellular automata rules directly from *experimental data*. In their case the data relate to the dendritic solidification of NH_4Br . The key theme here is to expand this concept to more complex multivariate scenarios.

Section 2 introduces spatiotemporal data mining through a formal treatment of the location prediction problem. In section 3 we motivate and present our evolutionary CA model. Section 4 details the experimental setup and contains preliminary results and analysis. Finally, section 5 concludes and discusses future work.

2 Spatiotemporal Data Mining

Spatiotemporal data mining is a relatively recent expansion of data mining techniques to take into consideration the dynamics of spatially extended systems for which large amounts of data exist. Given that all real world spatial data exists in some temporal context, and knowledge of this context is often essential in interpreting it, spatial data mining is inherently STDMM to some degree. Although time series and spatial data mining have existed as research fields for a number of years independently, they have been growing towards a synthesis that is inherently multi-disciplinary.

2.1 The Location Prediction Problem

This multi-disciplinary nature is borne out by the fact that a well known and thoroughly researched spatial data mining application is that of predicting nest site locations in an ecological data set. A particular data set was first considered by Ozesmi [5] in an ecological modelling context, and then taken up as a spatial data mining problem by Shekhar et. al. [6].

Following [6], the problem is defined as follows: Given

- A spatial framework S of sites $\{s_1, \dots, s_n\}$ for an underlying geographic space
- A collection X of explanatory functions $f_{X_k} : S \rightarrow \mathbb{R}^k, k = 1, \dots, K$. \mathbb{R}^k is the range of possible values for the explanatory functions
- A dependent class variable $f_C : S \rightarrow C = \{c_1, \dots, c_M\}, c_m \in \{0, 1\}$

Find: Classification Model $\hat{f}_C : \mathbb{R}^1 \times, \dots, \times \mathbb{R}^k \rightarrow C$

In [6], a number of approaches are compared both theoretically and experimentally. In order to gain insight regarding spatially dependent multivariate processes, and to motivate the use of an alternative technique, these approaches are also briefly reviewed here in order of increasing sophistication.

Linear Regression. The classical linear regression equation does not attempt to model spatial dependence. It is defined as

$$y = X\beta + \epsilon \tag{1}$$

where $\beta = (\beta_0, \dots, \beta_m)^T$, y is an n -vector of observations and X is an $n \times m$ matrix of explanatory data. This is then transformed from a real-valued variable to binary via the logistic function, $Pr(c_i|y) = e^y / (1 + e^y)$. When the samples are spatially related, the residual errors reveal a systematic variation over space.

Spatial Autoregressive Models. In order to model spatial dependence, a spatial contiguity matrix is used. The essential idea is that spatial locations which are neighbours are coupled through an entry in the matrix. The simplest form is a binary matrix where entries are either 1 or 0 depending on a Euclidean adjacency metric, however ‘neighbours’ can be defined in any fashion and non-zero elements are often scaled to sum to unity in each row. Figure 1 shows an example spatial framework and figure 2 a corresponding row-normalised contiguity matrix where cells that share an edge have double the weighting of cells that are diagonal neighbours. The modified regression equation is

$$y = \rho W y + X\beta + \epsilon. \tag{2}$$

The dependence on neighbouring classes is exerted in a static fashion through the contiguity matrix. The fact that it is at heart a logistic regression model means there is an assumption that the class conditional distribution belongs to the exponential family.

Markov Random Fields. A more sophisticated approach is to model the data as a Markov Random Field (MRF). A MRF-based Bayesian classifier is a non-parametric model which, unlike logistic regression, is not bound to make assumptions that the class conditional distribution belongs to a particular family. A MRF explicitly models the relative frequencies in the class prior term. Conceptually it is therefore more suited to the real-world nonlinearity of this

S1	S2
S3	S4

Fig. 1. An example spatial framework

	S1	S2	S3	S4
S1	0	0.4	0.4	0.2
S2	0.4	0	0.2	0.4
S3	0.4	0.2	0	0.4
S4	0.2	0.4	0.4	0

Fig. 2. Row-normalised contiguity matrix for the spatial framework shown in figure 1

problem and accordingly has the best experimental performance in the study in [6].

MRFs consist of a set of random variables with an interdependency relationship represented by an undirected graph which is analogous to the spatial contiguity matrix. The Markov property contends that a variable depends only on its neighbours and is independent of all other variables. For this problem, that means random variable l_i is independent of l_j if $W(s_i, s_j) = 0$, where W is the neighbourhood relationship contiguity matrix.

Using Bayes rule it is possible to predict l_i from feature value vector X and neighbourhood class label vector L_i as follows:

$$Pr(l_i|X, L_i) = \frac{Pr(X|l_i, L_i)Pr(l_i|L_i)}{Pr(X|L_i)} \quad (3)$$

Where L_i is the label vector for the neighbourhood of i .

2.2 The Curse of Dimensionality

Despite its theoretical ability to encode complex probabilistic relationships involving multiple variables without reliance on any assumptions about the distributions of class variables, the actual solution procedures for MRFs require large amounts of training data and/or a number of limiting assumptions on $Pr(X|l_i, L_i)$. This is because the graph of a Markov field must connect all pairs of variables that are conditionally dependent even for a single choice of values of the other variables [7]. In other words, it is hard to encode interactions that occur only in a certain context and are absent in all others. An example would be that the probability of a nest site occurring in a location with a certain combination of environment variables is very different when there happens to be three or more birds within a certain distance of that location. In many situations these assumptions may be acceptable, however if there is a high degree of attribute interaction this is not necessarily the case. Attribute interaction is a crucial theme in data mining [8] and is common in many complex real-world systems. The ‘curse of dimensionality’ is the situation where there is a high degree of attribute interaction over a sparse amount of data.

3 Evolutionary CA for Data Mining

The importance of approaching analysis of complicated systems with the goal of *understanding rather than prediction* is becoming more widely acknowledged. This places the emphasis on dynamic process rather static pattern, as noted by [9]. A model that has such a characteristic is Cellular Automata. CA are well suited to a high degree of attribute interaction and sparse data because they are not a statistical model of the system which requires a large amount of data to give ‘support’ to the probability distributions. Instead they are a model that can map every state of the environment to a state of the dependent variable. The

penalty for such flexibility however is that it is required to search a potentially massive parameter space for the correct rules.

Evolutionary methods are highly suited to such a problem and have been used to evolve CA transition rules for a number of computational tasks. The key theme of evolutionary approaches is that the behaviour of the system is defined implicitly by a fitness function measuring the difference between the candidate solution and the desired solution. No explicit knowledge is needed to show the system *how* to perform the task.

The line between a *forward* CA approach - conscious programming of CA rules with a resultant behaviour in mind - and a backward or *inverse* approach where no *a priori* knowledge is used and the desired behaviour is the only guide, is inherently indistinct. The nature of the relationship mirrors closely the mining-modelling one mentioned previously. In the CA literature this type of problem seems to have been predominantly approached in a forward sense. Examples of this abound especially in urban modelling [10,9] and, perhaps closer in form to our problem, in vegetation dynamics [11].

The problem is approached here in an inverse sense because from a data mining perspective this means there is no explicit requirement for prior domain knowledge. It also ensures the possibility to discover rules that are counter-intuitive and/or too complex to program consciously. Thus it is both more flexible and more general. In the evolution of a complex, high-dimensional system, it is often the case that there are only a relatively small number of degrees of freedom which contribute (above a certain threshold, which is indistinguishable from noise) to the dynamics, such that the attractor of such a system can be reconstructed in a greatly reduced state space. We see these two characteristics as analogous which motivates us to look for a parsimonious rule-set that will represent the important factors in nesting behaviour. Having said that, the immense size of state space is still a huge obstacle, and the method proposed here is intended only as a complementary approach: in combination with domain knowledge and other data mining techniques it offers the chance to discover nonlinear spatial and spatiotemporal relationships, even in the face of sparse data.

3.1 Cellular Automata Model Definition

The model we propose is a synchronous two-dimensional Cellular Automata where data belonging to each grid-cell is the basis for the state of a cell in the cellular lattice. More precisely, each cell's state can be split into two parts. The first is the three independent variables: distance to open water, vegetation durability, and water depth; these remain constant for the duration of each simulation run. The second state denotes the presence or absence of a nest site - this state will be more typical of a CA in that it will be recalculated at each time step according to the current rules. The first state can be collectively thought of as the landscape in which the simulations take place. The second can be thought of as the progressive spatial distribution of the birds as they are drawn towards the attractor which characterises their nest site behaviour.

Our CA can initially be defined as: $\mathbf{CA} = \langle \Omega, Q, N, I, f \rangle$

Where:

- $\Omega = \{(i, j) | 1 \leq i \leq L_x, 1 \leq j \leq L_y\}$ is the $L_x \times L_y$ lattice of cell sites
- Q is a finite set of cell state values
- N is a neighbourhood template
- $I : \Omega \rightarrow Q$ is the initialisation function
- $f : Q \times Q^{|N|} \rightarrow Q$ is the transition function

Incorporating the two-part cell state values we have $Q = \{Q_l, Q_n\}$, where:

$$Q_l = \{DOW, V, WD\}, Q_n = \{0, 1\} \tag{4}$$

Where Q_l is the state of the landscape consisting of *DOW* - distance to open water, *V* - vegetation durability, and *WD* - water depth. Q_n is the presence or absence of a nest site.

The transition function becomes:

$$f : Q_l \times Q_n \times Q_l^{|N_l|} \times Q_n^{|N_n|} \rightarrow Q_n \tag{5}$$

Where N_n is the neighbourhood for nest sites and N_l is the landscape neighbourhood. We may additionally wish to have different sized neighbourhood for each individual landscape variable depending on domain knowledge that certain variables have a larger influence.

3.2 Genetic Algorithm Model Definition

Similar to a number of evolutionary CA papers we use a standard CA-GA model of $\mathbf{GA} = \{P, T, \rho, G, E\}$, where P is population size of candidate rules, T is the number of CA time steps, ρ is the mutation rate, G is the number of generations and E is the number of elite rules. However, unlike the artificial computational tasks tackled by other evolutionary CA methods which aim to converge to a particular target configuration, we are looking to learn a complex probability distribution that represents the nest location ‘behaviour’ of the wading birds. In order to realistically rate the ability of a CA rule to do this, it is necessary to run the GA on a number of training configurations (of target nest site locations) which follow some predetermined probability distribution. Also, the number of CA time steps before the fitness is measure is nonstandard, and is discussed in the next section.

The fitness function of a rule is perhaps the most crucial ingredient in a successful GA. Our fitness function takes into account spatial accuracy so that configurations which are spatially near the target configuration are still rewarded. This may be crucial given the many potential sources of noise and uncertainty in spatiotemporal data sets, and also highly appropriate for discovering qualitative relationships. A modified version of the Spatial Accuracy Measure (SAM) outlined in [6] is used.

$$SAM = TPR - FPR = \frac{AnMPn}{AnMPn + AnMPnn} - \frac{AnnMP}{AnnMPn + AnnMPnn} \tag{6}$$



Fig. 3. Distance to Open Water, DOW

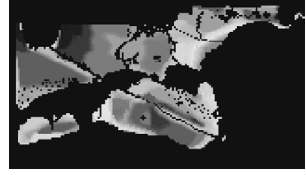


Fig. 4. Water Depth, WD

where TPR is the true positive rate, FPR is the false positive rate, $An[i] = f_C[s_i]$ and $Pn[i] = \hat{f}_C[s_i]$ are boolean vectors representing actual nest locations and predicted nest site locations respectively, and $Ann[i] = 1 - An[i]$ and $Pnn[i] = 1 - Pn[i]$ are their inverses. The spatial weighting is applied through $M = W + I$, the matrix addition of contiguity matrix W and identity matrix I .

4 Experimental Setup, Preliminary Results, and Analysis

In [12], high fidelity colour images of the three spatial data sets from the study in question are provided, along with a colour-intensity scale. Due to a number of factors, including: a desire to operate on the same data set for future comparisons, the qualitative nature of our approach at this initial stage, and an unavailability of other suitable data, we have used these images (rather than raw data) for our experiments. The relative values of the variables were preserved by creating a colour map based on the provided intensity levels, and the spatial integrity of the data was preserved by reducing the number of cells (pixels) in the image down to the corresponding number of data grid points in the original study.

In the Spatiotemporal Data Mining section it was shown that both regression and MRF models may be forced to make limiting assumptions, especially when there is a high degree of attribute interaction. Our first goal was to show that an evolutionary CA approach is highly suited to problems that are nonlinear in this way.

In order to investigate this in an experimental fashion, synthetic nest site locations were generated using a nonlinear generalisations of equation (2):

$$y_{syn1} = (I - \rho W)^{-1} \times (\beta \times \cos(X) + c \times \text{random}(\epsilon)) \quad (7)$$

$$y_{syn2} = (I - \rho W)^{-1} \times (\beta \times \cos(X) + W \times \cos(X) \times \gamma + c \times \text{random}(\epsilon)) \quad (8)$$

where W is an equally weighted, row-normalised 7×7 neighbourhood, X is the DOW explanatory variable, ρ is an arbitrary spatial weighting of 0.6, c is a relatively small noise weighting of 0.5, $\epsilon = N(0, 1)$, and β , the weighting on local explanatory data, has been set arbitrarily at 0.3. In equation (9), γ represents the dependence of the dependent variable on neighbouring explanatory data; this has been set high (relative to β) at 0.25 to generate a significantly more nonlinear problem.

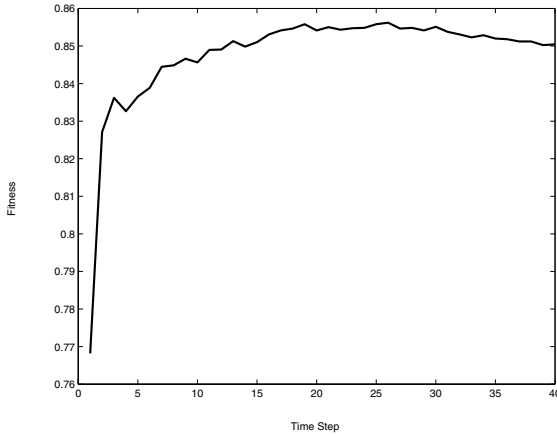


Fig. 5. Gradual fitness increase over CA time steps

Rule Types and Transients. The CA model defined above is very general and is likely in most situations to have a prohibitively large state space. In order to reduce this state space a number of different simplifications have been used. Firstly the explanatory data has been quantised so that $Q_t = \{0, 1, 2, \dots, 16\}$. The level of quantisation should be partly determined by the specific domain. More importantly, a number of rule types have been considered. Initially it was thought that an outer-totalistic rule would be sufficient to capture the extent of the nest-site location interactions, however in looking for a rule that produced an increase in fitness over a number of CA time steps (ie one that reached a higher ‘equilibrium’ fitness) it was discovered that symmetric rules tend to oscillate between fitness levels in a more extreme fashion. More complex rule types, while when averaged over a large number of specific rules may not have performed any better (plenty decreased in fitness over the CA time steps), seemed to produce a larger number of rules that purely increased in fitness. The mechanism behind this phenomena is more complex rules’ greater propensity for *information transfer* across the spatial lattice. In order to retain this property, while increasing the state space as little as possible, reflection symmetric (rather than rotation symmetric or completely symmetric) rules were preferred. Figure 5 shows the increase in fitness over a number of time steps for a von Neumann reflection symmetric rule on a single arbitrary target nest location data set. An empirical study of the transient behaviour of a large number of CA rules led to the requirement that at least Te CA time steps elapse before the spatial accuracy is measured, at which point it is averaged over at least the next Tav time steps, and $T = \{Te, Tav\}$.

Nonlinear Generalisation Capabilities. Thus far a reduced complexity version of the model outlined above has been implemented where the landscape

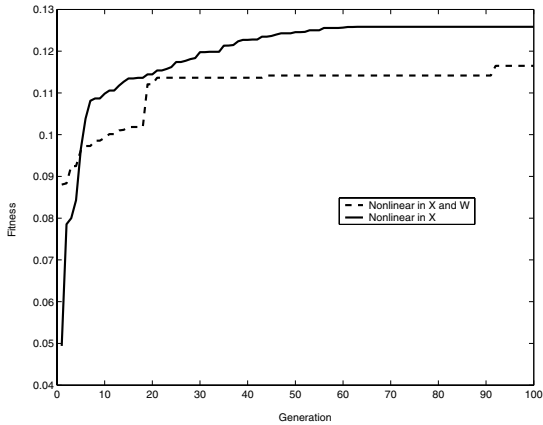


Fig. 6. Fitness of GA for two scenarios: one nonlinear in the local explanatory data only, the other additionally nonlinear in neighbouring explanatory data

consists only of the distance to open water variable. Although we have not yet performed enough experiments for statistical significance, so far results are encouraging. Figure 6 is a typical graph of fitness vs. generation where for both experiments: $\mathbf{GA} = \{P, Te, Tav, \rho, G, E\} = \{100, 30, 5, 0.01, 100, 20\}$. Each candidate CA rule is matched against 10 synthetically generated target nest site locations according to equation (8) for the first experiment and (9) for the second.

Both populations are performing at similar reasonable fitness levels, especially given the difficulty of the task. Secondly and more interestingly, in the population with the more nonlinear probability distribution to learn, the fitness increases take longer to come, but are larger when they do. Intuitively this matches our perceptions about the nature of the tasks and the GA's ability to learn nonlinear relationships for a CA model.

5 Conclusion and Future Work

We have used an analysis of statistical spatial models to give insight into the nature of attribute interaction and as the basis for experimental metrics. We have explored the potential for an evolutionary CA approach to STDM and presented some encouraging, though preliminary, experimental results.

Of the two goals set out at the beginning, more emphasis has been placed on the suitability of CA to nonlinear problems. Hopefully this has provided an experimentally justified foundation for the application of CA to a broader range of problems. The suitability of this technique for data mining specifically has been given less attention due to space constraints and future work will be needed to develop the interpretation of CA rules for data mining purposes in

greater depth. Also, the issue of sensitivity to noise needs to be investigated, possibly through the application of a probabilistic CA model.

Acknowledgements. This project is supported by an Australian Research Council Linkage grant in collaboration with the Built Environment Research Unit (BERU), Queensland Department of Public Works.

References

1. Jimenez-Morales, F., Crutchfield, J., Mitchell, M.: Evolving two-dimensional cellular automata to perform density classification: A report on work in progress. *Parallel Computing* **27** (2001) 571–585
2. Jimenez-Morales, F.: An evolutionary approach to the study of non-trivial collective behaviour in cellular automata. In Bandini, S., Chopard, B., Tomassini, M., eds.: *ACRI 2002*. Volume 2493 of LNCS., Geneva, Switzerland, Springer (2002) 32–43
3. Billings, S., Yang, Y.: Identification of probabilistic cellular automata. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics* **33** (2003) 225–236
4. Richards, F.: Extracting cellular automaton rules directly from experimental data. *Physica D* **45** (1990) 189–202
5. Ozesmi, U., Mitsch, W.J.: A spatial habitat model for the marsh-breeding red-winged blackbird (*agelaius phoeniceus* l.) in coastal lake erie wetlands. *Ecological Modelling* **101** (1997) 139–152 TY - JOUR.
6. Shekhar, S., Schrater, P., Vatsavai, R., Wu, W., Chawla, S.: Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia* **4** (2002) 174–188
7. Fridman, A.: Mixed markov models. *Proceedings of the National Academy of Sciences* **100** (2003) 8092–8096
8. Freitas, A.: Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review* **16** (2001) 177–199
9. Cheng, J., Masser, I.: Cellular automata based temporal process understanding of urban growth. In Bandini, S., Chopard, B., Tomassini, M., eds.: *ACRI 2002*. Volume 2493 of LNCS., Geneva, Switzerland, Springer (2002) 325–336
10. O’Sullivan, D., Torrens, P.M.: Cellular models of urban systems. In Bandini, S., Worsch, T., eds.: *ACRI 2000*, Karlsruhe, Germany, Springer (2000) 108–116
11. Bandini, S., Pavesi, G.: Simulation of vegetable populations dynamics based on cellular automata. In Bandini, S., Chopard, B., Tomassini, M., eds.: *ACRI 2002*. Volume 2493 of LNCS., Geneva, Switzerland, Springer (2002) 202–209
12. Chawla, S., Shekhar, S., Wu, W., Ozesmi, U.: Modeling spatial dependencies for mining geospatial data: An introduction. In Miller, H., Han, J., eds.: *Geographic data mining and Knowledge Discovery*. Taylor and Francis (2001) 338