2D1-5

A Study On The Viability of Evolutionary Computation Methods for the Generation of Earthquake Predictability Models

Claus de Castro Aranha*1 Bogdan Enescu*2

*1University of Tsukuba, Graduate School of Science and Information Engineering
*2University of Tsukuba, Faculty of Life and Environmental Sciences

Understanding the mechanisms and patterns of earthquake occurrence is of crucial importance for assessing and mitigating the seismic risk. In this work we make an initial analysis of the viability of Evolutionary Computation as a means of generating models for the occurrence of earthquakes. We use the Japan Meteorological Agency (JMA) earthquake catalog. Our proposal is made in the context of the "Collaboratory for the Study of Earthquake Predictability" (CSEP), an international effort to standardize the study and testing of earthquake forecasting models. As a starting point, we propose a methodology for encoding earthquake risk models in a Genetic Algorithm as a real valued genome, where each allele corresponds to a bin in the forecast model. A fitness function based on the log-likelihood of the evolved model against the earthquake occurrence data is used. Because of the high dimensionality of the chromosome, special attention has to be paid to crossover, mutation and validation operators. A careful discussion of these factors is conducted, and the results of our experiments are interpreted using a simple geophysical model.

1. Introduction

Earthquakes pose a great risk for human society, in their potential for large scale loss of life and destruction of infrastructure. In the last decade, large earthquakes such as Sumatra (2004), Kashmir (2005), Sichuan (2008) and Tohoku (2011) caused terrible amounts of casualties. Thus, it is important to understand the mechanisms behind the occurrence of Earthquakes. This knowledge may allow us to create better models of seismic risk, which can be used for mitigating damage through urban planning and emergency preparedness.

Surprisingly enough, Evolutionary Computation has so far played a limited role in the investigation of seismic events. In this work, we try to increase this dialogue by designing a simple Genetic Algorithm system applied to the task of generating earthquake forecast models based on catalog data. This task is based on the framework laid out by the *Collaboratory for the Study of Earthquake Predictability* (CSEP*1). The CSEP is an international effort for the standard and rigorous testing of predictability models [6]. The CSEP framework defines testing regions, periods, and statistical testing methodologies for model comparison. We use this information to orient our own experiments

To handle this task, our algorithm encodes a prediction model in its genome, and evolves this genome based on the value of the log-likelihood between the model and historical data taken from the Japan Meteorologial Agency (JMA) catalog. We compare the results obtained by the evolutionary algorithm with the Relative Intensity algorithm (RI), which is often used as a benchmark for this problem [5]. Our results indicate that the GA found effective models when compared to the RI, specially when considering inland events.

Although our target problem is to generate a prediction model, our ultimate goal is not to forecast the ocurrence of earthquakes. Instead, we aim at using the creative power of evolutionary computation to find new patterns in the mechanics of seismic events. This work is a first step towards that goal, a study case to highlight the considerations needed using EC in this field.

1.1 Literature Review

In the international literature, there is very little mention of the use of Genetic Algorithms for seismic research. For forecasting models, Zhou and Zu [10] recently proposed a combination of ANN and EC, but their system only forecasts the magnitude of earthquakes.

A somewhat more common approach is the use of EC for estimating parameter values in seismological models. For example, a few works use Evolutionary Computation to estimate the peak ground acceleration of seismically active areas [4, 1, 3]. Another variation of the same theme is the determination of Fault Model parameters of an earthquake using EC [7, 2].

2. GA Prediction Model

In this system, an individual's genome encodes a forecast model, as defined in the CSEP framework. The model consists of a series of *bins*, corresponding to locations in a geographical grid. For each bin, an integer denotes the number of *expected events*. In other words, each chromossome in the genome corresponds to the number of expected earthquakes in one bin (location).

While encoding the individual's genome in this way is a simple and direct method, we identified two concerns: 1- Testing Regions in the CSEP framework contain over a thousand bins, so the genome will be correspondingly large.

Contact: Claus Aranha, University of Tsukuba, Graduate School of SIE; Tennodai 1-1-1, Ibaraki, Tsukuba; 029-853-6574; caranha@cs.tsukuba.ac.jp

*1 http://www.cseptesting.org

2- Because of the very large number of parameters, a careful design of evolutionary operators is necessary, to avoid early overfitting.

2.1 Genome Representation

To avoid problems associated with integer values in the cromossomes, the genome is a real valued array. Each element in the array corresponds to one bin (geographical location) in the prediction model. An element x_i takes a value from [0.1). In the initial population, these values are sampled from a uniform distribution.

In order to transform these real values into integer forecasts, as required by the CSEP model, we extract Poisson deviates from the real values, using the following algorithm:

Algorithm 1 Obtain a Poisson deviate from a $[0, 1)$ value

Parameters $0 \leq x < 1, \mu \geq 0$
$L \leftarrow \exp(-\mu), k \leftarrow 0, prob \leftarrow 1$
repeat
  increment k
  $prob \leftarrow prob * x$
until $prob > L$
return k

2.2 Fitness Function

The basis for the fitness function in our system is the *Log Likelihood* between the forecast generated by an individual and the observed earthquakes in the training data, as described by Schorlemmer et al. [8]. Given an individual, Let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n | \lambda_i \in N\}$ be the forecast generated by this individual, with n bins. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n | \omega_i \in N\}$ be the observed numbers of earthquakes for each bin i in the training data (the catalog data). The log likelihood between an individual's forecast Λ_X and the observed data Ω is calculated as:

$$L(\Lambda_X|\Omega) = \sum_{i=0}^{n} -\lambda_i + \omega_i * \ln(\lambda_i) - ln(\omega_i!). \qquad (1)$$

However, in our early testing we observed that using a simple log-likelihood as the fitness parameter caused the population to degenerate into local optima. To avoid that, we apply a *time-slice* operation to the fitness function. We break up the training data into smaller slices, based on the chronology of the earthquakes in the data. The duration of each slice is the same as the duration of the test data.

Let's consider an example: the target period for the forecast is one year, from 1/1/2014 to 1/1/2015, and the training data is taken from the 10 year period between 1/1/2004 and 1/1/2014. To apply the time-slice log likelihood fitness function, we divide the training data into ten 1-year slices, from 2004 to 2005, 2005 to 2006, and so on. The final fitness of the individual will be the minimum value from all time slices.

2.3 Evolutionary Operators

Our system uses a regular generational genetic algorithm. For selection, we use Elitism and Tournament selection. We use a simple Uniform Crossover for the crossover operator. If a gene's value falls outside the $[0, 1)$ boundary, it is truncated to these limits. For the mutation operator, we sample entirely new values from $[0, 1)$ for each mutated chromosome.

Population Size	500
Generation Number	100
Elite Size	1
Tournament Size	50
Crossover Chance	0.9
Mutation Chance (individual)	0.8
Mutation Chance (chromosome)	$(\text{genome size})^{-1}$

Table 1: Parameters for GAModel

The parameters used for the evolutionary computation are described in Table 1. Because our focus is to show the viability of Evolutionary Algorithms for this application problem, we are not yet particularly concerned with the convergence speed of the system.

3. Current Results

To evaluate the performance of our proposed system, we execute a simulation experiment, and contrast its results to those obtained by the Relative Intensity (RI) method, and by an "unskilled" (random) forecast. For the three methods, we compare the final values of their log likelihood, and also the value for the ASS (Area Skill Score), which measures the number of false positives and false negatives in a forecast model [9].

The simulation is performed on catalog data made available by the Japan Meteorological Agency (JMA). It includes the time, magnitude, latitude, longitude and depth of the hypocenter. We consider events with magnitude above 2.5 and depth less than 100km, recorded in the period from 2000 to 2013. We selected two regions to investigate: Kanto and Northern Honshu. These areas are of interest because they show a good mix of inland and off-shore quakes, that make it easier to observe differing patterns in the methods. For each area, we performed 8 simulations, with target periods of one year (2005 to 2012), and training periods of 5 years prior to the target period.

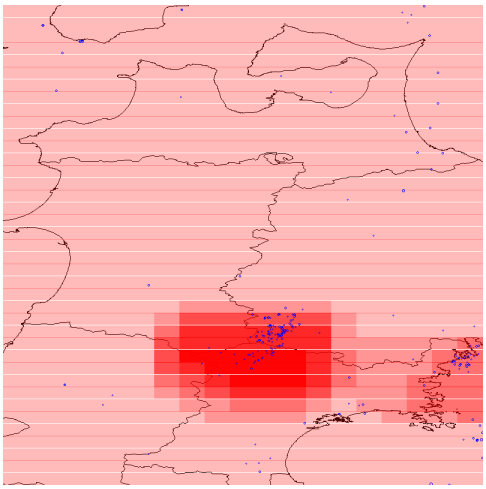
The results of the simulation experiments are summarized in Table 2. In this table, *Random* refers to the Random forecast, *RI* to the Relative Intensity algorithm, and *GA* to the proposed evolutionary system.

The GA column reports the average for 20 runs, and the standard deviation is reported in parenthesis. The *p-value* column indicates the result of a one-sided T-test where the alternate hypothesis is "The GA average is greater than the RI result".

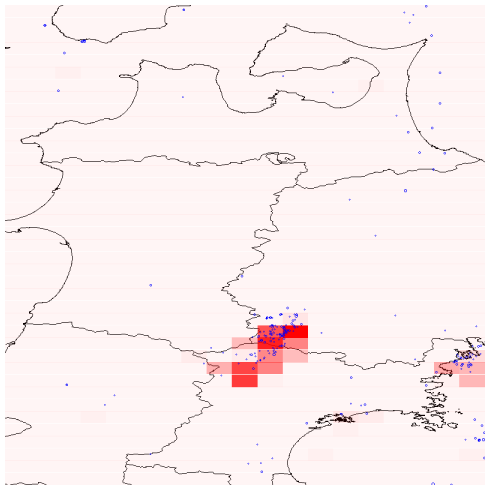
We note from the table that in general the evolutionary system has outperformed the RI in the Kanto area. In particular, we note that the GA based model forms smaller alarm clusters inland, while the RI marks large areas with its predictions.

Scenario		Log Likelihood				Area Skill Score			
		Random	RI	GA	p-value	Random	RI	GA	p-value
Kanto	2005	-3716.86	-2263.4	-2253.2 (16.5)	0.006	0.38	0.24	0.24 (0.04)	0.78
	2006	-3884.85	-2252.28	-2234.72 (14)	0.001 >	0.36	0.10	0.18 (0.01)	0.001 >
	2007	-3838.9	-2113.84	-2108.95 (11.1)	0.03	0.36	0.15	0.19 (0.02)	0.001 >
	2008	-3914.54	-2110.79	-2096.75 (11.8)	0.001 >	0.39	0.16	0.22 (0.3)	0.001 >
	2009	-4211.28	-2487.88	-2482.88 (10.3)	0.02	0.36	0.09	0.14 (0.01)	0.001 >
	2010	-4010.47	-2132.11	-2099.13 (16.3)	0.001 >	0.39	0.14	0.28 (0.03)	0.001 >
	2011	-17657.43	-20083.09	-19983.73 (144.4)	0.003	0.35	0.07	0.08 (0.02)	0.14
	2012	-10863.99	-3225.39	-4435.34 (248)	1	0.48	0.80	0.77 (0.01)	1
Northern Honshu	2005	-2552.61	-1067.38	-984.23 (84)	0.001 >	0.58	0.58	0.62 (0.01)	0.001 >
	2006	-2613.1	-1044.72	-1073.03 (154)	0.78	0.52	0.50	0.42 (0.03)	1
	2007	-2666.11	-1049.82	-999.64 (83.6)	0.007	0.51	0.51	0.41 (0.01)	1
	2008	-5124.54	-5007.49	-4704.15 (131)	0.001 >	0.36	0.05	0.18	0.001 >
	2009	-2737.47	-1049.22	-936.63 (60)	0.001 >	0.54	0.67	0.70 (0.01)	0.001 >
	2010	-2714.68	-1045.03	-1077.95 (136)	0.85	0.53	0.66	0.57	1
	2011	-3435.67	-2753.95	-2963.31 (88)	1	0.40	0.21	0.10 (0.01)	1
	2012	-3623.22	-1326.52	-1186.1 (45.3)	0.001 >	0.47	0.62	0.70 (0.05)	0.001 >

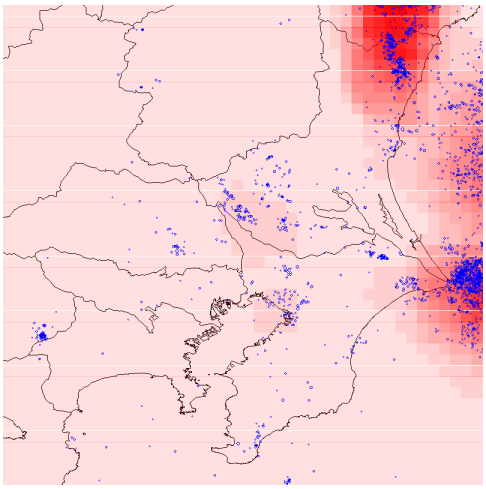
Table 2: Simulation Experiment Results



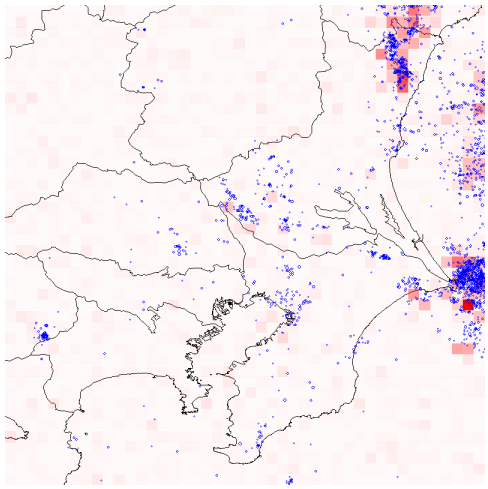
(a) Northern Honshu 2009, RI



(b) Northern Hoshu 2009, GA



(c) Kanto 2012, RI



(d) Kanto 2012, GA

Figure 1: Two simulation results. Red squares indicate the intensity of the forecast. Blue circles indicate actual earthquakes in the test data.

Even then, both models miss some clusters in this area, as indicated by their ASS value being under that of the random model. The ASS value is useful to estimate how much a forecast is suffering from over fitting - a low value indicates that a larger alarm area is necessary to reduce the miss rate of the forecast.

In the Northern Japan area we see a similar result. For example, figures 1a and 1b show that the proposed system is able to identify the two earthquake clusters more precisely than the RI, who casts a wide net, reducing forecast accuracy. The ASS score for this area is higher, which indicates both methods were able to learn the two "hot spots" for seismic activity.

We can also see that the results changed wildly in the aftermath of the 2011 M9 earthquake. That event caused a sudden large spike in seismic activity in both areas, including areas that never showed any seismic activity during the training period. In the following scenario (2012), both methods try to use this new data to reform their forecasts.

4. Discussion

Our initial objective was to test the feasibility of evolutionary approaches for common problems in the study of earthquakes. In this work, we presented a GA system for the generation of prediction model based on catalog data. While this system was very simple, it was able to perform competitively with a well know geophysical model.

Although we were concerned with the number of parameters, and the complex nature of the problem, these initial results indicate that there is promise in the use of Evolutionary Computation for this application field. The mechanisms of earthquake generation are still not fully understood, which motivates us to use self-adaptive methods such as Genetic Algorithms.

That said, our initial approach has highlighted several places where the system could be improved. Most important, the current fitness function has several problems, such as the fact that the quality of a bin does not take into account information from neighboring bins, and a general tendency of the system to overfit the data.

One way that we expect to mitigate that is by making the algorithm aware of data locality. While the RI algorithm uses a fixed smoothing pattern to make a high seismicity bin increase the forecast of neighborhood bins, we plan to develop a self-adaptive way of reaching the same goal. We are also very interested in finding ways to add domain knowledge into the system, such as the location of known faults, in order to improve the forecast ability.

5. Acknowledgments

We thank the Japan Meteorological Agency for providing the earthquake catalog used in this study.

References

[1] A. F. Cabalar and A. Cevik. Genetic programming-based attenuation relationship: An application of recent earthquakes in turkey. *Computers and Geosciences*, 35:1884–1896, October 2009.

[2] B. L. N. Kennet and M. S. Sambridge. Earthquake location genetic algorithms for teleseisms. *Physics of the Earth and Planetary Interiors*, 75(1–3):103–110, December 1992.

[3] T. Kerh, D. Gunaratnam, and Y. Chan. Neural computing with genetic algorithm in evaluating potentially hazardous metropolitan areas result from earthquake. *Neural Comput. Appl.*, 19(4):521–529, June 2010.

[4] E. Kermani, Y. Jafarian, and M. H. Baziar. New predictive models for the v_{max}/a_{max} ratio of strong ground motions using genetic programming. *International Journal of Civil Engineering*, 7(4):236–247, December 2009.

[5] K. Z. Nanjo. Earthquake forecasts for the csep japan experiment based on the ri algorithm. *Earth Planets Space*, 63:261–274, 2011.

[6] K. Z. Nanjo, H. Tsuruoka, N. Hirata, and T. H. Jordan. Overview of the first earthquake forecast testing experiment in japan. *Earth Planets Space*, 63:159–169, 2011.

[7] A. Nicknam, R. Abbasnia, Y. Eslamian, M. Bozorgnasab, and E. A. Mosabbeb. Source parameters estimation of 2003 bam earthquake mw 6.5 using empirical green's function method, based on an evolutionary approach. *J. Earth Syst. Sci.*, 119(3):383–396, June 2010.

[8] D. Schorlemmer, M. Gerstenberger, S. Wiemer, D. Jackson, and D. A. Rhoades. Earthquake likelihood model testing. *Seismological Research Letters*, 78(1):17–29, 2007.

[9] J. D. Zechar and T. H. Jordan. The area skill score statistic for evaluating earthquake predictability experiments. *Pure and Applied Geophysics*, 167(8–9):893–906, August 2010.

[10] F. Zhou and X. Zhu. Earthquake prediction based on lm-bp neural network. In X. Liu and Y. Ye, editors, *Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 1*, volume 270 of *Lecture Notes in Electrical Engineering*, pages 13–20. Springer Berlin Heidelberg, 2014.