

---

# Using Genetic Algorithms To Find Temporal Patterns Indicative Of Time Series Events

---

Richard J. Povinelli

Department of Electrical and Computer Engineering  
Marquette University, P.O. Box 1881, Milwaukee, WI 53201-1881, USA  
e-mail: Richard.Povinelli@Marquette.edu; www: http://povinelli.eece.mu.edu

## Abstract

A new framework for analyzing time series data called Time Series Data Mining (TSDM) is introduced. This framework adapts and innovates data mining concepts to analyzing time series data. In particular, it creates methods that reveal hidden temporal patterns that are characteristic and predictive of time series events. The TSDM framework, concepts, and methods, which use a genetic algorithm to search for optimal temporal patterns, are explained and the results are applied to real-world time series from the engineering and financial domains.

## 1 INTRODUCTION

The Time Series Data Mining (TSDM) framework is a fundamental contribution to the fields of time series analysis and data mining (Povinelli 1999). Methods based on the TSDM framework can successfully characterize and predict complex, nonperiodic, irregular, and chaotic time series. The TSDM methods overcome limitations (including stationarity and linearity requirements) of traditional time series analysis techniques by adapting data mining concepts for analyzing time series.

A time series  $X$  is “a sequence of observed data, usually ordered in time.” (Pandit and Wu 1983, p. 1)  $X = \{x_t, t = 1, \dots, N\}$ , where  $t$  is a time index, and  $N$  is the number of observations. Time series analysis is fundamental to engineering, scientific, and business endeavors. It may be applied to the prediction of welding droplet releases and stock market price fluctuations (Povinelli 1999; Povinelli and Feng 1998; Povinelli and Feng 1999a).

The novel TSDM framework has its underpinnings in several fields. It builds upon concepts from data mining (Fayyad et al. 1996), time series analysis (Pandit and Wu 1983; Weigend and Gershenfeld 1994), adaptive signal processing, genetic algorithms (Goldberg 1989; Povinelli and Feng 1999b), and chaos, nonlinear dynamics, and dynamical systems (Abarbanel 1996; Iwanski and Bradley 1998). From data mining comes the focus on discovering hidden patterns. Building on concepts from both adaptive

signal processing and wavelets, the idea of a temporal pattern is developed. From genetic algorithms comes a robust and easily applied optimization method (Goldberg 1989). From the study of chaos, nonlinear dynamics, and dynamical systems comes the theoretical justification of the method, specifically Takens’ Theorem (Takens 1980) and Sauer’s extension (Sauer et al. 1991).

## 2 PROBLEM STATEMENT

Figure 1 illustrates a TSDM problem, where the horizontal axis represents time, and the vertical axis represents observations. The diamonds show the time series observations, and the squares indicate observations that are deemed important – events. Although the following examples illustrate events as single observations, events are not restricted to be instantaneous. The goal is to characterize and predict when important events will occur. The time series events in Figure 1 are nonperiodic, irregular, and contaminated with noise.

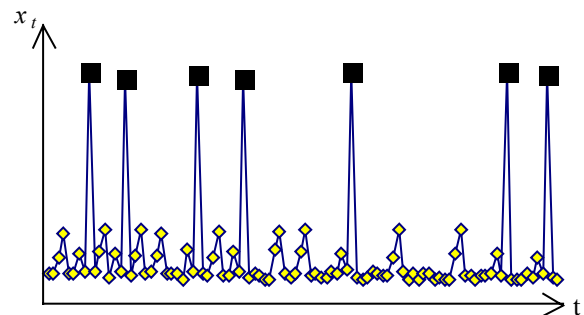


Figure 1: Synthetic Seismic Time Series

To make the time series more concrete, consider it a measure of seismic activity, which is generated from a randomly occurring temporal pattern, a synthetic earthquake, and a contaminating noise signal. The goal is to characterize when peak seismic activity (earthquakes) occurs and then use the characterizations of the activity for prediction.

### 3 DATA MINING

Weiss and Indurkha define data mining as “the search for valuable information in large volumes of data. Predictive data mining is a search for very strong patterns in big data that can generalize to accurate future decisions.” (Weiss and Indurkha 1998) Data mining evolved from several fields, including machine learning, statistics, and database design (Weiss and Indurkha 1998). It uses techniques such as clustering, association rules, visualization, decision trees, nonlinear regression, and probabilistic graphical dependency models to identify novel, hidden, and useful structures in large databases (Fayyad et al. 1996; Weiss and Indurkha 1998).

Others who have applied data mining concepts to finding patterns in time series include Berndt and Clifford (Berndt and Clifford 1996), Keogh (Keogh 1997; Keogh and Smyth 1997; Keogh and Pazzani 1998), and Rosenstein and Cohen (Rosenstein and Cohen 1999). Berndt and Clifford use a dynamic time warping technique taken from speech recognition. Their approach uses a dynamic programming method for aligning the time series and a predefined set of templates.

Rosenstein and Cohen (Rosenstein and Cohen 1999) also use a predefined set of templates to match a time series generated from robot sensors. Instead of using the dynamic programming methods as in (Berndt and Clifford 1996), they employ the time-delay embedding process to match their predefined templates.

Similarly, Keogh represents the templates using piecewise linear segmentations. “Local features such as peaks, troughs, and plateaus are defined using a prior distribution on expected deformations from a basic template.” (Keogh and Smyth 1997) Keogh’s approach uses a probabilistic method for matching the known templates to the time series data.

Other approaches to time series prediction include using genetic programming to learn the nonlinear generating function (Howard and Oakley 1994; Kaboudan to appear) and evolving a recurrent neural network (Torreale 1991) to learn temporal sequences.

The TSDM framework, initially introduced by Povinelli and Feng in (Povinelli and Feng 1998), differs fundamentally from these approaches. The approach advanced in (Berndt and Clifford 1996; Keogh 1997; Keogh and Smyth 1997; Keogh and Pazzani 1998; Rosenstein and Cohen 1999) requires *a priori* knowledge of the types of structures or temporal patterns to be discovered and represents these temporal patterns as a set of templates. Their (Berndt and Clifford 1996; Keogh 1997; Keogh and Smyth 1997; Keogh and Pazzani 1998; Rosenstein and Cohen 1999) use of predefined templates prevents the achievement of the basic data mining goal of discovering useful, novel, and hidden temporal patterns.

The next section introduces the key TSDM concepts, which allow the TSDM methods to overcome the limitations of traditional time series methods and the more recent approaches of Berndt and Clifford (Berndt and

Clifford 1996), Keogh (Keogh 1997; Keogh and Smyth 1997; Keogh and Pazzani 1998), and Rosenstein and Cohen (Rosenstein and Cohen 1999).

### 4 SOME CONCEPTS IN TIME SERIES DATA MINING

The fundamental TSDM concepts are event, temporal pattern, event characterization function, temporal pattern cluster, time-delay embedding, phase space, augmented phase space, objective function, and optimization.

In a time series, an event is an important occurrence. The definition of an event is dependent on the TSDM goal. In a seismic time series, an earthquake is defined as an event. Other examples of events include sharp rises or falls of a stock price or the release of a droplet of metal from a welder. A temporal pattern is a hidden structure in a time series that is characteristic and predictive of events. The temporal pattern  $\mathbf{p}$  is a real vector of length  $Q$ . The temporal pattern is represented as a point in a  $Q$  dimensional real metric space, i.e.,  $\mathbf{p} \in \mathbb{R}^Q$ . A temporal pattern cluster, the neighborhood of a temporal pattern, is defined as the set of all points within  $\delta$  of the temporal pattern.

$$P = \{a \in \mathbb{R}^Q : d(\mathbf{p}, a) \leq \delta\}, \quad (1)$$

where  $d$  is the distance or metric defined on the space. This defines a hypersphere of dimension  $Q$ , radius  $\delta$ , and center  $\mathbf{p}$ .

A reconstructed phase space (Abarbanel 1996; Iwanski and Bradley 1998), called simply phase space here, is a  $Q$ -dimensional metric space into which a time series is embedded. Takens showed that if  $Q$  is large enough, the phase space is homeomorphic to the state space that generated the time series (Takens 1980). The time-delayed embedding of a time series maps a set of  $Q$  time series observations taken from  $X$  onto  $\mathbf{x}_t$ , where  $\mathbf{x}_t$  is a vector or point in the phase space. Specifically,  $\mathbf{x}_t = (x_{t-(Q-1)\tau}, \dots, x_{t-2\tau}, x_{t-\tau}, x_t)$ .

To link a temporal pattern (past and present) with an event (future) the “gold” or event characterization function  $g(t)$  is introduced. The event characterization function represents the value of future “eventness” for the current time index. It is, to use an analogy, a measure of how much gold is at the end of the rainbow (temporal pattern). The event characterization function is defined such that its value at  $t$  correlates highly with the occurrence of an event at some specified time in the future, i.e., the event characterization function is causal when applying the TSDM method to prediction problems. Non-causal event characterization functions are useful when applying the TSDM method to system identification problems.

One possible event characterization function to address this goal is  $g(t) = x_{t+1}$ , which captures the goal of characterizing synthetic earthquakes one-step in the future. Alternatively, predicting an event three time-steps

ahead requires the event characterization function  $g(t) = x_{t+3}$ .

The concept of an augmented phase space follows from the definitions of the event characterization function and the phase space. The augmented phase space is a  $Q+1$  dimensional space formed by extending the phase space with  $g(\cdot)$  as the extra dimension. Every augmented phase space point is a vector  $\langle \mathbf{x}_t, g(t) \rangle \in \mathbb{R}^{Q+1}$ . The augmented phase space is illustrated in Figure 2.

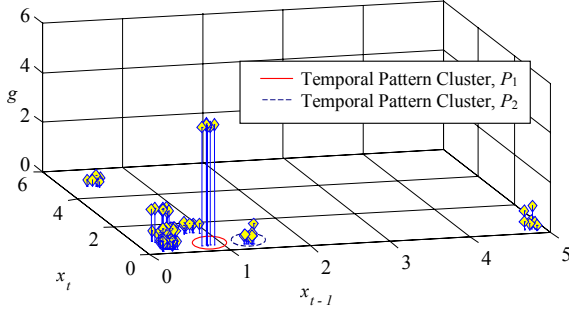


Figure 2: Synthetic Seismic Augmented Phase Space with Highlighted Temporal Pattern Clusters

The next concept is the TSDM objective function, which represents the efficacy of a temporal pattern cluster to characterize events. The objective function  $f$  maps a temporal pattern cluster  $P$  onto the real line, which provides an ordering to temporal pattern clusters according to their ability to characterize events. The objective function is constructed in such a manner that its optimizer  $P^*$  meets the TSDM goal.

The form of the objective functions is application dependent, and several different objective functions may achieve the same TSDM goal. Before presenting an example objective function, several definitions are required.

The index set  $\Lambda = \{t : t = (Q-1)\tau + 1, \dots, N\}$  is the set of all time indices  $t$  of phase space points, where  $(Q-1)\tau$  is the largest embedding time-delay, and  $N$  is the number of observations in the time series. The index set  $M$  is the set of all time indices  $t$  when  $\mathbf{x}_t$  is within the temporal pattern cluster, i.e.  $M = \{t : \mathbf{x}_t \in P, t \in \Lambda\}$ . Similarly,  $\tilde{M}$ , the complement of  $M$ , is the set of all time indices  $t$  when  $\mathbf{x}_t$  is outside the temporal pattern cluster.

The average value of  $g$ , also called the average eventness, of the phase space points within the temporal pattern cluster  $P$  is

$$\mu_M = \frac{1}{c(M)} \sum_{t \in M} g(t)$$

where  $c(M)$  is the cardinality of  $M$ . The average eventness of the phase space points not in  $P$  is

$$\mu_{\tilde{M}} = \frac{1}{c(\tilde{M})} \sum_{t \in \tilde{M}} g(t).$$

The corresponding variances are

$$\sigma_M^2 = \frac{1}{c(M)} \sum_{t \in M} (g(t) - \mu_M)^2, \text{ and}$$

$$\sigma_{\tilde{M}}^2 = \frac{1}{c(\tilde{M})} \sum_{t \in \tilde{M}} (g(t) - \mu_{\tilde{M}})^2.$$

Using these definitions, an objective function based on the  $t$  test for the difference between two independent means is defined below.

$$f(P) = \frac{\mu_M - \mu_{\tilde{M}}}{\sqrt{\frac{\sigma_M^2}{c(M)} + \frac{\sigma_{\tilde{M}}^2}{c(\tilde{M})}}},$$

where  $P$  is a temporal pattern cluster. This objective function is useful for identifying temporal pattern clusters that are statistically significant and have a high average eventness.

The key concept of the TSDM framework is to find optimal temporal pattern clusters that characterize and predict events. Thus, an optimization algorithm represented by  $\max_{\mathbf{p}, \delta} f(P)$  is necessary.

#### 4.1 OPTIMIZATION METHOD – GENETIC ALGORITHM

The simple genetic algorithm is adapted to the TSDM framework. These adaptations include an initial Monte Carlo search and hashing of fitness values. The genetic algorithm is described as follows.

- Create an elite population
- Randomly generate large population (n times normal population size)
- Calculate fitness
- Select the top 1/n of the population to continue
- While all fitnesses have not converged
  - Selection
  - Crossover
  - Mutation
  - Reinsertion

Initializing the genetic algorithm with the results of a Monte Carlo search has been found to help the optimization's rate of convergence and in finding a good optimum.

Typical population sizes range from 30 to 100. Both roulette and tournament selection have been used, but tournament selection has been found to be more adaptable because of its ability to handle negative fitness values. One point crossover is used with a uniformly random locus. Mutation rates range from 0 to 0.05. An elitism of one is typically employed.

A temporal pattern cluster, which is composed of the temporal pattern  $\mathbf{p}$  of dimension  $Q$  and radius  $\delta$ , is encoded into a binary string. Each component of the temporal pattern cluster is usually encoded with six to eight bits. Thus, each chromosome for a  $Q$  dimensional temporal pattern cluster is formed from  $(Q+1)n$  bits,

where  $n$  is the number of bit used to encode each component.

The hashing modification reduces the computation time of the genetic algorithm by 50%. This modification is discussed in detail in (Povinelli and Feng 1999b). Profiling the computation time of the genetic algorithm reveals that most of the computation time is used evaluating the fitness function. Because the diversity of the chromosomes diminishes as the population evolves, the fitness values of the best individuals are frequently recalculated. Efficiently storing fitness values in a hash table dramatically improves genetic algorithm performance (Povinelli and Feng 1999b).

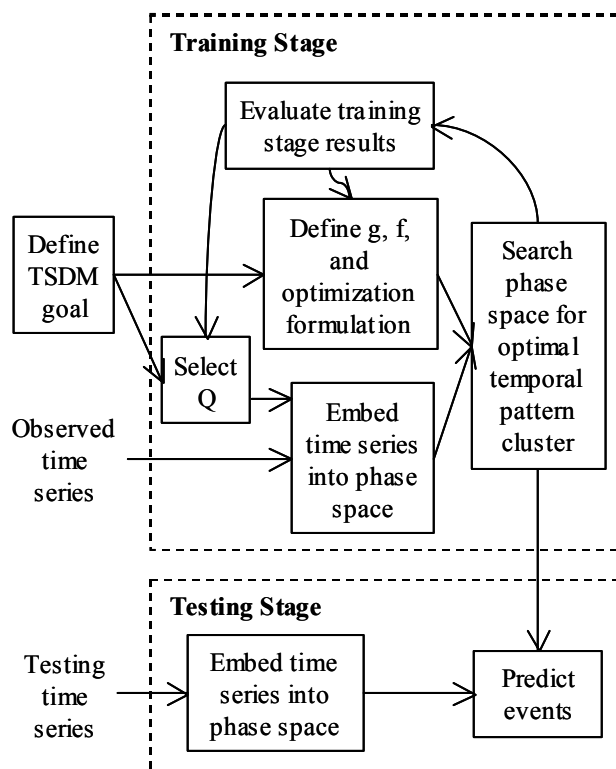


Figure 3: Block Diagram of TSDM Method

## 5 FUNDAMENTAL TIME SERIES DATA MINING METHOD

The first step in applying the TSDM method is to define the TSDM goal, which is specific to each application, but may be stated generally as follows. Given an observed time series  $X = \{x_t, t = 1, \dots, N\}$ , the goal is to find hidden temporal patterns that are characteristic of events in  $X$ , where events are specified in the context of the TSDM goal. Likewise, given a testing time series  $Y = \{x_t, t = R, \dots, S\}$   $N < R < S$  the goal is to use the hidden temporal patterns discovered in  $X$  to predict events in  $Y$ . The method is detailed in Figure 3.

The key to the TSDM method is that it forgoes the need to characterize time series observations at all time indices for the advantages of being able to identify the optimal local temporal pattern clusters for predicting important events. This allows prediction of complex real-world time series using small-dimensional phase spaces.

The results of the application of the method to the synthetic seismic time series are illustrated in Figure 4. The temporal pattern cluster discovered in the training phase is applied to subsequences of the testing time series. Using (1), events are detected by determining if embedded subsequences of length  $Q$  are in  $P$ . The pair of connected gray squares that match sequences of time series observations before events is the temporal pattern. The black squares indicate predicted events.

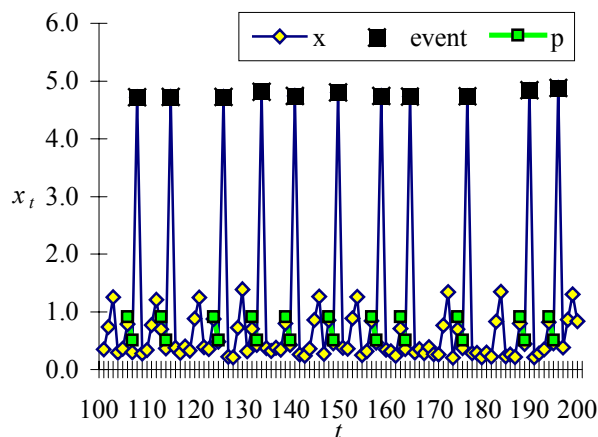


Figure 4: Synthetic Seismic Time Series with Temporal Patterns and Events Highlighted (Testing)

## 6 APPLICATIONS AND CONCLUSIONS

Two problems to which the TSDM method has been applied are briefly discussed here. The first is the prediction when a droplet of metal will release from a welder (Povinelli 1999). Using the time series generated from three sensors, a prediction accuracy of over 96% was achieved. The three sensors measured the voltage, current, and droplet stickout length. The second application is to the prediction of stock price increases (Povinelli 1999). The TSDM method was applied to the 30 Dow Jones Industrial components. It was able to achieve a return of 30% versus of 3% baseline return.

TSDM methods have been successfully applied to characterizing and predicting complex, nonstationary, chaotic time series events from both the engineering and financial domains. Given a multi-dimensional time series generated by sensors on a welding station, the TSDM framework was able to, with a high degree of accuracy, characterize and predict metal droplet releases. In the financial domain, the TSDM framework was able to generate a trading-edge by characterizing and predicting stock price events.

## References

- Abarbanel, H. D. I. (1996). *Analysis of observed chaotic data*, Springer, New York.
- Berndt, D. J., and Clifford, J. (1996). "Finding Patterns in Time Series: A Dynamic Programming Approach." *Advances in knowledge discovery and data mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthursamy, eds., AAAI Press, Menlo Park, California, 229-248.
- Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthursamy, R. (1996). *Advances in knowledge discovery and data mining*, AAAI Press, Menlo Park, California.
- Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley, Reading, Massachusetts.
- Howard, E., and Oakley, N. (1994). "The Application of Genetic Programming to the Investigation of Short, Noisy, Chaotic Data Series." *AISB Workshop*, Leeds, U.K., 320-332.
- Iwanski, J., and Bradley, E. (1998). "Recurrence plot analysis: To embed or not to embed?" *Chaos*, 8(4), 861-871.
- Kaboudan, M. (to appear). "Genetic Programming Prediction of Stock Prices." *Computational Economics*.
- Keogh, E. (1997). "A Fast and Robust Method for Pattern Matching in Time Series Databases." *9th International Conference on Tools with Artificial Intelligence (TAI '97)*.
- Keogh, E., and Smyth, P. (1997). "A Probabilistic Approach to Fast Pattern Matching in Time Series Databases." *Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, California.
- Keogh, E. J., and Pazzani, M. J. (1998). "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback." *AAAI Workshop on Predicting the Future: AI Approaches to Time-Series Analysis*, Madison, Wisconsin.
- Pandit, S. M., and Wu, S.-M. (1983). *Time series and system analysis, with applications*, Wiley, New York.
- Povinelli, R. J. (1999). "Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of Time Series Events," Ph.D. Dissertation, Marquette University, Milwaukee.
- Povinelli, R. J., and Feng, X. (1998). "Temporal Pattern Identification of Time Series Data using Pattern Wavelets and Genetic Algorithms." *Artificial Neural Networks in Engineering*, St. Louis, Missouri, 691-696.
- Povinelli, R. J., and Feng, X. (1999a). "Data Mining of Multiple Nonstationary Time Series." *Artificial Neural Networks in Engineering*, St. Louis, Missouri, 511-516.
- Povinelli, R. J., and Feng, X. (1999b). "Improving Genetic Algorithms Performance By Hashing Fitness Values." *Artificial Neural Networks in Engineering*, St. Louis, Missouri, 399-404.
- Rosenstein, M. T., and Cohen, P. R. (1999). "Continuous Categories For a Mobile Robot." *Sixteenth National Conference on Artificial Intelligence*.
- Sauer, T., Yorke, J. A., and Casdagli, M. (1991). "Embedology." *Journal of Statistical Physics*, 65(3/4), 579-616.
- Takens, F. (1980). "Detecting strange attractors in turbulence." *Dynamical Systems and Turbulence*, Warwick, 366-381.
- Torreele, J. (1991). "Temporal Processing with Recurrent Networks: An Evolutionary Approach." *Fourth International Conference on Genetic Algorithms*, 555-561.
- Weigend, A. S., and Gershenfeld, N. A. (1994). *Time Series Prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley Pub. Co., Reading, MA.
- Weiss, S. M., and Indurkha, N. (1998). *Predictive data mining: a practical guide*, Morgan Kaufmann, San Francisco.