TEMPORAL PATTERN IDENTIFICATION OF TIME SERIES DATA USING PATTERN WAVELETS AND GENETIC ALGORITHMS

RICHARD J. POVINELLI AND XIN FENG
Department of Electrical and Computer Engineering
Marquette University, P.O. Box 1881, Milwaukee, WI 53201-1881, USA
E-mail: povinellir@mu.edu
Ph: 414.288.6820
Fx: 414.288.5579

ABSTRACT:
A new method for temporal pattern matching of a time series is developed using pattern wavelets and genetic algorithms. The pattern wavelet is applied to the matching of an embedded time series. A problem-specific fitness factor is introduced in the new algorithm, which is useful to construct a fitness function of the feature space. A two-step process discovers the pattern wavelet that yields high fitness value. The best temporal pattern matches are found through a thresholding process. These matches are kept and the future time series data point is used in the genetic algorithm's fitness function. The algorithm has been successfully applied to the identification of statistically significant temporal patterns in financial time series data.

Keywords: Temporal Pattern Identification, Genetic Algorithms, Pattern Recognition, Time Series Analysis, Wavelets

INTRODUCTION

Data mining is the exploration of data with the goal of discovering hidden structure. In many real-world applications, it is important to study the change of temporal features of a non-stationary time series, and identify the ones that are representing the significance of time instances. For example, it is critical in stock market applications that the patterns relating to sudden stock price changes be identified. Generally such time series are considered non-stationary. Traditional time series analysis employs statistical methods to model and explain the data and predict future values of the time series. It is not easy, however, to identify the critical temporal patterns of the time series using these traditional methods.

Using a set of observations, in this paper, we present a new method for time series data mining. By introducing a pattern wavelet along with the use of a genetic algorithm (GA), temporal patterns can be effectively revealed in non-stationary time series.

The paper is organized as follows. After presenting the problem statement, traditional ARMA modeling is reviewed. The ideas of temporal pattern matching

and the pattern wavelet are then discussed. Next, a detailed discussion of the new algorithm is provided. Finally, a presentation of the results and conclusions is given.

PROBLEM STATEMENT

Let $Z = \{z_t, t = 1,\ldots, N\}$ be the non-stationary target time series, whose temporal features evolve over time. The task is to find an approach to characterize these changing temporal features.

Applying traditional time series modeling to this problem involves finding solutions to the Box-Jenkins difference equation (Bowerman and O'Connell 1993) .

$$\phi_p(B)z_t = \delta + \theta_q(B)a_t,$$

where $\phi_p(B)$ is the nonseasonal autoregressive operator of order p, $\theta_q(B)$ is the nonseasonal moving average operator of order q, z_t is the time series, a_t is a sequence of random variables, δ is a constant term, and B is the backshift operator. The Box-Jenkins method is limited by the requirement of stationarity of the time series and normality and independence of the residuals. However, in most applications, these conditions are not met. One of the most severe drawbacks of this approach is the loss of the non-stationary characteristics we desire to identify.

Our method takes a new approach. Let

$$\mathbf{z}_t^T = \left(z_t,\ldots,z_{t+Q-1}\right), \qquad t = 1,\ldots, N-Q+1$$

be the set of sub-time series of length Q embedded in Z, where $Q \le N$. Clearly, $\mathbf{z}_t \subseteq Z$, which may represent the changing temporal features or patterns of Z. We propose that by studying the embedding $\mathbf{z}_t$, the temporal features of Z may be identified.

The method for eliciting the temporal features from the embedding $\mathbf{z}_t$ arises from a study of wavelets and the wavelet transform. The wavelet transform is a natural extension of Fourier's work done in the early 19th century. Where Fourier's transform can find frequency information with no time reference or time information with no frequency, the wavelet transform provides both time and frequency information.

Generally speaking, the wavelet transform matches a compactly supported function, called a wavelet, across both scale (frequency) and translation (time) (Polikar 1996). The Fourier transform matches an infinitely supported function across frequency (scale). Both use convolution of the basis function and the original time series. For the wavelet transform, it is provided for all scales.

Next we introduce the so called “pattern wavelet” and “pattern wavelet transform”. This transform is an extension of a discrete form of the wavelet transform applied specifically to identifying temporal features.

PATTERN WAVELETS

By relaxing the restrictions of the wavelet transform, the pattern wavelet transform is derived. Where the wavelet transform uses the convolution of the wavelet and the

time series, the pattern wavelet transform uses a subset of the convolution of the pattern wavelet and the time series. Also, where the wavelet is required to have a zero mean, the pattern wavelet is not. These relaxations yield a transform that identifies the temporal features discussed in the problem statement. A detailed explanation of the algorithm follows.

Let $f(\mathbf{p},\delta,Z,g)$ be the pattern wavelet transform, where $\mathbf{p} \in P \subseteq \Re^Q$ is the pattern wavelet, $\delta \in \Re$ is a threshold parameter, and $g = g(z_t)$ is a measure of fitness of the temporal feature. We want to find the optimal solution to the following problem

$$\max_{\mathbf{p},\delta} \left\{ f(\mathbf{p},\delta,Z,g) \;\middle|\; \mathbf{p} \in P \subseteq \Re^Q, \delta \in \Re \right\}. \tag{1}$$

The pattern wavelet transform $f(\mathbf{p},\delta,Z,g)$ is the fitness of pattern $\mathbf{p}$ with threshold δ applied to time series Z with fitness measure g. The following definitions are needed for f.

$$\begin{aligned} r_t &= \langle \mathbf{p}, \mathbf{z}_t \rangle, \qquad t = 1,\ldots, N-Q+1 \\ \mu_r &= \frac{1}{N-Q+1} \sum_{t=1}^{N-Q+1} r_t \\ \sigma_r^2 &= \frac{1}{N-Q+1} \sum_{t=1}^{N-Q+1} (r_t - \mu_r)^2 \\ M &= \{ t:\ r_t \geq \mu_r + \delta\sigma_r \} \end{aligned}$$

The vector $\mathbf{z}_t \subseteq Z$ is the embedded series of length Q, where $Q \leq N$. The pattern factors r_t, $t = 1,\ldots, N\text{-}Q+1$, are elements of the vector $\mathbf{r} \in \Re^{N-Q+1}$ which consists of $N\text{-}Q+1$ inner products of the pattern wavelet $\mathbf{p}$ and the embedded time series $\mathbf{z}_t$. Also μ_r denotes the mean of r_t, σ_r is the standard deviation of r_t, and M is the pattern match set, which is defined as the set of all time instances t where the pattern factor r_t is greater than or equal to the threshold $\mu_r + \delta\sigma_r$. Finally, the pattern wavelet transform f is defined as the mean of $g(z_t)$ for $t \in M$.

$$f(\mathbf{p},\delta,Z,g) \equiv \mu_M = \frac{1}{c(M)} \sum_{t \in M} g(z_t) \tag{2}$$

where $c(M)$ is the cardinality of M. Also σ_M is the standard deviation of $g(z_t)$ at times $t \in M$.

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

It should be noted that the selection of fitness operator g in (2) is problem specific and is independent of the algorithm. It should be chosen a priori based on the types of hidden temporal features to be discovered.

Because the maximization problem in (1) is complex and nonlinear, it is difficult to solve using traditional numerical optimization methods. To overcome these limitations, a roulette wheel based GA with elitism (Goldberg 1989) searches for the optimal **p** and δ. Ideally $\mathbf{p} \in \Re^Q$ and $\delta \in \Re$, for efficiency purposes $\mathbf{p} \in [-\varepsilon, \varepsilon]^Q$ and $\delta \in [\delta_1, \delta_2]$. These ranges are discrete due to the nature of the GA with a possible 2^b unique values, where b is the number of bits used to represent p_i and δ. The parameters for the GA are Q, Z, g, b, and the population size. The parameter b is usually in the range of 4 to 16 and the population size is set to 30. The most elite individual is maintained from generation to generation without change. No mutation is used. The GA is shown below.

Pattern Finding Genetic Algorithm

1. Create an elite population
   a) Randomly generate large population (10 times normal population size)
   b) Calculate fitness
   c) Select the top 10th of the population to continue
2. While all fitness have not converged
   a) Perform roulette selection, save elite individual
   b) Crossover population
   C)Calculate fitness

APPLICATION RESULTS

The goal of this application is to find hidden temporal patterns in a certain stock time series. Our experimental time series is the daily open stock price of the Quantum (QNTM, traded on the NASDAQ) time series $Z = \{z_t, t = 1,\ldots, N\}$ with N=3,761. See Figure 1 for illustration. Obviously, this time series is non-stationary. Our special interest is to identify the temporal pattern that is related to a significant price change.

ARMA Model

Two ARMA models of the time series reveal essentially the same random walk characteristics. The models are

Figure 1- Quantum Corp stock time series

$$\hat{z}_t = \hat{\phi} z_{t-1} + \varepsilon_t \tag{3}$$

$$\hat{z}_t = \frac{z_{t-1}^{1+\hat{\phi}}}{z_{t-2}^{\hat{\phi}}} + \varepsilon_t \tag{4}$$

$$\hat{z}_t = z_{t-1} + \varepsilon_t \tag{5}$$

where $\hat{\phi}$ = 0.99933 in (3) and $\hat{\phi}$ = 0.045948 in (4). The $\hat{\phi}$ in both models is statistically significant, but the autocorrelations of (3) show strong evidence of non-stationarity and the Ljung-Box test of the residuals indicates a lack of independence. The model (4) Ljung-Box test of the residuals indicates independence. By seeing that the $\hat{\phi} \cong 1$ in (3) and $\hat{\phi} \cong 0$ in (4), both models become equivalent (5).

The ARMA models provide little insight into hidden structure in the time series; the series is a random walk. On the other hand the method presented by the authors finds statistically significant structure as presented below.

Pattern Wavelet Model

In building the pattern wavelet model, the fitness operator *g* in (2) is chosen as

$$g(z_t) = \frac{B^{-Q} - B^{-(Q+1)}}{B^{-Q}} z_t .$$

In our case we want to find features that indicate a fit Δ% after the end of the pattern match.

We found $c(M)$ to be between 138 and 314, depending on the support of the pattern wavelet. The statistics for eight patterns are given in Table 1. The change in the stock price after a pattern match was between +0.7% and +1.5%, whereas the average change was +0.12%. This shows that there is a correlation between the patterns and the price changes. The standard deviation, though, is between 3% and 4% for the patterns and 3% for the average day. The μ_M of the matched patterns is between 5 to 12 higher than $\mu_{g(Z)}$ of the whole time series. Two statistical tests are used to show significance of the results. The first test is the runs test. The test hypothesis is H_0: There is no difference between the matched time series and the remaining time series. H_A: There is significant difference between the matched time series and the remaining time series. Our test uses a 1% probability of Type I error ($\alpha = 0.01$). Table 1 shows that the null hypothesis can easily be rejected in all cases.

The second statistical test is the difference of two independent means. The two populations are the transformed series and the whole time series. Although the two populations are probably dependent, this can be ignored because it makes the statistics more conservative, i.e., it will tend to overestimate the Type I error. The test hypothesis is H_0: $\mu_M - \mu_{g(Z)} = 0$, H_A: $\mu_M - \mu_{g(Z)} > 0$. This test uses a 1% probability of Type I error ($\alpha = 0.01$). Again, Table 1 shows that the null hypothesis can be very confidently rejected for all the patterns. The mean fitness of the time series $\mu_{g(Z)} = 0.001179$, and the $\sigma_{g(Z)} = 0.032931$.

TABLE 1 – STATISTICAL SIGNIFICANCE OF RESULTS					
Q	c(*M*)	μ_M	σ_M	Runs test α	means test α
1	238	0.00736	0.0385	$< 1.00 \times 10^{-17}$	8.81×10^{-3}
2	167	0.00834	0.0375	$< 1.00 \times 10^{-17}$	7.58×10^{-3}
3	357	0.00746	0.0336	$< 1.00 \times 10^{-17}$	3.64×10^{-4}
4	185	0.00913	0.0417	4.78×10^{-10}	5.30×10^{-3}
19	201	0.01057	0.0416	$< 1.00 \times 10^{-17}$	8.28×10^{-4}
21	144	0.01397	0.0362	$< 1.00 \times 10^{-17}$	1.51×10^{-5}
27	190	0.01276	0.0406	4.44×10^{-16}	5.55×10^{-5}
39	210	0.01113	0.0348	$< 1.00 \times 10^{-17}$	2.56×10^{-5}

CONCLUSIONS

In this paper, a new method for temporal data mining is proposed. Using a pattern wavelet transform as a data mining tool has yielded meaningful results. Instead of forcing the wavelet to match everywhere, it matches only when there is a high similarity between the pattern wavelet and the underlying time series. To find such pattern wavelets, a genetic algorithm is used. Even with a complex, non-stationary time series like stock price, the algorithm detected interesting patterns. Across all tested Q the patterns found were statistically significant.

The algorithm is flexible in that by using an alternative g, fitness function, different structures can be found. The g used in this research was for positive changes, but just as easily

$$g(z_t) = -\frac{B^{-Q} - B^{-(Q+1)}}{B^{-Q}} z_t$$

which would find negative changes. Also, a more complicated g could be used that could take into account the standard deviations of the matches.

Future research directions will include exploring combinations of patterns, looking for patterns in shorter segments of the time series, and adding additional factor dimensions such as volume.

REFERENCES

Bowerman, B. L., and O'Connell, R. T. (1993). *Forecasting and Time Series: An Applied Approach*, Duxbury Press, Belmont, California.

Ghoshray, S. (1996). "Hybrid prediction technique by fuzzy inferencing on the chaotic nature of time series data." *Artificial Neural Networks in Engineering, Proceedings*, 725-730.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*, Addison-Wesley Pub. Co., Reading, Mass.

Lin, C. T., and Lee, C. S. G. (1996). *Neural Fuzzy Systems - A Neuro-Fuzzy Synergism to Intelligent Systems*, Prentice-Hall, Upper Saddle River, NJ.

Polikar, R. (1996). "The Engineer's Ultimate Guide To Wavelet Analysis - The Wavelet Tutorial." .

Weigend, A. S., and Gershenfeld, N. A. (1994). "Time Series Prediction: Forecasting the Future and Understanding the Past." , Addison-Wesley Pub. Co., Reading, MA.