Prof. Rupesh Mahajan, Prof. Arivanantham Thangavelu, Prof. Minal Shahakar / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 6, November- December 2012, pp.185-187

Data Mining Techniques for Identifying Temporal Patterns of Time Series Data

Prof. Rupesh Mahajan, Prof. Arivanantham Thangavelu, Prof. Minal Shahakar
Department of IT, Pad. Dr.DYPIET, Pune -18
Department of IT, Pad. Dr.DYPIET, Pune -18
Department of Computer,Pad. Dr.DYPIET, Pune -18

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. It creates a set of methods with the growing deployment of a large Number of sensors, telemetry devices and that reveals hidden temporal patterns that are characteristic and predictive of time series events. Traditional time series analysis methods are unable to identify complex (nonperiodic, nonlinear, irregular and chaotic) characteristics. TSDM methods overcome limitations of traditional time series analysis techniques.

1. Introduction

Time series databases consist of sequence of values or events over repeated measurement of time. The values are typically measured at equal time intervals (e.g. hourly, daily, and weekly). Time series databases are popular in many applications, such as stock market analysis, economics and sale forecasting, budgetary analysis, utility studies, inventory studies, work load projections, process and quality control, observation of natural phenomenon (such as atmosphere, temperature, wind, earthquake), scientific and engineering experiments and medical treatment [2].

With the growing deployment of a large number of sensors, telemetry devices, and other on-line data collection tools, the amount of time series data is increasing rapidly, often in order of gigabytes per day (such as stock trading) or even per minute (such as from NASA space programs). It is challenging to analyze such a huge number of time series to find similar or regular patterns, trends and outliers.

2. Example of Business Operations

The restaurant franchise to be described is one of world's largest multi-brand fast-food restaurant chains with more than 30,000 stores worldwide. Due to Information Technology, this restaurant franchise has modernized its business operations at store level using relatively inexpensive pc-based servers, and at corporate level using highly scalable parallel processing architectures.

The data collection involves a POS (Point-of-sale) system. Each time a customer order is placed and the information is keyed into front register, the transaction is automatically processed through the POS system, time stamped and stored in back-office database. Each restaurant tracks all menu items in addition to the ingredients that go into producing the menu items. This yields several hundred time-ordered series. The centralized corporate office also collects higher level data from each individual restaurant on regular basic and stores the information in a data warehouse. The sales and transaction data collected by restaurant chains may be explored and analyzed at both the store level and corporate level. At store level, exploring or mining the large amount of transaction data allows each restaurant to improve its operations management (such as labor scheduling) and product management (such as product preparation scheduling), thereby reducing restaurant operating expenses and increasing food quality. At the corporate level, mining information across the restaurant could greatly facilitate corporate strategic planning [1]. The adaptation of data mining on time series promises to assist the restaurant industry in several ways. Data mining (1) provides a method to process large amount of data in an automated fashion (2) provides a method of distilling vast amount of data into information that is useful for inventory planning, labor scheduling and food preparation planning and (3) offers a consistent, reliable and accurate method of forecasting inventory and product depletion rates over set temporal provides commonly used in business planning [1].

3 High Performance Time Series Analysis

Time series analysis is often associated with the discovery and use of patterns (such as periodicity, seasonality or cycles), and prediction of future values (termed as forecasting in time series context). Traditionally people have tried to build models for time series data and then fit the actual observations of

sequence into these models. If the model is successful in interpreting the observed time series can be predicted provided that the model's assumptions continue to hold in future [3]. As a result of developments in automatic massive data collection and storage technologies, we are living in an age of data explosion. Many applications generate massive amounts of time series data, for example (1) In mission operations for NASA's Space shuttle, approximately 20,000 sensors are telemeter one per second for mission control. (2) In telecommunication, the AT&T long distance data stream consists of approximately 300 million records per day from 100 million customers. As extremely large data sets grow more prevalent in a wide variety of setting, we face the significant challenges of developing more efficient time series analysis method. Following methods are used for flexible time series analysis.

3.1 Data Reduction

Because time series are observations made in sequence, the relationship between consecutive data items in a time series gives
data analyst the opportunity to reduce the size of data without substantial loss of information. It will provide synopsis of the data. Many data reduction techniques can be used for time series data.

3.2 Indexing Method

Indexing Methods provide a way to organize data so that the data with the interested properties can be retrieved efficiently. The indexing methods also organize the data in a way so that the I/O cost can be greatly reduced.

3.3 Transforms on Time Series

To discover patterns from time series, data analysts must be able to compare time series in a scale & magnitude independent way. Hence, shifting and scaling of the time series amplitude, time shifting and scaling of time series and dynamic wrapping of time series are some useful techniques.

4 ARIMA Time Series Analysis

In real life research and practice, patterns of the data are unclear, individual observations involves considerable error, and we still need not only to uncover the hidden patterns in the data but also generate forecasts. The ARIMA (Autoregressive Integrated Moving Average) Methodology developed by Box and Jenkins [5], is use for this purpose. ARIMA is complex technique, it is not easy to use, it requires a great deal of experience, and although it often produces satisfactory results. ARIMA methodology involves finding solutions to the difference equation.

$$\Phi p(B)\ \Phi p\ (BL)\ xt = \partial + \Phi q\ (B)\ \Phi Q\ (BL)\ at$$

- Nonseasonal autoregressive operator Φp (B) of order P models low-order feedback responses.
- The Seasonal autoregressive operator *Φp (BL)* of order P models feedback responses that occur periodically at seasonal intervals.
- The nonseasonal moving average operator Φq *(B)* of order q models low- order weighted average response.
- The Seasonal moving average operator *ΦQ (BL)* of order Q models seasonal weighted average response.
- The term *xt, ∂, at* are the time series, a constant and a sequence of random shocks respectively.

5 Time Series Data Mining Framework

The Time Series Data Mining (TSDM) framework is a fundamental contribution to the fields of time series analysis and data mining. Methods based on TSDM framework are able to successfully characterize and predict complex, non-periodic and irregular time series. The TSDM methods overcome limitations of traditional time series analysis techniques by adapting data mining concepts for analyzing time series. It creates a set of methods that reveal hidden temporal patterns that are characteristic and predictive of time series events [1].

The TSDM framework innovates data mining concepts for analyzing time series data. The TSDM framework focuses on predicting events, which are important occurrences. It is commonly assumed that the ARIMA time series models developed with past data will apply to future prediction. This is stationary assumption that models will not need to very through time. ARIMA model can be defined by linear difference equations, but system generating the time series are not necessary linear or stationary. In contrast, the TSDM framework and methods built upon it can handle nonlinear and non-stationary time series [1].

5.1 Time Series Data Mining Method

The first step in applying the TSDM method is to define the TSDM goal, which specific to each application. The goal is to find hidden temporal patterns that are characteristics of events in time series where events are specified in context of the TSDM goal. Given a TSDM goal, an observed time series to

be characterized, and a testing time series to be predicted, the steps in the TSDM methods are

I Training Stage (Batch Process)

1. Frame the TSDM goal in term of event characterization function, objective function, and optimization formulation
    a. Define the events characterization function g
    b. Define the objective function f.
    c. Define the optimization formulation including independent variable over objective function.
2. Determine Q i.e. the dimension of the phase space and the length of temporal patterns.
3. Transform the observed time series into the phase space using the time delayed embedding process.
4. Associates with each time index in phase space & form the augmented phase space.
5. In the augmented phase space, search for optimal temporal pattern cluster.
6. Evaluate training stage result repeat training stage necessary.

II Testing Stage (Real Time or Batch Process)

1. Embed the testing time series into the phase space.
2. Use the optimal temporal patterns cluster for predicting events.
3. Evaluate testing stage results.

6. Time Series Data Transformation for Classification

Data in Time Series has lots of variations, as in telecom industry, some data sequences are long, consisting over 20 month of billing data, while others are short, consisting of perhaps only three or four months of billing data. Furthermore, each data item in a time series is a multidimensional vector, instead of just a single number as in some stock market analysis data sets. These problems pose particular difficulties for many time series analysis methods, because most standard classification method such as decision tree, maximum like hood and SVM methods requires that the input data be consist of equal length vectors of attributes value pairs. General approach for data transformation is to follow a two-step process. In first step, we transform the data into equal-length vectors. A Key issue is how to maintain as much key temporal information as we can. To this end, we will apply a model based clustering method to help us transform the data. In the second step, we apply and compare different standard classification method to achieve high level of AUC metric.

Conclusion

In this paper, we presented an approach on time series data mining in which automatic time series model identification and automatic outlier detection are employed. Although modern business operations regularly generate a large amount of data, we have found a very little published work that link data mining with time series modeling and gain forecasting applications. By using automatic procedures, we can easily obtain appropriate models for time series and gain increased knowledge regarding the homogeneous pattern of time series as well as anomalous behavior associated with known & unknown events.

References

[1]. R.J.Povinelli, Time Series Data Mining: Identifying Temporal Patterns for Characterization and Prediction of time series events, Ph.D. Dissertation, Marquette University, 1999.
[2]. J.Han and M.Kamber, "Data Mining Concepts and Techniques", 2001, Academic Press, pp 489-499.
[3]. S.M. Pandit and S-M. Wu, "Time Series and System Analysis, With Application", Newyork, Wiley, 1983.
[4]. R.J.Povinelli and X.Feng," Temporal Pattern Identification of Time Series Data Using Patterns Wavelets and Genetic algorithm ", Proceeding of Artificial Neural Network, st. Locuis, 1998, pp 691-696.
[5]. Chang, I, Tiao, G.C and Chan C (1988) "Estimation of Time Series Parameters in Presence of outlier", Technometrics 30.
[6]. Liu, L-M, "Identification of Seasonal ARIMA Models Using Filtering Method", Communication in Statistic A18, 2279-2288.
[7]. Chiu, B.Keogh, E & Lonardi (2003) "Probabilistic Discovery of Time Series Analysis", In 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 493-498.
[8]. Xiong. Y & D. Yeung(2002) "Mixtures of ARIMA Models for Model-based Time Clustering", In Proceeding of IEEE International Conference on Data Mining, pp 717-720.
[9]. Agrawal,R Faloustsos, C. & Swami, A Efficient Similarity Search in Sequence database, Proc of 4th Conference of Foundation of Data Organization.