Forecasting of economic data sets and there trends using time series data modeling

SUNIL BHASKARAN
Department of Computer Science
Army Cadet College Wing
Indian Military Academy, Dehradun
(+91) 9412347407
bsunil2003@gmail.com

Abstract

Towards the end of the 20th century, we have seen an improved interest among Statisticians and Computer engineers to explore and extract data from data sources, by considering time as standard of measurement. However, some of the techniques used remains the same as that used in conventional data mining. As in the case of traditional data base systems, in time series data systems too the methods employed in capturing, indexing, representing and storing the data remains as the key issue. In time series data mining the indexing problem become very critical under the noisy conditions. The indexing problem, however, both in noisy and non noise conditions have exploded the database size. In time series data analysis mathematical/statistical models, that provide descriptions for sample data, (like data collected on global warming, flood forecasting system etc) are used. The method is also used to provide a statistical arrangement for describing the nature of a continues stream of data that fluctuate in a random fashion with respect to the time. A time series can be further defined as a collection of random variables, indexed according to the order they have been extracted. Hence we can assume a time series as a sequence of random variables t1,t2,t3,t4 …., where the random variables t1,t2 etc are the observed values with respect to the time. It is already been proved that statistical methods such as moving average can be effectively used in smoothening data flow. Financial market, equity markets are essentially non-linear in nature. My approach in predicting financial time series is tested in simulation studies using non-linear models. It is shown to have a good success rate of correct forecasting.

Key words : Time Series Data Ming - TSDM, random shock, lag values, information retrieval- IR, SENSEX, Time Series Knowledge Representation – TSKR, Gross Domestic Production - GDP

1. Introduction

In time series data mining, usually the values are observed and recorded at a regular intervals. Depending on the nature of data being observed, the time interval may be in seconds, hours, days, months, years etc. However, the sampled data may be irregular[1].

Popular cases of time series data is Bombay Stock exchange 30 share index (BSE Index – SENSEX), Jones Industrial Average, GDP, vehicle pollution level in a city over a period of time, and rain fall/snow fall pattern recorded over a long period of time. Another important area where information retrieval and time series data mining is employed, on a large scale is financial markets, such as equity market, credit market, banking and mutual fund industry. The data analysis in time series data systems are performed in two distinguished steps, as given in 2(i), 2(ii) [2].

2. Objectives of Time Series Analysis

To device a probabilistic models which could describe the flow of the data in a continues stream with respect to the time. Various parameters considered in the model can be estimated. Conduct trial runes of the model for correctness of the output.

(i) Construction of a model that represents the behavior of the time series under consideration – It can be considered as a primary stage of the data preparation.

(ii) Use the model as discussed above to find the hidden relationship and predict (forecast) future values and trends[4].

Graphical representation of data, is a very powerful and effective mechanism for the objects, events, patterns and concepts in many discipline of science, manufacturing, industry, sports, military, software industry etc. Though the graphical representation of events and patterns are effective many operations in graphs are computation very costly. For instance, computing of similarities in the up side movements and down side movements of stock market of two Asian countries at any given point time from the opening time of the markets, requires exponential computational power. In TSDM patterns are represented with the help of a new language called Time Series Knowledge Representation (TSKR). The TSKR has the following advantages.

- Robustness
- Expressivity
- Comprehensibility

The hidden patterns and there relationships are discovered with suitable algorithms from the individual item-set, which are interlinked. Human interaction is extensively employed for mining, analyze and till the validation of the intermediary results. External interaction is continued till the beginning of further processing of these intermediary results[5,6].

Time series some times has got regular pattern. In such cases a value of the time series should be a function of its previous values. Such time series are called linear time series. When, T is the target value under model framing and prediction. T_t is the value of T at time *t*. Here the goal is to create a model of the form:

$$T_t = f(T_{t-1}, T_{t-2}, T_{t-3}, \ldots, T_{t-n}) + e_t \qquad (1)$$

T_{t-1}, is the value of T for the first observation, T_{t-2} is the value second observations, etc., In the expression (1) e_t denotes a noise that does not follow a per-defined pattern. It is known as random shock[7]. Data points occurring prior to the current observation is known as lag values. When a time series follows a regular pattern, then the value of T_t is generally related with $T_{t\text{-cycle}}$ where, cycle denotes the number of observations in the regular cycle. For instance, weekly readings with respect to a monthly cycle often can be represented by

$$T_t = f(T_{t-12}) \qquad (2)$$

The purpose of constructing a time series model is to build a model such that the deviation from the predicted value and the actual value is as negligible as possible. The fundamental difference between time series models and other similar types of models is that lag values of the target variable are used as predictor variables. However traditional models use external variables as predictors. In such case the concept of a lag value doesn't arise. It is due to the fact that the observations don't reflects a chronological order[8].

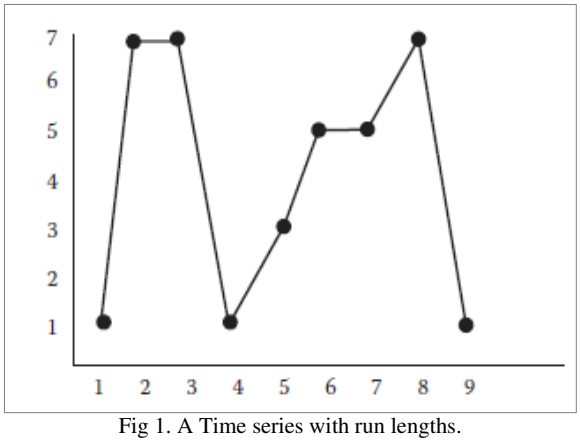

Fig 1. A Time series with run lengths.

3. ARMA and modern types of models

Conventionally, time series analysis, prediction and medaling uses Box-Jenkins ARMA (Auto-Regressive Moving Average) technique. The auto-regressive (AR) part of ARMA predicts the target variable as a linear function of lag values. Moving average (MA) part effect the recent random shock values. Though the ARMA models are extensively used, there application is limited by the linear basis function[8,9].

2.1. 4. Creations of a time series and its subsiquent analysis

4.1 Input variables

When constructing a time series sequence, the input must contains values for one target variable and one or more variables under prediction. The input may consist

of values for a single variable. Let us consider the following example

Table 1. The upward or downward movement of a share of a company in the stock market.

Observation day	Observed value in Rs
Day 1	111
Day 2	116
Day 3	113
Day 4	119
Day 5	119
Day 6	121
Day 7	123

The time between observations are days closing values for seven continues working days of the exchange. In financial markets the data may be missing due to general holidays and non-working periods due to computer system failures.

Table 2. The movement of a share of a company in the stock market with missing values.

Observation day	Observed value in Rs
Day 1	111
Day 2	116
Day 3	113
Day 4	119
Day 5	? Missing value
Day 6	121
Day 7	123

5. Similarity identification

Identifying the similarities in a set of selected data domain as mentioned in table 1 above, is the fundamental requirement for a successful time series data mining in the financial market. Unlike traditional databases, in time series data, the similarity is measured in numerical and continues nature. The same can be interpreted as an approximation technique also. similarity measure is typically carried out in an approximate manner.

Let us take the stock market time series and a typical query to find out all stocks traded in the market, which are having a similar (but non-linear) performance on a particular day of the month or week.

Closing day behavior of the future and Option (F&O) index on a month wise. In the Indian equity market the F&O closing is on the last Thursday of every month and they have been known as November series, December series etc.

6. Intervention variables

An unexpected event and a variation from the calculated flow of data occurring in a time series is popularly called intervention. Such interventions may be caused due to an increase or decrease in interest

rates, a natural calamity or an employee problem[10].

Figure 2. Performance of (NIFTY-India), National stock exchange 100 share index.

(Source, http://www.nseindia.com) – accessed on 19/02/2013.

7. Framing of rules for equity market

A decision on share trading to buy or to sell, rise or fall, can be determined on the basis of pre-established rules, which are established on the basis of the relationships obtained through, the analysis of historical data (previous data). The historical data received are non-volatile in nature. These rules may be in the form of a group of conditional statements.

if <condition A>

&
<condition B>
&
<condition C>
then
<decision>

On the basis of the availability of an appropriate historical data values, a decision tree model can be constructed as mentioned above. It can be used to either validate a rule, or to generate new rules.

Suppose that a decision is fitted using historical data. Each internal node in the tree is a test on one of the variables used to predict the outcome in the historical data. If the variable, say X1, takes

continuous values, this test is either

(X1 >= VALUE)
or
(X1 < VALUE),

where X1 and VALUE are determined by the algorithm that fits the decision tree. If the variable, say
X2, can take one of m discrete values, this test is one of {(X2 = i)}, for i=1, 2, m, where X2 and i are chosen by the fitting algorithm. Leaf nodes contain the actions, eg. buy or sell. Thus tracing down the tree from the root to each leaf node will give a set of rules. For example, historical data could be collected and a decision tree built to test the validity of the following rule[11].

"When the 10-day moving average crosses above the 30-day moving average and both moving averages are increasing it is time to buy"

9. Conclusion and suggestions for future work

In this paper, I am trying make an effort to review some mechanisms followed in the financial market for the prediction, based on time series data analysis. The research conducted here is insufficient for the practical implementation, as the same involves a very high amount of capital risk. However, the priliminary study can be extended to its advanced stages. Representation and storage of time series data is still remaining as a major research area.

10. Some applications of time series data mining:

• Financial service sector: Share rates, trading risk analysis, Forward market analysis etc.

• Data related with economic matters : Analysi of gross Domestic Production - GDP, Consumer Price Indexes - CPI, Whole Sale Price indexes - WPI, Inflationary data analysis etc.

• Very long term data analysis of scientific and environmental data: Rainfall, deforestation, climate variation, formation and disappearance of rivers etc.

• Engineering and Pharmaceuticals : Brain mapping, DNA Decoding, VLSI analysis etc.

11. References

1. Sheng Chang, Wynne Hsu and Mong Li Lee. (2006). *Mining Dense Periodic Patterns in Time Series Data,* Proceedings of the 22nd International conference on data engineering (ICDE'06), 8-7695-2570-9/06, IEEE.
2. Jose Zubcoff , Jesús Pardillo and Juan Trujillo. (2009). *A UML profile for the conceptual modeling of data-mining with time-series in data warehouse*. Information and Software Technology 51(2009) 977–992, Science Direct, ELSEVIER.
3. Juan Trujillo. (2011). *A review on time series data mining - Engineering Applications of Artificial Intelligence*, 24 (2011) 164–181. ELSEVIER.
4. Xiao Hu, Peng Xu, Shaozhi Wu, Shadnaz Asgari and Marvin Bergsneider. (2010). *A data mining framework for time series estimatio*n. Journal of Biomedical Informatics, 43 (2010) 190–199. ELSEVIER.
5. Das P.K, Maya Nayak, Senapati. M.R and Lee I.W.C. 2007. *Mining for similarities in time series data using wavelet-based feature vectors and neural networks*. Engineering Applications of Artificial Intelligence, 20 (2007) 185–201. Science Direct, ELSEVIER.
6. Zhe Song, Xiulin Geng, Andrew Kusiak, and Chang Xu. 2011. *Mining Markov chain transition matrix from wind speed time series data. Expert Systems with Applications* xxx(2011)xxx–xxx. Science Direct, ELSEVIER.
7. Chun-Hao Chen, Tzung-Pei Hong and Vincent S. Tseng. 2009. *Mining fuzzy frequent trends from time series. Expert Systems with Applications*, 36 (2009) 4147–4153. Science Direct, ELSEVIER.
8. Huei-Wen Wu and Anthony J.T. Lee. 2009. *Mining closed Knowledge Engineering*, 68 (2009) 1071–

1090, Science Systems, 24 (2011) 492–500. IEEE, Science Direct, ELSEVIER. ELSEVIER.
9. Hailin Li and Chonghui Guo . 2011. *Piecewise cloud patterns in multi-sequence time-series databases. Data & approximation for time series mining*. Knowledge-Based systems 24(2001) 202–215. Science Direct, ELSEVIER.
10. Dash P.K, Behera H.S and Lee I.W.C 2009. *Time sequence data mining using time–frequency analysis and soft computing techniques*. Applied Soft Computing, 8 (2008) 202–215. Science Direct, ELSEVIER.
11. W. N. Venables, D. M. Smith and the R *Development Core sequence data mining using time–frequency analysis and Team , "An Introduction to R. Manual of R language*", soft computing techniques. Applied Soft Computing, Institute for Statistics and Mathematics, 2007.

Acknowledgements

I take this option to express my sincere thanks to Prof Sandeep Vijay, HOD, Department of Electronics, Dehradun Institute of of Technology (DIT) for encouraging me to write this research paper and get it published. He is always an inspiration for all his students to undertake deep research activity in there respective discipline. Also I am thankful to my wife for providing me a conducive atmosphere and all other background support, without which, it was not possible for me to complete this work.