Volume 3, Issue 11, November 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Using Mining and Intelligent Approach for Time Series Forecasting Problems

Ashwini N, Bhavya G
Assistant Professor
Dept. of ISE, B.M.S.I.T, Bangalore, India

Abstract: In today's real time economic environment, there is ample opportunity to clout the numerous sources of time series data available for decision making. This time ordered data can be used to improve business profit if the data is converted to information and then into knowledge which is called knowledge discovery. Most challenging business processes like trading in retail, financial markets, weather changing data , etc. during their course of activity mount up large databases of time-ordered observations and these data can be presented as time series. There is significant value in the interdisciplinary notion of data mining for forecasting when used to solve time series problems. .The intention of this discussion is to describe how to get the most value out of the numerous available time series data by utilizing data mining techniques specifically oriented to data collection over time, methodologies to extract information i.e. knowledge using intelligent approach and how to create best possible forecast.

Keywords: Neural Networks, Prediction, Self-organizing map, Time series data, Forecasting

I Introduction

Data mining is the process of finding previously unknown and potentially interesting patterns and relations in large databases. Currently electronic data repositories are growing quickly and contain huge amount of data from real time sources related commercial, scientific research, and other domain areas. The ability for collecting and storing all kinds of data totally exceed the abilities to analyze, log and extract knowledge from this logged data. Numerous data mining methods have recently been developed to extract knowledge from huge databases. In most of the cases it is necessary to solve the problem of evaluation and selection of the most appropriate data-mining method or a group of the most applicable methods. Regularly the method selection is done statically without analyzing each particular instance. If the method selection is done dynamically taking into account characteristics of each instance of time, then data mining generally gives better results.

Data mining for forecasting offers the opportunity to leverage the numerous sources of time series data, from internal and external sources, is now readily available for the business decision maker, into actionable strategies that can directly impact profitability. Deciding what to make, when to make it, and for whom is a complex process. Understanding what factors drive demand and how these factors (e.g. raw materials, logistics etc.,) interact with invention processes or demand, and change over time, is keys to deriving value in this context.

Time series data is an ordered sequence of values of a variable at equally spaced time intervals. The time series prediction aim is to observe or model the existing data series to enable future unknown data values to be forecasted precisely. Examples of data series include financial data series (stocks, indices, rates, etc)/physically observed data series (weather etc) and mathematical data series (Fibonacci sequence, differential equations in integral terms, etc). The phrase Time series usually refers to several data sequences, whether or not the data are dependent on a certain time increment. To do time series prediction we can use many different forecasting techniques, which are based on historical time series data and the trends the data reveal. In modern research, different methods taken from a variety of fields are employed for this task.

Time series predation is one of the most important aspects for the practical usage of scientific and engineering knowledge, including physical science, business, engineering processes, control systems, bioengineering, environmental systems, daily temperature, management and econometrics. Real life problems and systems that adequately model such problems are often characterized by a large number of dependent and independent variables, interactions and parameters, resulting in highly complex non-linear dynamics, and in chaotic or random performance. On the other side, the amount of the existing data collected to characterize the system is always limited. The problem of time series predation consists of predicting the next value of a series known up to a definite time, using the known past data values of the timely series. Mostly, time series prediction can be considered a modeling problem. The first step is establishing a mapping between input and output. Generally, the mapping is nonlinear and chaotic. After such a mapping is set up, future values are predicted based on current and past observations.

Time series forecasting (TSF),the forecast of a time ordered variable, turns on into a decisive tool in problem solving, since it allow one to model complex system where the goal is to predict the system's behavior and not how the system works. Indeed in the last few decades an increasing focus has been put over this field. Contributions from the

arenas of operational research, statistics, and computer science as lead to solid TSF methods like exponential smoothing or regression that can replace the old fashioned ones, where were primarily based on institution. An alternative approach for TSF arises from that computational intelligence field, where one has observed a trend to look at nature for inspiration, when building problem solving models. In particular, studies on biological evolution influenced the loom of powerful artifacts, such as Genetic and evolutionary computation that enriched the potential use of intelligence in a broad set of scientific and engineering problem such as the ones of combinatorial and numeric optimization.

Evolutionary algorithms are suited for combinatorial optimization, where the exhaust of all possible solutions require enormous computational power, heuristically finding solution where other seems to fail. The use of ENN in TSF is expected to increase in importance, motivated by advantages such as explicit model representation and adaptive evolutionary search, which escape from unsatisfactory local minima.

There are three fundamentals to consider in the successful implementation of a data mining for time series approach understanding the usefulness of forecasts at different time horizons, differentiating planning and forecasting and, finally, getting all stakeholders on the same page in forecast implementation.

II Literature survey

Many real time organizations require data in different range, hence reference purposes we can include short ranged forecasts are defined herein as one to three years,medium range forecasts are defined as 3 to 5 years and long term forecasts are defined as greater than years. Greater than 10 years should be considered a historical data used for testing rather than a forecast. Several organizations like strategy groups delta with strategic planning are always in need for medium to long range forecasts. Sales and marketing groups demand medium range forecasts for resource planning. Business managers rely heavily on short and medium term forecasts for their own businesses data. Purchasing organization demands high quality forecasts for costs associated with raw materials, logistics, materials, as well as services.

Current time series forecasting methods generally fall into two groups: methods based on statistical concepts and computational intelligence techniques such as neural networks (NN) or genetic algorithms (GA). Hybrid methods combining more than one technique are also commonly found in the literature [1].Statistical time series forecasting methods can be subdivided into the following categories:

(i)Exponential smoothing methods (ii) Regression methods (iii) Autoregressive integrated moving average (ARIMA) method (iv)Threshold methods (v) Generalized autoregressive conditionally heteroskedastic (GARCH) methods.

The first three categories can be considered a linear method that is methods that employa linear functional form for time series modeling and the last two are non-linear methods [2].In exponential smoothing a forecast is given as a weighted moving average of recent time series observations. The values assigned decrease exponentially as the observations get older on time basis. In regression, a forecast is given as a linear function of one or more explanatory variables. ARIMA[7] methods give a forecast as a linear function of past observations (or the differences of past observations) and error values of the time series itself and past observations of zero or more explanatory variables. See Makridakis et al. (1998) for a discussion of smoothing, regression, and ARIMA methods.Threshold methods assume that extant asymmetric cycles are caused by distinct underlying phases of the time series and that there is a transition period (either smooth or abrupt) between these stages. Commonly, the individual stages are given a linear functional form and the transition period (if smooth) is modeled as an exponential or logistic function. GARCH methods are used to deal with time series that display non-constant variance of residuals (error values). In these methods, the variance of error values is modeled as a quadratic function of past variance values and past error values. In Makridakis et al. (1998), McMillan (2001) and Sarantis (2001), various threshold methods are detailed while Akgiray (1989), Bollerslev (1986) and Engle (1982) describe GARCH methods. The literature documenting statistical forecasting methods is vast. Many forecasting studies employ a variation on one of the techniques described above. Some examples include Baille and Bollerslev (1994), Chen and Leung (2003), Cheung and Lai (1993),Clements and Hendry (1995), Dua and Smyth (1995), Engle and Granger (1987), He et al. (2010), Hjalmarsson (2010), Masih and Masih (1996), Ramos (2003), Sarantis and Stewart (1995), Shoesmith (1992), Spencer (1993), Stock and Watson (2002) and Tourinho and Neelakanta (2010). Some studies employ statistical techniques to handle demand time series with unusual characteristics. In Ozden et al. (2009), regression trees are used to handle forecasting demand for products influenced by promotions. Dolgui and Pashkevich (2008) use a Bayesian method to forecast demand for products with very short demand histories. A system described in Chern et al. (2010) uses statistical measures to help users manually select a forecasting model for a particular demand series.

Computational intelligence methods for time series forecasting generally fall into two major categories:(i) Methods based on NN; and (ii) Methods based on evolutionary computation.

Some examples of recent NN forecasting studies include Yu and Huarng (2010) and Zou et al. (2007). General descriptions of NN can be found in Gurney (1997) and White (1992). Back (1996), Michalewicz (1992) and Mitchell (1996) give detailed descriptions of GA while Chambers (1995), Chiraphadhanakul et al. (1997), Goto et al. (1999), Ju et al.(1997), Kim and Kim (1997) and Venkatesan and Kumar (2002) provide examples of GA applied to forecasting. Some examples of EP forecasting experiments include Fogel et al. (1966, 1995), Fogel and Chellapilla (1998) and Sathyanarayan et al. (1999). Some recent examples of GP forecasting applications include Chen and Chen (2010), Dilip (2010) and Wagner et al. (2007).Prevalent in recent literature are forecasting studies which make use of a hybrid model that employs multiple methods. NN are commonly involved in these hybrid models. Examples of hybrid models combining statistical and NN techniques include Azadeh and Faiz (2011), Mehdi and Mehdi (2011), Sallehuddin and

Shamsuddin (2009) and Theodosiou (2011). Examples of models combining GA and NN techniques include Araujo (2010), Hong et al. (2011) and Wang et al. (2008). Johari et al. (2009) provide a hybrid model that combines EP and NN while Lee and Tong (2011) and Nasseri et al.(2011) provide hybrid models that combine GP with an ARIMA model and a Kalman filter, respectively[3]. Sayed et al. (2009) provide a hybrid model that combines GA and statistical techniques while Wang and Chang (2009) provide a hybrid model that combines GA and diffusion modeling.

III Methodology

In the univariate time series prediction problem, the input data consists of a sequence of values and the next values should be estimated by analysis of the previous observations. Auxiliary factors are not involved in this process. When the predicted values are obtained using the past and current, frequently confidence levels are presented along with the results. Some of the most commonly used methods for time series prediction are averages, autoregressive methods, decomposition, exponential smoothing or regression, trend extrapolation, neural networks, etc.

In this research we are planning to develop computational Intelligence concept based on neural network, evolutionary computation etc. in solution development of forecasting models and rules, research will also have direction to conducted experiments in none quadratic error functions for neural networks training or asymmetric costs for export evaluation in prediction theory, however, neural networks theory as traditional prediction theory focus on quadratic error function and least square predictors which implies a symmetric and quadratic cost relationship.

In one of the method of prediction by neural networks by obtaining subseries from the initial time series is considered. This subseries which are obtained through a sliding window that traverses the time series and thus generates input-output training patterns for the network. Most of the regression models use this technique. Neural network forecasting and classification functions help users discover relationships and valuable information in vast quantities of data. Neural Networks provide a highly flexible forecasting approach that delivers accurate forecasts under a variety of conditions including short time series, complex data systems, diversified categorical and continuous data, a large number of variables and noisy data. An important matter is the determination of the window size. There are two main methods for solving this problem. Utmost often autocorrelation and partial autocorrelation functions are analyzed for this purpose. Occasionally because of the complexity of their structure brute force searching can be done to reduce the complexity. The search starts with a window [21] containing a single value.After searching model building is performed then followed by test prediction and error calculation. After that the size of the window is increased by one element and the process is repeated until the size of the window is equal to the size of initial time series or satisfactory error is reached. In the prediction stage the sliding window is situated at the end of the series to obtain input values to feed the neural network. The output of the neural network is considered as the first expected value. Then this first prediction is considered as a part of the initial time series and the window is moved again in order to predict the second value. This procedure (recursive prediction) is repeated until the desired time prospect is reached. This process is graphically shown on Fig.1. The estimate begins when the window is moved N-p steps, where N is the initial time series length, p is the window size (number of values shown in the grey rectangle).

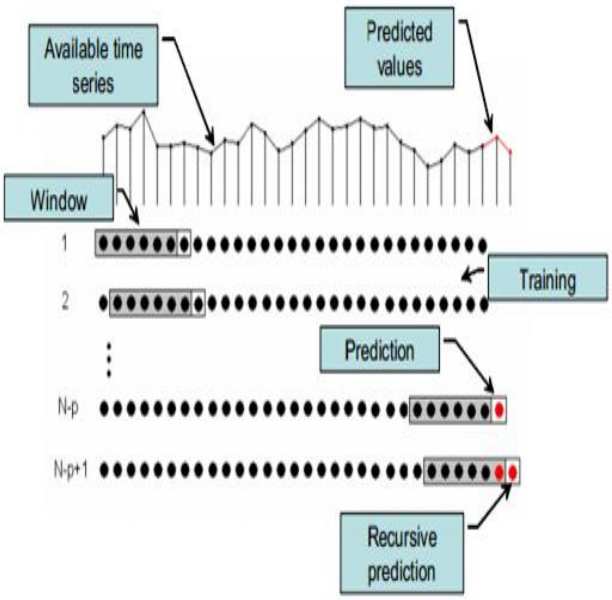

Fig 1: Subseries approach.

Some preprocessing operations should be done before the training process. The subjective of the preprocessing is to transform the initial series into a form needed for the neural network to work accurately. This stage consists mainly of the following operations: normalization, diverse techniques for trend removing and sometimes seasonality removing. They can be used in different order according to the time series characteristics.

Because many real world scientific and engineering problems are irregular, deterministic and enumerative search techniques are then unsuitable. Stochastic search and optimization approaches were developed as alternative approaches for solving these unbalanced problems. Stochastic methods require a function assigning fitness values to possible solutions, and encode /decode (planning) mechanism between the problem and algorithm domains. In general stochastic methods provide good solutions to a wide range of optimization problems which traditional deterministic search methods find difficult. In order to provide the required solution the data can be processed through content using some clustering techniques to identify the patterns which can be predicted.

Clustering techniques[20] are mainly used for preprocessing the data existing. Preprocessing the data means the data existing that is the initial set which has both historical data and real time updated data. Hence in order to separate the data in terms of some is varying independent variable as a output of some cluster which consist of training data set to train the system to forecast and actually data to be forecasted as shown in Fig 2.

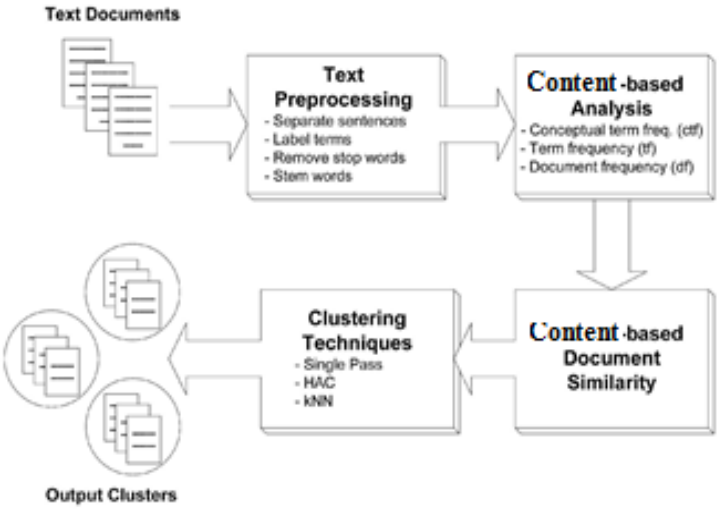

Fig 2: Clustering approach.

IV. Present Challenges and Possible solution

Neural networks have been used in applications of prediction for time series. Many different neural networks and learning methods has been applied to the times series prediction problem with varying degrees of success .The multi-layer perceptron's is probably the most frequently used type of neural networks. However, due to its multi-layered structure and the greedy nature of the back propagation training algorithm, the training processes using Back-Propagation may produce problems in training, as converge to local minima of the error surface or converge too slowly so as to affect the reliability and accuracy of prediction model With the intelligent integration of different architecture and evolution, ANNs can be able to approximate data with nonlinear and non-stationary properties and to conduct self-learning adaptive structure identification and parameter estimation in modeling process.

Many physical and artificial problems can be described by time series. The prediction of such problems could be as complex as interesting. Most of the time series forecasting methods look for universal rules to predict the whole series. The key problem is that time series usually have local behaviors that don't allow forecasting the time series by general rules. In this research, focus will be given for finding a new local prediction rules. Those local prediction rules can attain better general prediction accuracy. The method in this research will be based on the evolution of a rule system.

V. Objectives and Possible Outcomes

Chaotic behavior in deterministic dynamical systems is an intrinsically non-linear phenomenon. A characteristic feature of chaotic systems is an extreme sensitivity to changes in initial conditions, whiles the dynamics at least for so-called dissipative systems, is still constrained to a finite region of state space called a strange attractor. Time hints of the state variables of such systems display a seemingly stochastic behavior. In this research, concept of evolution will applied to evolve neural network at various level for predicting the time series with higher accuracy. Various benchmarks of chaotic time series like Logistic differential equation, Micky-glass and Lorenz system will take for experiment along with some other time series data set. Also In this research, method for finding local prediction rules will be developed with evolutionary approach [19] to attain better general prediction accuracy.

References

[1] Akgiray, V. (1989), "Conditional heteroskedasticity in time series and stock returns: evidence and forecasts", Journal of Business, Vol. 62, pp. 55-80.

[2] Araujo, R. (2010), "A quantum-inspired evolutionary hybrid intelligent approach for stock market prediction", International Journal of Intelligent Computing and Cybernetics, Vol. 3,pp. 24-54.

[3] Azadeh, A. and Faiz, Z. (2011), "A meta-heuristic framework for forecasting household electricity consumption", Applied Soft Computing, Vol. 11, pp. 614-20.
[4] Back, T. (1996), Evolutionary Algorithms in Theory and Practice: Evolution Strategies,Evolutionary Programming, and Genetic Algorithms, Oxford University Press,New York, NY.
[5] Baille, R. and Bollerslev, T. (1994), "Cointegration, fractional cointegration, and exchange rate dynamics", Journal of Finance, Vol. 49, pp. 737-45.
[6] Bass, F. (1969), "A new product growth model for consumer durables", Management Science,Vol. 15, pp. 215-27.
[7] Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasticity", Journal of Econometrics, Vol. 31, pp. 307-27.
[8] Chambers, L. (Ed.) (1995), Practical Handbook of Genetic Algorithms: Applications, CRC Press,Boca Raton, FL.
[9] Chen, A. and Leung, M. (2003), "A Bayesian vector error correction model for forecasting exchange rates", Computers & Operations Research, Vol. 30, pp. 887-900.
[10] Chen, S. and Chen, J. (2010), "Forecasting container throughputs at ports using genetic programming", Expert Systems with Applications, Vol. 37, pp. 2054-8.
[11] Chern, C., Ao, I., Wu, L. and Kung, L. (2010), "Designing a decision-support system for new product sales forecasting", Expert Systems with Applications, Vol. 37, pp. 1654-65.
[12] Cheung, Y. and Lai, K. (1993), "A fractional cointegration analysis of purchasing power parity",Journal of Business and Economic Statistics, Vol. 11, pp. 103-12.
[13] Chiraphadhanakul, S., Dangprasert, P. and Avatchanakorn, V. (1997), "Genetic algorithms in forecasting commercial banks deposit", Proceedings of the IEEE International Conference on Intelligent Processing Systems, Beijing, China, Vol. 1, pp. 557-65.
[14] Clements, M. and Hendry, D. (1995), "Forecasting in cointegrated systems", Journal of Applied Econometrics, Vol. 10, pp. 127-46.
[15] Dilip, P. (2010), "Improved forecasting of time series data of real system using genetic programming", GECCO '10 Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, Portland, OR, USA, Vol. 1, pp. 977-8.
[16] Engle, R. (1982), "Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation", Econometrica, Vol. 50, pp. 987-1008.
[17] Lee, Y. and Tong, L. (2011), "Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming", Knowledge-Based Systems, Vol. 24, pp. 66-72.
[18] McMillan, D.G. (2001), "Nonlinear predictability of stock market returns: evidence from nonparametric and threshold models", International Review of Economics and Finance, Vol. 10, pp. 353-68.
[19] Fogel, L., Angeline, P. and Fogel, D. (1995), "An evolutionary programming approach to self-adaptation on finite state machines", Proceedings of the 4th Annual Conference onEvolutionary Programming, San Diego, CA, USA, Vol. 1, pp. 355-65.
[20] D. Gildea and D. Jurafsky, "Automatic Labeling of Semantic Roles," Computational Linguistics, vol. 28, no. 3, pp. 245-288, 2002.
[21] Zhang, G. Neural Networks in Business Forecasting. Idea Group Publishing,2004.

