APPLICATION OF GENETIC ALGORITHM TO TIME SERIES DATA MINING

Mrs.R.A.Fadnavis
Assistant professor
Yeshwantrao Chavan college of Engineering, Nagpur-440010

Abstract: Time series data mining plays a major role in organizations decision making. Time series data mining involves extraction of all meaningful knowledge and patterns from data collected over a period of time. Various data mining techniques like classification, clustering, segmentation, anomaly detection, prediction are applied to different type of time series data to find useful and interesting patterns from it. This paper focuses on one such important mining task performed on time series data - prediction. Various statistical methods used for time series prediction includes regression, exponential smoothing, moving averages, autoregressive and moving average models, etc. Many times complexity underlying these methods precludes its use by those less acquainted with them. So it is possible to use heuristic method which gives approximated, but satisfactory solution for such class of problems. This paper proposes heuristic approach for time series data mining task-prediction. The approach used is based on evolutionary computation technique- genetic algorithm. Prediction results are evaluated on the basis of R^2 and Root mean square error value (RMSE).

Index terms: Time series, Time series data mining, prediction, genetic algorithm

I. Introduction

A time series is collection of observations often recorded at fixed time intervals. Examples of time series data could be sequence of annual sales of products, monthly inventory levels, weekly exchanged rates and so on. Many business and economic applications of forecasting involve time series data.

1.1 Major tasks in time series data mining:

Following are the various data mining tasks that can be applied on time series data:

a. Clustering:

Clustering task involves finding group of similar observations, called clusters, in a dataset. The main objective is to find the most similar clusters that are as distinct as possible from other clusters. More formally, the grouping should maximize inter-cluster distance and minimizing intra-cluster distance. It is an unsupervised approach.

b. Classification:

The classification task involves assigning labels to each series of a set. The main objective is to learn about the distinctive features distinguishing classes from each other and, when an unlabeled dataset is inputted into the system, it can automatically determine which class each series belongs to. The main difference in classification over clustering task is that classification follows supervised approach. In classification classes are known in advance and the algorithm is trained on an example dataset.

c. Segmentation:

The segmentation (or summarization) task includes creating an accurate approximation of time series, by reducing its dimensionality while retaining its essential features. The objective of this task is to minimize the reconstruction error between a reduced representation and the original time series

d. Anomaly detection:

The anomaly detection task involves finding abnormal subsequences in a series. It has numerous applications like intrusion detection, biosurveillance etc.

e. Prediction:

The Prediction task involves predicting the value of future observations based on past/ historical values of time series data . While predicting time series data, different kind of data patterns that exist within a time series need to be considered. Four types of data patterns can be identified: horizontal, trend, cyclical, and seasonal.

Data patterns in Time Series Data:

Horizontal: When data observations fluctuate around a constant level or mean, a horizontal pattern exists. This type of series is called stationary in its mean [2]. For instance, monthly sales for a product that do not increase or decrease consistently over time would be considered to have a horizontal pattern.

Trend: When data observations grow or decline over an extended period of time, a trend pattern exists. [2]. Examples of basic forces that affect and help explain the trend of a series are population growth, price inflation, technological change, consumer preferences and productivity increases.

Cyclical: When observations exhibit rises and falls that are not of a fixed period, a cyclical pattern exists [2]. The cyclical component is wavelike fluctuation around the trend that is usually affected by general economic conditions. Cyclical

fluctuations are often influenced by changes in economic expansions and contractions, commonly referred to as Business cycles.

Seasonal: When observations are influenced by seasonal factors, a seasonal pattern exists [2]. The seasonal component refers to pattern of change that repeats itself year after year. For a monthly series, the seasonal component measures the variability of the series each January, February, and so on. For a quarterly series, there are four seasonal elements, one for each quarter.

II. Literature Survey

This section focuses on the different methods used for predicting time series data and introduction to an evolutionary computation technique - genetic algorithm

2.1 Time Series prediction models:

For predicting time series data various methods are available like Naïve methods, n-period moving average, Exponential smoothing, Holts Linear Exponential smoothing, Winters multiplicative smoothing technique, Autoregressive models , Box Jenkins ARIMA model ,Regression techniques. This paper focuses mainly on regression techniques and specific case of application of genetic algorithm on autoregressive and moving average model.

2.1.1 Regression Models:

Regression models are used to predict one variable from one or more other variables. Regression models provide a powerful tool, allowing predictions about past, present, or future events to be made with information about past or present events. In order to construct a regression model, both, the information which is going to be used to make the prediction and the information which is to be predicted, must be obtained from a sample of objects or individuals. The relationship between the two pieces of information is then modelled with a linear transformation. Then in future, only the first information will be necessary, and the regression model will be used to transform this information into the prediction. In statistical modelling of N data observations two types of variables are usually defined. One is the response variable or variate, usually denoted by Y, and the other is the explanatory variable or covariate X. The number or explanatory variables can classify a regression model as Simple regression model when there is only one explanatory variable, and only one response variable.

2.1.2 Autoregressive moving average model:

This model includes autoregressive as well as moving average parameters, and explicitly includes differencing in the formulation of the model. Specifically, the three types of parameters in the model are: the autoregressive parameters (p), the number of differencing passes (d), and moving average parameters (q) [3]. In the notation introduced by Box and Jenkins, models are summarized as ARIMA (p, d, q); so, for example, a model described as (1, 1, 1) means that it contains 1 autoregressive (p) parameters and 1 moving average (q) parameters which were computed for the series after it was differenced once(d).

2.2 Introduction to Genetic Algorithms

Genetic algorithms attempt to mimic computationally the process by which natural selection operates and apply them to solve business and research problems. [1] Genetic algorithms are search and optimization algorithms based on the principles of natural evolution.They have numerous applications in business, scientific and engineering optimization problems. In the world of genetic algorithms, the fitness of various potential solutions are compared and the fittest potential solution evolve to produce ever more optimal solutions.

2.2.1Basics of genetic algorithms

In Genetic algorithms, a chromosome refers to one of the candidate solutions to the problem, a Gene is a single value of candidate solution, and an allele is a particular instance of the value.

Genetic algorithm is composed of three operators [4]:-

(1) Reproduction / Selection (2) Cross Over (3) Mutation

The balance, which critically controls the performance of Genetic Algorithm are the Cross over and mutation probabilities and population sizes.

Reproduction

Reproduction is a process in which individual strings are copied according to their objective function values f. (the function is called as fitness function). The strings are copied according to their fitness value which means that the strings with a higher value have a higher probability of contributing one or more offspring in the next generation and the strings with a lower value(less than average) have a probability of not contributing any offspring in the next generation.

Cross over

Crossover (simple) may proceed in two steps. First, member of newly reproduced string in the mating pool are mated in random. Second, each pair of strings undergoes crossing over.

Mutation

Mutation decides the probability of changing the bits from 1 to 0 and vice versa. With a low probability rate, if the expected mutation is very less, no mutation is carried out. Mutation introduces new information to the genetic pool and protects against converging too quickly to a local optimum.

Most genetic algorithms functions by iteratively updating a collection of potential solutions called a population. Each member of population is evaluated for fitness on each cycle. A new population replaces old population using above said operators, with the fittest members being chosen for reproduction.

The standard genetic algorithm in pseudo-code format is given as:

Pseudo-code of the standard genetic algorithm

```
Begin GA
  g:=0  { generation counter }
  Initialize population P(g)
  Evaluate population P(g)  { i.e., compute fitness values }
  while not done do
    g:=g+1
    Select P(g) from P(g-1)
    Crossover P(g)
    Mutate P(g)
    Evaluate P(g)
  end while
end GA
```

III. Research Methodology

Following section focuses on Application of Genetic Algorithm to Autoregressive and moving average model.

Most of the time series prediction methods are based on Autoregressive and moving average derived methods. Many times complexity underlying these methods precludes its use by those less acquainted with them. On the other hand it is possible to use heuristic method that gives approximated, but satisfactory solutions for such class of problems. This paper proposes heuristic method for time series prediction using Autoregressive and moving average. The method is based on evolutionary computation technique, genetic algorithm. Considering the time series prediction as a particular problem of Autoregressive and moving average, different series found in the literature were modeled using Autoregressive and moving average method. Performance was measured using the R^2 and Root mean square value. For each data series time series two prediction methods were applied. Prediction results were evaluated on the basis of R^2, RMSE. Results show that the heuristic method here proposed is appropriate for modeling time series using Autoregressive and moving average model of order 2.

Basic Model

The proposed model is Autoregressive and moving average of order 2 is given by equation 3.1[2] [3]

$$Y_t = \phi_0 + \Phi_1 * Y_{t-1} + \Phi 2 * Y_{t-2} - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} \qquad (3.1)$$

Genetic algorithm is used to determine the coefficients of the above model. The approach is as follows.

Initial values for the coefficients are decided randomly by Genetic Algorithm. The fitness function taken as RMSE. Genetic Algorithm is applied to get the optimum value of RMSE and R^2 and to find the best coefficients of the model. The termination criterion is the number of iterations.

IV. Results and discussion

Two methods of prediction -Simple regression and genetic algorithm applied on auto regression and moving average of order 2 are evaluated for different datasets. The forecasted values obtained by applying these two models on some of the data sets are given below:

Table 4.1 Time series prediction results when applied on different data sets using Simple regression and genetic algorithm applied on auto regression and moving average of order 2

Dataset	METHOD	MSE	RMSE	R^2	EQUATION
The data gives the monthly sales of pairs of jeans in the UK (in thousands) from January 1980 to December 1985. There are 72 observations	Normal Regression	49871.4057	223.319067	0.1127	Y=1932.491+3.82976 3X
	Genetic Algorithm (Auto regression and moving average model of order 2)	40443.097	201.104692	0.997288	

This data set contains selected monthly mean CO_2 concentrations at the Mauna Loa Observatory from 1974 to 1987. Each line contains the CO_2 concentration (mixing ratio in dry air, expressed in the WMO X85 mole fraction scale, maintained by the Scripps Institution of Oceanography.)	Normal Regression	4.71875117	2.17226867	0.8701487	Y=329.33+0.1209X
	Genetic Algorithm(Auto regression and moving average model of order 2)	1.58697267	1.25975104	0.9995382	
This data gives quarterly production of sulphuric acid in Australia: thousand tones. Mar 1956 - Jun 1994.	Normal Regression	13615.6211	116.685994	0.0061724	Y=380.4475+0.2095X
	Genetic Algorithm (Auto regression and moving average model of order 2)	4412.79382	66.4288628	0.9922444	
Monthly international airline passenger traffic from 1949-1956. There are 96 observations.	Normal Regression	932.5292	30.5373411	0.8178	100.474+2.334X
	Genetic Algorithm (Auto regression and moving average model of order 2)	535.7291	23.1458225	0.9917	

Depending on the objectives of the forecast and preciseness needed therein, it is important to choose a method of prediction that suits the most in a given situation. A method that fits well with the available data and gives the best results should be used for making the forecast. Different time series found in the literature were modeled by auto regression and moving average of order2 using genetic algorithm . For each of these data series other time series prediction method normal regression is also applied. Forecasting results were evaluated on the basis of $R^2$and RMSE. The paper showed the results obtained by the application of genetic algorithm. Results shows that the heuristic method proposed here is more appropriate for predicting time series

References

[1] Daniel T. Larose, Data Mining Methods and Models, Wiley, second reprint 2007.
[2] John. Hanke and Dean w. Wichern , Business Forecasting, 8th edition, PHI publication, 2006.
[3] Box, G. E. P., Jenkins, G. M. and Reinsel, G. C., Time Series Analysis, Forecasting and Control, 3rd ed, pearson education,2004.
[4] David E. Goldberg ,"Genetic algorithm in search , optimization and machine learning", Pearson education, low price edition.
[5] "A review on time series data mining", Tak-chungFu, Engineering Applications of Artificial Intelligence 24 (2011) 164– 181
[6] "Mining Time Series Data", Chotirat Ann Ralanamahatana ,Jessica Lin, Dimitrios Gunopulos, Eamonn Keogh, Michail Vlachos, Gautam Das, Data Mining and Knowledge Discovery Handbook pp 1069-1103.

[7] Sam Mahfoud , Ganesh Mani," Financial Forecasting Using Genetic Algorithms Applied Artificial Intelligence", 10:543± 565, 1996Copyright € 1996 Taylor & Francis 0883-9514/96 $12.00 + .00

[8] Heitor. S. Lopes,Wagner.R.weinert, " A gene expression programming system for time series modeling ", proceedings of XXV Iberian Latin American Congress on computational methods in engineering, November2004.

[9] Douglas J.Dalrymple, "Using Box-Jenkins techniques in sales forecasting",Journal of Business Research, Volume 6, Issue 2, May 1978, Pages 133-145.

[10] Chorng-Shyong Ong , Jih-Jeng Huang , Gwo-Hshiung Tzeng " Model identification of ARIMA family using genetic algorithms", C.-S. Ong et al. / Appl. Math. Comput. 164 (2005) 885-912.http://ntur.lib.ntu.edu.tw/bitstream/246246/84973/1/12.pdf

[11] Ion Dobre, Adriana AnaMaria Alexandru, "Modeling Unemployment Rate Using Box-Jenkins Procedure ", Journal of applied quantitative methods, Volume 3, No.2 summer 2008