Book Loan Quantity Prediction Using Time Series Data Mining

Yuqing Shi[1,2, a], Yuelong Zhu [2,b]

[1]Library, Hohai University Nanjing, China

[2] College of Computer and Information Engineering, Hohai University Nanjing, China

[a]shiyuqing@hhu.edu.cn, [b]ylzhu@hhu.edu.cn

Keywords: Digital Library; Book Loan Quantity; Time Series Data Mining; Event prediction

Abstract. This paper shows a new method to book loan quantity prediction using time series data mining which unites data mining and chaos theory to characterize and predict events in nonperiodic, complex and chaotic time series. Intelligent library management, including inquire, borrowing and reading, typify a class of nonlinear systems named chaotic, in which the relationships between variables in a system are disproportionate and dynamic, nevertheless entirely deterministic. Chaos theory offers an integrated explanation for anomalies and irregular behavior in systems that are not internally stochastic. The showed time series data mining technique concentrate on prediction of events where book loan quantity comprises the events in a library daily time series. The technique is demonstrated using data collected at library of Hohai University in China. Results associated with the impact of earliness of prediction and the prediction accuracy vs. acceptable risk-level is presented.

Introduction

Traditional library book loan quantity statistical methods have become discommodious and low-efficiency in assisting library users. The purpose of data mining is to enhance the quality of the mutual effect between the library and its users. The collected data contain valuable information that can be used to improve library decisions, and can be integrated into the library's strategy. We need an automatic discovery and analysis tool for extracting useful knowledge from huge amounts of untreated library data. The goal of this paper is to illustrate how time series data mining technology is a good approach to accomplish users' requirements.

In recent years, many studies from fields of librarianship [1], statistics [2] data mining [3], and event prediction [4] have contributed to research in digital libraries. Among these methods, linear and nonlinear time series models are called black-box models in which prediction is used. The black-box models try to found connection between model outputs and inputs.

Nonlinear time series methods such as Hidden Markov Models (HMM) [5], Artificial Neural Networks (ANN) [6] and Nonlinear Prediction (NLP) [7] have been applied to digital libraries.

The rest of the paper is organized as follows: the next segment introduces the TSDM (time series data mining) technique and third segment presents the application of the TSDM to the book loan quantity prediction problem. Fourth segment describes the results and discussions regarding the prediction accuracy associated with the TSDM approach, and finally conclusions are consisted in the last segment.

Time Series Data Mining Technique

The TSDM technique obeys the time delayed imbedding procedure to predict future happenings of events. [8] TSDM unites the method techniques of data mining and phase space reconstruction to uncover concealed patterns prognostic of future events in nonstationary and nonlinear time series. The steps contained in the TSDM technique discoursed in the following section.

Phase space reconstruction is a technique method that offers a simplified, multi dimensional representation of frequently a simplex dimensional nonlinear time series. Attractors are the status conditions towards which a system develops when starting from begin with sure initial conditions. Because of the dynamic of the chaotic system is undiscovered, the original theoretic attractor that produces the observed time series cannot be manufactured. Inverse, a phase space is manufactured

where the attractor is remanufactured from the invariant observed data that conserves the scalar peculiarities of the original undiscovered attractor depicted by the time delay technique to come close the status space from a simplex time series.

The remodeled phase space is an M-dimensional metric space into which a time series is inset. It is a vector space for the system such that appointing a status of the system appoints the dot in this space at any specific instant and vice-versa. Time delayed insetting maps a group of M observations obtained from time series X onto x_t, where x_t is a dot or vector in the phase space. The time series is delegated by $\{x_{t-(M-1)s}, \ldots, x_{t-2s}, x_{t-s}, x_t\}$ where x_t delegates the present observation, and $(x_{t-(M-1)s}, \ldots, x_{t-2s}, x_{t-s})$ are the bygone observations. Where t is the present time index, then t + s is a time index in the future, and t - s is a time index in the bygone. The implanting delay is the delta-T in number of hourly bases between neighbor components of delay vectors and the implanting dimension (M) is the number of dimensions of reconstructed phase space. Any ulterior analysis of conclusive attributes of a nonlinear time series hinges on the prerequisite of a triumphant reconstruction of a state space of the potential process.

In the back of an implanting is accomplished, the following procedure is to define a mathematical function that facilitates confirm the events of interest and projects them in the phase space to confirm the temporal patterns clusters and temporal pattern. The event characterization function is an application dependent function where the value that signals the equation at time t correlates to the value of that event in the future. These types of event characterization functions are useful in prediction type problems and are classified as causal.

The objective function is used to confirm which temporal pattern cluster is competent in its competence to characterize events and is tie in with the TSDM target. The search for first-rank temporal pattern cluster is implemented using Genetic Algorithms.

If the first-rank pattern clusters determined in the training phase are adequately precise in predicting the events in the training time series, these clusters are utilized for the testing time series. Else the training phase is reduplicated for alternate event first-rank formulations, characterization functions or objective functions. In the testing phase, the testing time series is implanted in the renewed stage space using the same implanting parametric quantities as the training time series. At any time a dot in the testing time series phase space subsides inside the cluster identified in the training phase, an event is predicted. Testing results are appraised by measuring the number of events properly predicted and identified.

Application of TSDM to Book Loan Quantity Prediction

The TSDM technology is applied to special book loan quantity forecasting at the library of Hohai University in China. The daily amount of borrowing/returning time series is obtained from the Huiwen digital library management system, covering the period from May 2000 to August 2015, consisting of 11,040 data points. Before the TSDM technology is applied to special book loan quantity forecasting, the presence of nonlinearity in the amount of borrowing/returning time series is obtained from the Huiwen digital library management system, covering the period from May 2000 to August 2015, consisting of 11,040 data points. Time series is confirmed using the surrogate data method.

As phase spaces are renewed in higher embed implanting dimensions for an infinite data set, a point will be reached after which increasing the number of dimensions will not have any significant effect on the correlation dimension.

An important consideration in the accurate identification and prediction of special book loan quantity events is the selection of thresholds. Selection of a high threshold will cause events to be missed and conversely, a low threshold will result in a large number of false positives. Historical records indicate that special book loan quantity have occurred during the years 2001, 2005, 2006, 2008, 2012 and 2014. The minimum value of amount of borrowing/returning, during the reported special book loan quantity periods is 2,000copy/day. This metric has been chosen as the threshold for identifying events, i.e. a value of amount of borrowing/returning exceeding 2,000copy/day is

considered a special book loan quantity and hence constitutes an event. The time series is split into two parts, the first 8,000 data points make up the training time series and, the rest is used in testing. The time series is implanted in a two dimensional phase space with a delay of one.

The event characterization function $f(x_t) = x_{t+i}$, captures the goal of characterizing a special book loan quantity i time steps in the future is selected. The renewed phase space is augmented by the $f(x_t)$ value for each corresponding phase space point. Since the event event characterization function is a step-ahead function, the points that have the highest $f(x_t)$ value in the augmented phase space are the phase space points that have high discharge i steps ahead in the future depending on the step-ahead function used.

The objective function is defined in a manner that incorporates important aspect of the special book loan quantity prediction problem. Based on the historical data, a minimum borrowing/returning value that results in a special book loan quantity can be identified and each event can be classified as a false positive (fp) or true positive (tp). If an event identified by the cluster as a special book loan quantity is actually not a special book loan quantity, then it constitutes a false positive. If an event identified by the cluster is a special book loan quantity, then it is called a true positive.

For every tp included in the cluster, the objective function is rewarded by multiplying the summation of the event of points inside the cluster by the number of tp's inside the cluster increasing the value of the objective function. An optimization formulation that maximizes the objective value, would include all the points in the data set in the cluster, which in essence would predict every event as a special book loan quantity. The user defined parameter value controls the number of points in the cluster based on the planners' choice of acceptable risk. In other words, a low value of parameter minimizes the number of points inside the cluster that are not actual special book loan quantity, minimizing the number of false alarms or false positives, but at the same time increases the risk of missing an actual special book loan quantity. On the other hand, a high value of parameter leads to a large number of false positives to be included in the cluster but decreases the risk of missing an actual special book loan quantity, effectively incorporating the planners' desired level of risk. As a result, the maximization of the objective function results in a cluster that tries to include all true positives along with other phase space points with high event of the point i inside the cluster values, and the number of points in the cluster is restricted by parameter.

The optimization formulation is modeled as a multi-objective formulation with the two objectives as: minimize the radius of the cluster and maximize the value of objective function. An unsupervised clustering technique, the genetic algorithm is used in the search process for optimal temporal pattern cluster. The minimization objective is required to select the crispest cluster in order to minimize the number of false positives in the testing phase. The genetic algorithm toolbox in Matlab is used for the search and the output is the cluster center and its radius.

To determine the prediction accuracy of the clusters and measure their ability to predict special book loan quantity in the training and testing phases, the following set of performance parameters is measured:

1. True positives (tp). If the event identified by the cluster as a special book loan quantity is actually a special book loan quantity it is called a true positive.

2. False positive (fp). If an event identified by the cluster is not a special book loan quantity it constitutes a false positive or a false alarm.

3. Positive prediction accuracy is the percentage of true positives in the cluster. Since the events are classified either as false positives or true positives, the positive prediction accuracy of a cluster is calculated as (tp/(tp + fp)) *100.

4. Correct prediction percentage is the percentage of true positives predicted with respect to actual events and is calculated as tp*100.

TSDM methodology results and discussion

This section presents the results of the application of the TSDM technology to the borrowing/returning data set from the Huiwen digital library management system of Hohai University. The optimization process is able to identify clusters which include all 15 tp's and have a 100% correct prediction percentage. No cluster is identified for parameter value of 0.05. These results are indicative of appropriate event characterization and objective functions. The clusters for different values of parameter identified in the training stage are used to predict special book loan quantity in the testing stage. For every tp included in the cluster, the objective function is rewarded by multiplying the summation of event values of points inside the cluster by the number of tp's inside the cluster increasing the value of the objective function.

Conclusions

This paper presents the application of the TSDM technology in the prediction of special book loan quantity. The technology is applicable to the Huiwen digital library data at the library of Hohai University with the target of predicting special book loan quantity accurately and as early as possible. The training phase results indicate that the selected event objective functions and the characterization are successful in predicting special book loan quantity and the beginning of all special book loan quantity are successfully predicted in the testing phase. An inverse relationship exists between the earliness of prediction and the correct prediction percentage where, as the prediction horizon increases, the correct prediction accuracy decreases. Ideally, one would like to predict a special book loan quantity as early as possible, however associated with the earliness of prediction is the risk of missing the start of special book loan quantity.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No. 51079040/E090101) and Hohai university library '2015' key research projects (No.20150323).

References

[1] H.M. Osborne and A. Cox: Program-Electronic Library and Information Systems Vol. 49-1 (2015), p. 23-45

[2] E. Cerrillo-Cuenca and M. Sepulveda: Journal of Archaeological Science Vol. 55 (2015), p. 197-208

[3] M. Shoaib, A. Daud and M.S.H. Khiyal: Arabian Journal For Science And Engineering Vol. 40-6 (2015), p. 1591-1605

[4] S. Lee, H. An, S. Yu, and J.J. Oh: Environmental Earth Sciences Vol. 72-8 (2014), p. 3111-3128

[5] L. L. Presti and M. L. Cascia: Multimedia Tools and Applications Vol. 68-3 (2014), p. 777-803

[6] Y.D. Bu and J.C. Pan: Monthly Notices Of the Royal Astronomical Society Vol. 447-1 (2015), p. 256-265

[7] Z. Wang, R. Murgai and J. Roychowdhury: IEEE Transactions on Computer-Aided Design Of Integrated Circuits and Systems Vol. 24-1 (2005), p. 56-64

[8] A. Zuccala, M. Thelwall, C Oppenheim and R. Dhiensa: Journal of Documentation Vol. 63-4 (2007), p. 558-589