Integrated Cluster-based Rule Induction Mining of Temporal Data for Time-Series Analysis

K. Poongothai1*, M. Mathivanan2, T. Senthilkumaran2 and S. Revathi3

1Selvam College of Technology, Namakkal - 637003, Tamil Nadu, India; sakthipoongi@gmail.com
2A.C.S. College of Engineering, Bengaluru - 560074, Karnataka, India; tmmathiv@yahoo.co.in, senthilvts@gmail.com
3B. S. Abdur Rahman University, Chennai - 600048, Tamil Nadu, India; srevathi@bsauniv.ac.in

Abstract

Objectives: To deliver excellent performance when compared to existing method Mining Comprehensible Classification Rules for Time-Series (MCCR-TS) using zoo dataset collected from UCI repository comprising of more than 1000 records.

Methods/Analysis: Research works on Time-series data has been evolving as a new trend in current scenarios due to the wide range of applications involved in it. One of the widely researched topics is the web usage mining and effective evaluation of time-series movements and transaction associated with them. Most of the studies conducted earlier, focused on identifying the time-series data from the entire logs. However, this type of patterns may not be accurate enough for evaluation due to differentiated behaviors of user patterns are not taken into consideration. Findings: In this paper, we examine the time series for temporal data using integrated cluster-based rule induction mining, by presenting a novel algorithm, namely, Integrated Cluster-based Rule Induction Mining for Time-Series analysis (ICRIM-TS). The algorithm discovers temporal data and the similarities between users are evaluated by the proposed measure, Time Series. To our best knowledge, this is the first work on mining and identification of user behaviors based on the time series data with preferences being given for both user relations and temporal data at the same time. Improvement: Through experimental evaluation, under various settings, the performance of web usage mining is evaluated in terms of precision of users, error threshold value and number of rules extracted.

Keywords: Cluster-based Rule Induction Mining, Temporal Data, Time-series Analysis, User Relations, Web Usage Mining

1. Introduction

Temporal data mining addresses on certain tasks comprising of classification, clustering, segmentation, forecasting, and prediction. Web usage Mining has been divided into three sub-groups.

(1) Web Content Mining
(2) Web Structure Mining and
(3) Web Usage Mining

This kind of classification is represented in Figure 1. The Web Content contains the actual textual and other information related to multimedia, the Web structure concentrates on the structure of the Web documents. Conventionally, information provided by integrating web content combined with the web structure has been widely used for searching process and ranking pages returned by a search process for a specific query. The third kind of Web data, Web Usage reveals the behavior of user surfing patterns that has been a
great deal of attention for different types of applications. Owing to the complexity involved in analyzing different user behaviors, the web usage mining has received much attention in the recent times to read the human behavior.

The Web Usage Mining is further divided into three categories based on the nature of data. The first one refers to web server data that relates to the user logs accumulated at web server which comprises of IP address from which the request was placed, the timestamp, the URIs of the requested and referral documents. The second one refers to the application server log that is created dynamically by different application servers. Finally, the third model refers to the application level log which is provided by the user for a specific application.

In this paper, we describe a new approach to discover and apply Integrated Cluster-Based Rule Induction Mining (ICRM) for time series. Our approach combines the following properties in a unique way.

(i) ICRM framework incorporates the clustering model and Induction based decision rule model
(ii) Evolutionary clustering model discovers web data clusters and analyzes the visitor web site behavior and optimally segregate similar user interests
(iii) Rule Induction mining generates inferences and implicit hidden behavioral aspects in the web usage mining to investigate at the web server and client logs
(iv) Achieves flexibility by using time series analysis based on the ICRM framework with varied timing constraints with efficient time series segmentation techniques.

To achieve the properties mentioned above, the work ICRM uses Fuzzy C Means algorithm (FCM), back propagation algorithm and rule induction mining technique using tree structure, by sorting them down the tree from the root node to leaf node illustrated in Figure 1. In addition, we derive a novel similarity measure for time series using Sliding Window Segmentation to support the complexity involved in clustering and to avoid the “curse of dimensionality” problem. The remainder of the article is structured as follows.

In this section some related work is discussed. The novel idea for time series analysis, which we call ICRM-TS, is introduced in Section 2. Section 3 makes a detailed analysis of experiments conducted using zoo data set derived from the UCI repository. Finally, Section 4 summarizes the results and discussions and gives an outlook to future work. With the advancement in hardware and communication technologies, time series is gaining more popularity than ever in different fields ranging from data processing, monitoring intrusion and anomaly detection. A time series in integrated cluster and induction rule mining represents a collection of data values gathered at definite interval of time to determine various characteristics of an entity.

Three different approaches, polynomial, Discrete Fourier Transform (DFT), and probabilistic, to identify the unknown values and answer similarity queries on the basis of the predicted data were proposed1. An efficient algorithm using suffix tree as the underlying data structure is presented2 to detect all the three types of periodicity patterns in data mining. The author presents an algorithm to detect symbol, sequence (partial), and segment (full-cycle) periodicity in time series. The work on periodicity detection in time series data3 is a challenging problem as it is of great importance in many applications. Most work focused on mining synchronous periodic patterns the author focus was on presenting more flexible model of asynchronous periodic pattern using two-phase algorithm. It shows that this algorithm cannot only provide linear time complexity with respect to the length of the sequence but also achieve space efficiency.

The structural analysis of temporal data and the prediction of future data values on the basis of time series are the most important problems that data analysts face in web mining. A new approach was pre-
presented to forecast the behavior of time series based on similarity of pattern sequences by using clustering techniques with the aim of grouping and labeling the samples from given dataset. A novel weighted consensus function by using clustering validation techniques is used to define initial partitions to further divide into consensus partitions from numerous perspectives. An analyzes of an incremental system for clustering streaming time series which comprises of a process Online Divisive-Agglomerative Clustering (ODAC) system which continuously maintains a tree-like hierarchy of clusters on basis of top-down strategy.

Numerous techniques have been designed to extract complex classification rules from time series. A new technique for temporal data mining on the basis of classification rules was provided which can easily be understood by domain experts using the two techniques Time Series Segmentation and Segment Representation. A rule-based classifier by using a generalized Fuzzy-Rough Set (FRS) framework using heuristic algorithms which achieve optimal attribute value to build a rule-based classifier. The use of temporal fuzzy chains for designing dynamical systems also discussed. An unsupervised ensemble learning approach to time series clustering by integrating Rival-Penalized Competitive Learning (RPCL) networks with different form of time series was presented. To perform the evaluation of web mining for time-series analysis based on precision and number of rules extracted using time series segmentation techniques. The brief explanation of the proposed work is explained in section 2.

2. Materials and Methods

Integrated Cluster-based Rule Induction web usage mining for time-series involves the analysis of web log data to identify and represent time series data.

It also obtains the similarity values derived from the patterns which improve the pre-fetch pages. The proposal ICRIM-TS provide with the method of how the time series are segmented and the similarity values for extracting useful information based on these models are evaluated. The framework of ICRIM-TS discussed in Figure 2 involves three major phases which comprises of

1. preprocessing of access log file
2. pattern discovery for time-series data and
3. pattern analysis of access log files.

The mentioned sequential patterns are also applied to web log data.

**Figure 2.** Framework of ICRIM-TS.

2.1 Pre-processing of Access Log

In the first phase, the pre-processing of access log file is performed using raw series that are obtained as input. These raw data may be noisy due to the errors involved in it. The input data are henceforth smoothed, to remove noise. Our work uses weighted average method based on the assumption that the window time size WT and DistWT represent the distance in time between two successive logs. For example, each log value

\[
\text{CurrSmoothedVal} = \frac{(\text{Prev} + \text{Curr})}{\text{Avg}}
\]

Where

\[
\text{Prev} = \frac{(\text{PrevSmoothedVal} \times \text{WT})}{\text{WT}}
\]

\[
\text{Curr} = \frac{(\text{CurrVal} \times \text{DistWT})}{\text{WT} \times \text{DistWT}}
\]

\[
\text{Avg} = \frac{(\text{WT} \times \text{DistWT})}{\text{WT} \times \text{DistWT}}
\]

From the equation 3.1, it is clear that the CurrSmoothedVal is obtained by the weighted average of Prev (product of PrevSmoothedVal and WT), Curr (product of CurrVal and DistWT). The pre-processing
Integrated Cluster-based Rule Induction Mining of Temporal Data for Time-Series Analysis

of access log file is described using pseudo code as illustrated in Figure 3.

```
Input WT, DistWT
// Pre-processing ICRIM-TS
Step 1: Select the window time size WT
Step 2: Select the value for distance between two successive logs
// Remove noise
Step 3: for each logs
   //Evaluate previous smooth logs
   Step 4: Compute Prev = (PrevSmoothedVal * WT)
   // Evaluate current log value
   Step 5: Compute Curr = (CurrVal * DistWT)
   // Evaluate average logs
   Step 6: Compute Avg = (WT * DistWT)
Step 7: Compute CurrSmoothedVal = (Prev + Curr) / Avg
```

Figure 3. Pseudo code for pre-processing of access log files.

The first step involved in pre-processing of ICRIM-TS consists of selecting the window size and the distance between two logs. The second step involved is the evaluation of previous smooth logs, current value and average log obtained from web server.

2.2 Pattern Discovery for Time-series Data

The second phase involved in ICRIM-TS is to discover the pattern for the output log files (pre-processed) obtained in 3.1. Pattern discovery for ICRIM-TS includes segmentation techniques. From the different segmentation algorithms which are in existence, the ICRIM-TS uses sliding-window method to discover similar patterns which is suitable for segmenting time-series obtained in real time. Two of the important threshold factors controlling the segmentation technique are the Error Tolerance Value (ETV), which corresponds to the derived information loss and cost of segment. If the value of ETV = 0, implies no information loss, whereas higher ETV represents the loss related accordingly. The next factor to be evaluated while pattern discovery is to find the cost of segment derived from the least square error.

A simple data set log files consists of m values \( (a_i, b_i) \), where \( i = 1, 2, ..., m \) where \( a_i \) denotes independent variable and \( b_i \) denotes dependent variable whose value is obtained through web servers. The cost of segmentation is obtained through least square method when \( C_{Seg} \) is minimum. From equation 3.2

\[
C_{Seg} = \sum R_i^2
\]

The squared residual value \( (R_i^2) \) is derived as the difference obtained between the dependent variable \( (bi) \) and the predicted model \( (a_i, \beta) \).

Where

\[
R_i^2 = b_i - Fun(a_i, \beta)
\]

The pattern discovery of access log file described below using pseudo code from the equation 3.3 is illustrated in Figure 4.

The pattern discovery for ICRIM-TS is evaluated through three steps. The first step is to select the error threshold value. The second step is to evaluate the cost of the segment and the process is iteratively performed till all the log files are accessed and similar patterns are discovered.

```
Input window size (WS) and value for ETV
Step 1: Select the Error Threshold Value ETV
Step 1: Design a segment with the log values falling in window
Step 2: Calculate Cost of segment C_{Seg}
Step 3: if C_{Seg} < ETV
   // if cost of segment is less than the error threshold value increase the segment value forward
   Seg = seg + WS
Step 4: Else
Step 5: Continue with step 1
```

Figure 4. Pseudo code for pattern discovery.
2.3 Pattern Analysis of Access Log Files

The third phase consists of analyzing the pattern to identify the user interest of log data collected in a web server on basis of time series. Latest trend in web mining for time series involves converting numeric time series to symbolic representation using Symbolic Aggregate Approximation algorithm which requires less storage space. The distance between strings of time series $A'$ and $B'$ is defined in the equation 3.4 as

$$MinDist (A', B') = \sqrt{\sum (dist(A', B'))^2}$$ 3.4

where, the dist function is evaluated using the lookup table for the particular log files. The pattern analysis of access log file is described below using pseudo code is illustrated in Figure 5.

![Pseudo code for pattern analysis.](image)

In the first step the normalization of time series log files from web server are performed. In the second step, the dimensionality of the time series for log file is reduced using PAA which is divided into equal sized segments. In the third step, the discretization of PAA (piecewise aggregate approximation) is obtained. This is done by identifying the number and the location of the breakpoints, determined by statistical lookup tables. The last step involved is evaluating the similarity measure as given in equation 3.4.

3. Results and Discussion

The proposed ICRIM for time series based on its precision of users and inductive rules generated are implemented in Java. The experiments were run on an Intel P-IV machine with 2 GB memory and 3 GHz dual processor CPU. In this section, the performance evaluation tests aimed at comparing the most important properties of time series with respect to precision value, inductive rules generated and error threshold value.

Evaluation is made for ICRIM-TS framework by way of using Zoo dataset for more than 1000 records. For example, in Zoo dataset, the number of persons visited to the zoo are evaluated on the basis of working days and hours and evaluated for different time series with detailed comparison made. The forthcoming section discusses about the results of ICRIM-TS framework.

In this work, we have seen how web mining for ICRIM-TS is designed for time series temporal data to capture the user behaviors for both user relations and temporal data written in mainstream languages such as Java. We run independent tests with growing number of web server and client’s logs, and constant number of service requests sent by each user to web services. The performance graph and table describes the evaluation of the performance of ICRIM-TS using Sliding Window Segmentation (SWS).

The Table 1 describes the performance of the ICRIM-TS using SWS analysis based on precision rate.

| Number of Transaction | Precision (%) |
|-----------------------|---------------|
|                       | Proposed ICRIM-TS | MCCR-TS |
| 100                   | 60             | 53      |
| 200                   | 63             | 55      |
| 300                   | 68             | 60      |
| 400                   | 72             | 65      |
| 500                   | 75             | 70      |
| 600                   | 80             | 75      |

Figure 6 describes the precision for performing the web client and server logs observed from the web server using time-series. Varied number of transactions are used in the experimentation to validate the ICRIM-TS using segmentation techniques. Comparison result of the
ICRIM-TS using Sliding Window Segmentation (SWS) with an existing MCCR-TS based on Gaussian Mixture Model (GMM), measured in terms of precision. When number of transactions obtained from the zoo data set increases, the effectiveness of the process of precision obtained is high in the ICRIM-TS using SWS contrast to an existing GMM. The performance graph of the proposed ICRIM-TS using SWS is shown in the figure 6. The variance in the precision for similarity retrieval would be 10-15% high in the proposed ICRIM-TS using HDM.

![Figure 6. Number of transactions vs precision level.](image)

The Table 2 describes the performance of the ICRIM-TS using SWS based on Inductive rules.

| Data Size (KB) | Proposed ICRIM-TS | MCCR-TS |
|---------------|-------------------|---------|
| 200           | 5.3               | 4.8     |
| 400           | 6.2               | 5.9     |
| 600           | 6.8               | 6.2     |
| 800           | 7.5               | 7.1     |
| 1000          | 7.8               | 7.3     |
| 1200          | 8.2               | 7.9     |

Figure 7 describes the inductive rules generated using zoo dataset. Different ranges of data size are used in the experimentation to validate the ICRIM-TS using SWS. Comparison result of the ICRIM-TS using SWS with an existing GMM, measured in terms of inductive rules. When the data size increases the inductive rules generated also gets increased, the information obtained for similarity retrieval is high in the ICRIM-TS using SWS contrast to an existing MCCR-TS based on SWS. The performance graph of the proposed ICRIM-TS using SWS is shown in the figure 7. The variance in the inductive rules generated for purpose of identifying similarity retrieval is 12-15% high in the proposed ICRIM-TS using SWS.

![Figure 7. Data size vs inductive rules.](image)

The Table 3 described the performance of the proposed ICRIM-TS using SWS based on Error Threshold Value.

| Decisive Rules | Error Threshold Value |
|----------------|-----------------------|
|                | Proposed ICRIM-TS     | MCCR-TS     |
| 2              | 3                      | 5           |
| 4              | 5                      | 6           |
| 6              | 7                      | 9           |
| 8              | 9                      | 12          |
| 10             | 11                     | 14          |
| 12             | 15                     | 18          |

One of the important metrics to be considered for performing ICRIM-TS using SWS is the allowed ETV which indicates the loss of information to provide solution to “curse of dimensionality”. The value of ETV to be 0 symbolizes that no information loss occurs. Higher value of ETV results in raise in loss of information while performing similarity retrieval. Figure 8 describes the number of decisive rules obtained using the sliding window segmentation. Comparison result of the work ICRIM-TS using SWS with an existing MCCR-TS based on GMM, measured in terms of Error Threshold Value (ETV) is made. An increase in the number of decisive rules
results in the higher ETV value but when compared to the existing model, MCCR-TS, the ETV value lessens due to the application of SWS. The performance graph of the proposed ICRIM-T using SWS is shown in the figure 8. The variance in the ETV for obtaining similarity measure would be 20% low in the proposed ICRIM-TS using SWS.

Finally, it is observed that the proposed ICRIM-TS is used for measuring the similarity value obtained during different time intervals on the basis of temporal data. The process of ETV and cost of segmentation is done effectively by the SWS and the web server and client logs similarity search is done efficiently and optimally segregates user interest by solving curse of dimensionality.

Figure 8. Decisive rules Vs ETV.

4. Conclusion

In this paper, ICRIM-TS is efficiently implemented using Sliding Window Segmentation Model for identifying the user interest by evaluating the web server and client logs. The transactions are accessed to perform the similarity measure for different number of transactions by solving curse of dimensionality. The Integrated Cluster Rule Induction Mining using temporal data for time series analysis is measured with metrics such as precision, inductive rules generated and allowed Error Threshold Value to test the effectiveness of information retrieval. The results showed that the proposed ICRIM-TS using Sliding Window Segmentation is 70% better in performing the users’ tasks based on the information retrieval process amongst the web server and client logs available in the web server.

5. References

1. Suyun Z, Eric CCT, Degang C, XiZhao W. Building a rule-based classifier—a fuzzy-rough set approach. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(5):624–38.
2. Alhajj FRMAR. Efficient periodicity mining in time series databases using suffix trees. IEEE Transactions on Knowledge and Data Engineering. 2010; 23(1):79–94.
3. Alvarez MF, Lenguajes AD, Sevilla S, Spain S. Energy time series forecasting based on pattern sequence similarity. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(8):1230–43.
4. Lian X, Chen HKL. Efficient similarity search over future stream time series. IEEE Transactions on Knowledge and Data Engineering. 2008; 20(1):40–54.
5. Wang W, Yu YPS. Mining asynchronous periodic patterns in time series data. IEEE Transactions on Knowledge and Data Engineering. 2003; 15(3):613–28.
6. Rodrigues PP, Porto P, Gama J, Pedroso J. Hierarchical clustering of time-series data streams. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(5):774–87.
7. Fisch D, Gruber T, Sick B. Swift rule: Mining comprehensible classification rules for time series analysis. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(5):774–87.
8. Moreno-Garcia J, Castro-Schez T, Jimenez LJ. A fuzzy inductive algorithm for modeling dynamical systems in a comprehensible way. IEEE Transactions on Fuzzy Systems. 2007; 15(4):652–72.
9. Yang Y, Chen K. Temporal data clustering via weighted clustering ensemble with different representations. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(2):307–20.
10. Yang Y, Chen K. Time series clustering via rpcl network ensemble with different representations. IEEE Transactions on Systems Man and Cybernetics, Part C: Applications and Reviews. 2011; 41(2):190–9.
11. Syed SA, Kumaran TS, Ahmed AS. Energy efficiency distributed routing algorithm based on HAC clustering method for WSN. Indian Journal of Science and Technology. 2014; 7(57):66–75.
12. Hariharan R, Mahesh C, Prasenna P, Kumar RV. Enhancing privacy preservation in data mining using cluster based greedy method in hierarchical approach. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–8.
13. Murugananthan V, Kumar BLS. An adaptive educational data mining technique for mining educational data models in e-learning systems. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–5.
14. Sivakumar S, Venkataraman S, Selvaraj R. Predictive modeling of student dropout indicators in educational data mining using improved decision tree. Indian Journal of Science and Technology. 2016 Jan; 9(4):1–5.

15. Undavia JN, Dolia P, Patel A. Customized prediction model to predict post-graduation course for graduating students using decision tree classifier. Indian Journal of Science and Technology. 2016 Mar; 9(12):1–7.