A Survey on Periodic pattern mining Techniques in Time Series Database

[1]Savaliya Hemali, [2]Khachane Hetal, [3]Dharmesh Bhalodiya

[1,2]PG Students, [3]Assistant Professor
Department of Computer Engineering, BHGCET, Rajkot

***Abstract*: Time series data mining is become new emerging topic for research in knowledge discovery. Time series database consist of the sequence of values on data point where usually their order is define by the time when they are recorded. In this paper brief overview of some recently important time series data mining methods and applications are presented.**

Keywords: Data mining, Time series, periodic pattern mining

INTRODUCTION:

Time series is a set of observations each one of being records of a specific time period. A time series database consists of sequence of eventsor values obtained over repeats measurement of time. The values are measures at equal time intervals. [1]Mining in time series data considered as an important analysis in many applications like finance, whether forecasting, economics, stock market, shopping mall, medical and fraud detection.

The various issues in mining time series are [2]:

- How to find the relationships within the time- series data?
- How to represent and index the time series data?
- How to find the similar patterns or trends after analysis of huge amount of complex time series data?
- How to store the data as data increases linearly with time. Due to which the need for the storage will also increase and the pattern search process will slow down exponentially**.**

I. MINING IN TIME SERIES DATA

The final goal of the data mining system to find the hidden information or knowledge from the original or the transformed data. Mining in time series include verity of task like classification, clustering, prediction, segmentation, pattern mining and Association rule mining.

A. CLASSIFICATION

Classification is most familiar and most popular data mining Technique. Classification of unlabeled time series to existing classes is a further traditional data mining task. Classification maps input data into predefined groups. It often referred to as supervised learning, as the classes are determined prior to examining the data; a set of predefined data is used in training process and learn to recognize patterns of interest.
Two most popular methods in time series classification include the Nearest Neighbor classifier -Nearest Neighbor method applies the similarity measures to the object to be classified to determine its best classification based on the existing data that has already been classified. Decision trees- a set of rules are inferred from the training data, and this set of rules is then applied to any new data to be classified

Time-series classification
Distance based classification k-nearest neighbor (KNN)[3], Euclidean distance (ED)[4] ,Feature based classification :feature-extraction technique is to transform the time-series data into the frequency do-main, where data dimensionality can be reduced.DFT(DiscreteFourier Transform), DWT (Discrete Wavelet Transform) and SVD (Singular Value Decomposition)[5].

B. CLUSTRING

Clustering is a datamining technique where similar data are placed into related homogeneous groups without advanced knowledge of the groups definitions [6] The goal of clustering is to identify structure in an unlabeled data set by objectively organizing data into homogeneous groups where the within-group object similarity is minimized and the between-group-object dissimilarity isMaximized. [7]

General methods of time series clustering are:

- Partitioning Clustering methods arek-means algorithm [8,9], the k-medoidsalgorithm [10], and the fuzzy c-meansalgorithm [11].
- Hierarchical Clustering methods areChameleon [12], CURE [13] and BIRCH [14].

- Density-basedclusteringmethods are DBSCAN [15] and OPTICS [16].
- Grid-based clustering methods are STING [17].

C. PREDICTION

Time series prediction is used to predict future values of a time series data based on the past observations. For prediction it is necessary to build a model that describe the behavior of the observed variable over time. A regression model is the model where the values of time series are fitted to a curve with some certain error. The most popular linear model fit where the values of time series are fitted to a straight line.Higher-level models usedif the next value of a variable depends on the previous [18].Most popular stochastic time series models is the ARIMA -Autoregressive Integrated Moving Average [19,20]ARIMA model is further explored for other subclasses of models like Autoregressive (AR),Moving Average (MA) and Autoregressive Moving Average (ARMA). Box and Jenkins [21] had proposed verysuccessful variation of ARIMA model for knownSeasonal time series forecasting, the Seasonal ARIMA (SARIMA).

D. SEGMENTATION

Time-series segmentation is a method in which an time-series data divided into a sequence of discrete segments to reveal the underlying properties of its source. [22] This technique is used to alternate the way the time series are represented for faster and easier access.For example, piecewise aggregate approximation(PAA) divides time series intoSubparts of equal length and each block of data points is substituted with one value[23] Another approximation technique is symbolic Aggregate approximation SAX[24]uses the pricewise aggregate approximation of time series as a part of its algorithm.

E. ASSOIATION RULE

Association rules discovery is the task of Finding existing relations or patterns in the data. In time series data mining this task is similar to find relations or patterns between different attributes of the time series. To properly define these attributes is a crucial step in association rule mining. Association rules are interpretable by humans because they are close to natural language. For that reason, experts can gain valuable knowledge about the application domain. [25]

F. PATTERN MINING:

Pattern Discovery is the task of finding the pattern of interest. Pattern is the local structure in the data. It would typically be like a substring with a don't care character etc. The problem of pattern discovery is to find all the interesting patterns in the data. There are many methods in time series for determining patterns which one can look for in the data like periodic pattern mining, partial periodic pattern, recurring pattern etc.

Many time series datasets have an inherent periodic structure. Therefore, detecting periodicity is another classical pattern discovery task. Besides classical time series analysis methods for handling seasonality and periodicity. Time series data mining community produced techniques for huge data sets. [26] used for mining for partial periodic patterns as in many data applications full periodic patterns appear not that frequently.[27]aim to mine for the periodicity rate of a time series database.[28] used for find the recurring pattern in time series database.

II. LITRATURE SURVEY:

A. Periodic-frequent Pattern:

Periodic-frequent Pattern tree is a novel concept of mining periodic-frequent patterns from transactional databases. In this method an efficient tree-based data structure used it called Periodic-frequent pattern tree (PF-tree in short), that captures the database contents in a highly compact manner and enables a pattern growth mining technique to generate the complete set of periodic-frequent patterns in a database for User-given periodicity and support thresholds. PF-tree provides, a highly compact tree structure to capture the database content, and a pattern growth based mining technique to discover the complete set of periodic-frequent patterns on the user-given maximum periodicity and minimum support thresholds over a transactional database [29].

B. MaxCPF-Tree

The basic model of periodic-frequent patterns is based on the notion of "single constraints", use of this model to mine periodic-frequent patterns containing both frequent and rare items the main dilemma with the model called the "rare item problem." To confront the problem, an alternative model based on the notion of "multiple constraints" has been proposed in this model. The periodic-frequent patterns discovered with this model do not satisfy downward closure property. MaxCPF-Tree is efficient model based on the notion of "multiple constraints." The periodic-frequent patterns discovered with this model satisfy downward closure property so that periodic-frequent patterns can be efficiently discovered. A pattern Growth approach has also been proposed for efficient mining of periodic-frequent patterns,

periodic-frequent patterns mined with the proposed model satisfy downward closure property. This increases the search space, which in turn increases the computational cost of mining the patterns[30]

C. QPF-Growth algorithm

QPF-Growth algorithm Introduce a new class of user-interest based frequent patterns, called quasi-periodic-frequent patterns. A frequent pattern is said to be quasi-periodic-frequent if most of its occurrences are periodic in a database. The proposed model and a pattern-growth algorithm used for discover these patterns. In this method three pruning techniques to reduce the computational cost of mining the patterns [31].

Technique	Advantage	Disadvantage
Periodic-frequent Pattern tree	Efficient technique to mine periodic frequent pattern	Require more search space
MaxCPF- tree	Satisfy downward closure property and solve the problem for rare item	Computational cost is high
QPF-growth algorithm	Reduce the complexity of time and memory	Do not satisfy downward closure property
PFP-Growth ++ algorithm	Provide greedy search method to prune non-periodic frequent pattern	Computational cost is high
RP- Growth algorithm	Solve the rare item problem	Tree Need more space and memory

D. PFP-Growth ++

In this method applied the greedy search on a pattern's tid-list to determine whether it is a periodic-frequent or a non-periodic-frequent Using the concept of local-periodicity, in this introduce two novel pruning techniques and extend them to improve the performance of PFP-growth. The algorithm called as PFP-growth++. The proposed techniques facilitate the PFP-growth++ to prune the non-periodic-frequent patterns with a suboptimal solution, while finds the periodic-frequent patterns with a global optimal solution. Thus, to not miss any knowledge pertaining to periodic frequent patterns [32]

E. RP –growth algorithm

In this introduced a new class of partial periodic patterns known as recurring patterns, it used for discovering recurring patterns. For mine the recurring pattern proposed a novel pruning technique to reduce the computational cost of finding Recurring patterns. To mine therecurring pattern RP-growth algorithm used to discover the recurring patterns effectively [28].

I. COMPARISON OF THE TECHNIQUES

V. CONCLUSION

Analysis and applications of time series have become progressively more important in a variety of fields of research in business, economics, engineering, medicine, social science, environ metrics, politics, and others. Time series analysis, modeling and forecasting are fundamental importance in time series. In the paper we presented several contemporary, innovative, and important methods and discussed the possible applications of time series

REFERENCES

[1] Book (Mining Stream, Time-Series, and Sequence Data).

[2] Garima, Sangeeta Rani," Review on Time Series Databases and recent research trends in Time Series Mining" 5th International Conference- Confluence The Next Generation Information Technology Summit,IEEE2014.

[3] Xing, Z., Pei, J., and Keogh, E. (2010).A brief survey on sequence classification.ACM SIGKDD Explorations Newsletter, 12(1):40{48.

[4] Keogh, E. J. and Kasetty, S. (2003). On the need for time seriesdata mining benchmarks: A survey and empirical demonstration. Data Min. Knowl.Discov.7(4):349{371.

[5] Yang, K. and Shahabi, C. (2004). A pca-based similarity measurefor multivariate time series. In ACM International Workshop On Multimedia Databases:Proceedings of the 2 nd ACM international workshop on Multimedia databases, volume 13,pages 65{74.
[6] T.WarrenLiao,"Clustering of time series data—a survey",Pattern Recognition 38 (2005) 1857 – 1874.
[7] K. Krishna, M.N. Murty, Genetic *k*-means algorithms, IEEETrans. Syst. Man Cybernet.-B: Cybernet. 29 (3) (1999)433–439.
[8] L. Meng, Q.H. Wu, Z.Z. Yong, A genetic hard *c*-meansclustering algorithm, Dyn. Continuous Discrete ImpulsiveSyst. Ser. B: Appl. Algorithms 9 (2002) 421–438.
[9] V. Estivill-Castro, A.T. Murray, Spatial clustering for datamining with genetic algorithms, http://citeseer.nj.nec.com/estivillcastro97spatial.html.
[10] L.O. Hall, B. Özyurt, J.C. Bezdek, Clustering with agenetically optimized approach, IEEE Trans. EvolutionaryComputat. 3 (2) (1999) 103–112.
[11] G. Karypis, E.-H. Han, V. Kumar, Chameleon: hierarchicalclustering using dynamic modeling, Computer August (1999)68–75.
[12] S. Guha, R. Rastogi, K. Shim, CURE: an efficient clusteringalgorithm for large databases, Proceedings of the 1998 ACMSIGMODInternational Conference on Management of Data,Seattle, WA, June 1998, pp. 73–84.
[13] T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: an efficientdata clustering method for very large databases, Proceedingsof the 1996 ACM-SIGMOD International Conference onManagement of Data, Montreal, Canada, June 1996, pp.103–114.
[14] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A densitybasedalgorithm for discovering clusters in large spatialdatabases, Proceedings of the 1996 International Conferenceon Knowledge Discovery and Data Mining (KDD'96),Portland, OR, 1996, pp. 226–231.
[15] M. Ankerst, M. Breunig, H.-P. Kriegel, J. Sander,OPTICS: ordering points to identify the clustering structure,Proceedings of the 1999 ACM-SIGMOD InternationalConference on Management of Data, Philadelphia, PA, June1999, pp. 49–60.
[16] W. Wang, J. Yang, R. Muntz, R., STING: a statisticalinformation grid approach to spatial data mining, Proceedingsof the 1997 International Conference on Very Large Data Base(VLDB'97), Athens, Greek, 1997, pp. 186–195.
[17] Tan Yan, Liudmila Ulanova, Ye Ouyang and Fengyuan Xu," Data Mining in Time Series: Current Study and Future Trend", Journal of Computer Science 10 (12): 2358-2359, 2014.
[18] Zhang G.P. (2003) Time series forecasting using a hybrid ARIMAand neural network model.Neurocomputing 50: 159–175
[19] Cochrane J.H. (1997) Time Series for Macroeconomics andFinance.Graduate School of Business, University of Chicago.
[20] Box G.E.P., Jenkins G. (1970) Time Series Analysis, Forecastingand Control. Holden-Day, San Francisco, CA.
[21] Eamonn Keogh Selina Chu David Hart Michael Pazzani," Segmenting Time Series: A Survey and Novel Approach", Department of Information and Computer Science University of California, Irvine, California 92697.
[22] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. (2001)Dimensionality reduction for fast similarity search in large timeseries databases. Knowledge and Information Systems, vol. 3, no.3, pp. 263–286
[23] J. Lin, E. Keogh, L. Wei, and S. Lonardi, (2007) "Experiencingsax: a novel symbolic representation of time series," Data Miningand Knowledge Discovery, vol. 15, no. 2, pp. 107–144.
[24] SRIVATSAN LAXMAN and P S SASTRY," A survey of temporal data mining",S¯adhan¯aVol. 31, Part 2, April 2006, pp. 173–198.
[25] Han J, Dong G, Yin Y 1999 Efficient mining of partial periodic patterns in time series database. InProc. 15th Int. Conf. on Data Engineering, (ICDE'99), Sydney, pp 106–115
[26] Laxman S, Sastry P S, Unnikrishnan K P 2005 Discovering frequent episodes and learning hiddenmarkov models: A formal connection. IEEE Trans. Knowledge Data Eng. 17: 1505–1517
[27] R. Uday Kiran, Haichuan Shang, Masashi Toyoda and Masaru Kitsuregawa, "Discovering Recurring Patterns in Time Series", Proc.18thInternational Conference on Extending Database Technology (EDBT), March 23-27, 2015
[28] Syed KhairuzzamanTanbeer, Chowdhury Farhan Ahmed, Byeong-Soo Jeong, and Young-Koo Lee," Discovering Periodic-Frequent Patterns in Transactional Databases", PAKDD 2009, LNAI 5476, pp. 242–253, 2009.,Springer-Verlag Berlin Heidelberg 2009.
[29] AkshatSurana, R. Uday Kiran, and P. Krishna Reddy," An Efficient Approach to Mine Periodic-Frequent Patterns in Transactional Databases", Cao et al. (Eds.): PAKDD 2011 Workshops, LNAI 7104, pp. 254–266, 2012., Springer-Verlag Berlin Heidelberg 2012.
[30] R. Uday Kiran and Masaru Kitsuregawa," Discovering Quasi-Periodic-Frequent Patterns in Transactional Databases", BDA 2013, LNCS 8302, pp. 97–115, 2013.," Springer International Publishing Switzerland 2013.
[31] R. Uday Kiran and Masaru Kitsuregawa," Novel Techniques to Reduce Search Space in Periodic-Frequent Pattern Mining", S.S. Bhowmick et al. (Eds.): DASFAA 2014, Part II, LNCS 8422, pp. 377–391, 2014, Springer International Publishing Switzerland 2014.