International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) ISSN: 2395-5325
Volume 3, Special Issue 3 (September '2017)
Proceeding of - International Conference on Recent Trends in Computational Life Science and Information Technology,
Conference held at Tiruppur Kumaran College for Women,Tirupur,Tamilnadu, India.

ANALYSIS OF SEQUENTIAL PATTERN MINING FOR EVENT DETECTION USING APRIORI ALGORITHM

G.Divya,
Department of Computer Science & Application,
Faculty of Park's College (Autonomous),
Chinnakarai, Tamilnadu, India.

R.Nithyaananthi,
Department of Computer Science & Application,
Faculty of Park's College (Autonomous),
Chinnakarai, Tamilnadu, India.

Abstract: In this paper we propose a pattern mining technique to study event detection representation from difficult multivariate temporal data, such as electronic health reports. Pattern recognition is seen as a main challenge within the field of data mining and knowledge discovery. In this paper, we propose a comprehensive data mining framework for event detection DP miner, which functions in a distributed and parallel manner (data in a partitioned database processed by one or more sensor processors) and is able to extract a pattern of sensors that may have event information with a low communication cost. To achieve this, we introduce a new sensor behavioral pattern mining technique called sequential data mining. The task of sequential pattern mining is a data mining task specialized for analyzing sequential data, to discover sequential patterns. More specifically, it consists of discovering interesting subsequences in a set of series, where the interestingness of a subsequence can be calculated in terms of a range of criteria such as its occurrence frequency, length, and profit. In order pattern mining has several real-life applications due to the actuality that data is naturally encoded as series of symbols in many fields such as bioinformatics, e-learning, market basket analysis, texts, and web page click-stream analysis. An Apriori algorithm has been proposed for data preparation, to generate sequential sensor patterns. Evaluation results show better trade-off between Sequential data mining and Differential data mining. An analysis for communication cost is also evaluated here.

Keywords: Apriori, DP miner, Multi variant Temporal data, Pattern Mining.

I. INTRODUCTION

By data the definition refers to a set of facts (e.g. records in a database), whereas pattern represents an expression which describes a subset of the data, i.e. any structured representation or higher level description of a subset of the data. The term process designates a complex activity, comprised of several steps, while non-trivial implies that some search or inference is necessary, the straightforward derivation of the patterns is not possible. The resulting models or patterns should be valid on new data, with a certain level of confidence. Also, we wish that the patterns be novel at least for the system and, ideally, for the analyst and potentially useful, i.e. bring some kind of benefit to the analyst or the task. Ultimately, they need to be interpretable, even if this requires some kind of result transformation. A significant idea is interestingness, which generally quantifies the added value of a pattern, merge novelty, validity, utility and simplicity.

This can be uttered either implicitly, or explicitly, through the ranking carried out by the DM scheme on the returned patterns. Recently, extracting knowledge from sensor data has received a great deal of attention in the data mining community. Traditional data mining schemes focusing on association rules, frequent patterns, sequential patterns, clustering, and classification have been successfully used on sensor data. These mining schemes are usually centralized and computationally expensive, and they focus on disk-resident transactional data. A decent number of data mining algorithms have been developed with less computational complexity, and the process of forming pat-terns and producing association rules is straightforward. Metrics, rules, binary patterns, and frequent patterns are often used as indicators to find interesting knowledge. They require excessive interactions and rule exchanges, leading to massive amounts of communication. Such a change leads to further interaction in the cyber systems in terms computation, communications, etc. Thus, in an intended CPS, physical system aspects and cyber system aspects should be tightly combined. The wireless sensors in the CPS produce a huge volume of dynamic, geographically-distributed and heterogeneous data when deployed in these applications.

The raw data, if accurately analyzed and transformed to usable information through data mining, can facilitate automatic and intelligent decision-making on specific events of interest (e.g., damage in aerospace vehicles, chemical explosion), while optimizing the resource efficiency of cyber systems. Hence, it is vital to develop methodologies to mine sensor data. A cluster of sensors shares data patterns so that each individual sensor can calculate a sensor pattern. The sensors in a cluster coordinate with their cluster head (CH), and together, they develop a differential data pattern tree structure, called DP-Tree in a distributed and parallel manner for data mining. The CH, along with its sensors, find an initial differential sensor pattern (DSP) via the DP-Tree. After mining all initial DSPs, the CH provides a confirmed DSP that can ensure whether an event has occurred around some sensors or the cluster, even offering a value (e.g., e V

> 1) as the event indicator. In DPminer, the sensors which are not in a DSP are dropped with their data from further pattern mining, thereby reducing the communication cost. Instead of finding binary frequent patterns, we find sensor patterns that come from the consideration of different rates of frequencies and values in the Cases and Controls. Generating such a DSP from a network can be very useful in a wide range of applications that require fine-grained monitoring of physical environments.

We assume that the whole event detection time (Qw) is divided into Q periods. Each period includes further q slots, i.e., {t 1 ,t 2 ,...,t q } such that tk +1 - t k = τ , which is the length of each time slot. We assume that a sensor database DB can be partitioned into d sub-databases, i.e., DB 1,DB 2 ,.. . ,DB d. One of the sub-databases (e.g., DB 1) of a sensor contains prepared data that is collected in period Q. This arriving dataset is a large dataset which we call Cases. Other sub-databases are shared with the neighbors in a cluster. Each sensor mines both different values/items (V) and different frequencies (F) f these values from Cases and determines a set of tuples within time slot tk. Here, we maintain a Controls database/dataset which contains the healthy data (for comparison with the data in Cases). If there is an event, each tuple may have frequent values with higher event intensity. This can be determined through comparison with data in Controls. In Cases, a set of tuples denoted by H h (h =1, 2 ... , n) is defined as a subset of data (frequencies and values) of a particular sub-database. From Cases , we first find a rate of frequencies (rf) and median values (mv) in H h , a subset S s of sensors, and DB i.

II.RELATED WORK

There are various data mining techniques outlined in the literature, including frequent patterns, sequential patterns, clustering, and classification. They already address numerous issues in data mining including execution time, complexity, and rule or query processing needed to mine stored (static) and/or stream data [3]–[6], [9], [10]. In the recent decades, mining association rules have been used in transactional databases. Recently, they have been applied to data mining schemes in sensor networks. Mining the associations among sensor values that co-exist temporally in large-scaled WSNs and mining spatial temporal event patterns from sensor data are proposed in [9], [11]. A behavioral pattern named Target-based Association Rules (TARs) for point-of-coverage in WSNs which aims to discover the correlation among a set of targets monitored by a WSN and uses confidence metrics is proposed [9].
In TARs, every sensor maintains an additional flash memory that increases the deployment cost. An interesting data mining technique in wireless ad hoc networks uses a tree-based structure called Positional Lexico-graphic Tree (MAR-PLT for short) to mine association rules in which the event-detecting sensors are the main objects [5]. It follows a FP-growth-like pattern growth mining technique, but the two database scanning requirements and the extra MAR-PLT update operations during mining limit efficient use of this technique in handling WSN data. Association rules-based growth trees do not show satisfactory performance in WSNs in terms of communications.
A method which captures association-like co-occurrences as well as temporal correlations (linked with such co-occurrences) is used to mine associated patterns from sensor data streams [7].

A regular frequent pattern is proposed to find frequent sensor patterns that occur after a certain interval in the sensor database. Most of these techniques consider a binary (0/1) occurrence of the patterns in the database. Binary value (0/1) is also used for frequent pattern association. Such a binary occurrence or pattern association may fail to detect events in practice. In addition, they still require significant communication costs in terms of excessive message transmission in the WSN. Also, there is a lack of analysis of the costs in WSNs. We observe that current data mining schemes using association rules, associated pattern, data clustering, and so on do not show satisfactory performance in terms of communication and event detection in sensors of IoT. These issues have not been specifically addressed before. Our framework DPminer is an attempt to overcome these shortcomings while detecting an event through a DSP. In another work [9], generating context association rule over an online sensor/actuator transactional data stream is suggested. This is used to invoke proper operations of actuators relevant to values of the sensors. It organize frequent context item sets over the current data steam, such that a set of frequently co-occurred sensors and actuators items is arranged.

2.1. Sp-Tree Performance

A data mining technique to generate association rules from WSN by using a prefix-tree called sensor pattern tree (SP-tree). Although SP-tree shows better performance than PLT [16], it still generates a large number of rules, many of which may not be useful enough, resulting in a significant communication cost and computation costs. A method which captures association-like co-occurrences as well as temporal correlations (linked with such co-occurrences) is used to mine associated patterns from sensor data streams [12]. A regular frequent pattern is proposed to find frequent sensor patterns that occur after a certain interval in the sensor database.

Most of these techniques consider a binary (0/1) occurrence of the patterns in the database. Binary value (0/1) is also used for frequent pattern association. Such a binary occurrence or pattern association may fail to detect events in practice. In addition, they still require significant communication costs in terms of excessive message transmission in the WSN. Our work is most associated to pattern discovery from sequential data, which include time series, event sequences, and spatio-temporal trajectory. Mannila et al. [10] inspect the discovery of frequent episodes from event sequences. An episodes is a (incompletely or completely) ordered list of events, thus is a alternate of sequential pattern. A fixed sliding window w is used to extract segments (i.e., subsequences) in the event series, and the contribution of every section to each candidate episode's frequency is counted. The segments supporting one episode may overlap, which is reasonable

since episodes try to capture the appearing order of instantaneous events

2.2. ***Segmentation***

Though, this methodology may not get satisfactory results in finding spatio-temporal patterns, for several reasons. First, the window restrictions the length of the patterns. Second, pattern supports may not be counted correctly. E.g., the object's association is aabbcdefg , where each character a, b , etc. corresponds to a spatial region. The occurrence of the pattern abc should be 1, since the object moves from a to c, once. How-ever, if w is 5, pattern abc has support 4 due to the contribution of 4 segments (a b c, ab c, abc, and a bc). Third, as opposed to well-defined categorical values for event in-stances, object locations do not repeat themselves exactly in pattern instances, for they are usually ordinal and inexact. Yang et al. investigated mining long sequential patterns in [13], also dealing with event series with noise. Previous work on detecting patterns from time-series (e.g, [2, 7]) converted the problem to finding subsequences in lists of categorical data (e.g., event sequences), by pre-processing the original sequence to a string. A window w of fixed size is slided along the sequence, and a subsequence with length w is extracted for every position.

In [2], the subsequences are clustered based on their shapes, and each cluster is given an id. In [7], some features are extracted from each subsequence (e.g., the slope of the best fitting line of the sub-series, the mean of the signal, etc.). The feature space is divided into groups of similar values, and every subsequence is converted to a group-id. The raw sequence is then transformed to a string of cluster-ids or group-ids. The use of the window may over-count the patterns due to the reason explained above. In addition, since w is fixed, the ex-tracted subsequences have the same length, which may affect the resultant patterns. Furthermore, for spatio-temporal data, even when we extract the subsequences using a slid-ing window and get simple features from these segments, we cannot directly group these features using methods in [2] and [7].

III.BACKGROUND STUDY

The trouble of mining association rules over basket data was introduced. An example of such a rule might be that 98% of customers that obtain tires and auto accessories also get automotive services completed. Searching all such policies is precious for cross-marketing and attached mailing function. Other functions consist of catalog design, add-on sales, store layout, and customer segmentation based on buying patterns.

The databases concerned in these applications are very large. It is imperative, consequently, to have fast algorithms for this job. Algorithms for find out huge item sets make various passes over the data. In the initial pass, we count the support of individual items and verify which of them are big, i.e. have least support. In each subsequent pass, we initiate with a seed set of item sets originate to be huge in the previous pass. We utilize this seed set for producing new potentially large item sets, called candidate item sets, and calculate the actual sustain for these candidate item sets through the pass over the data. At the finish of the pass, we locate out which of the candidate item sets are actually large, and they happen to the seed for the after that pass. This method carry on until no new large item sets are found.

IV. PROPOSED METHODOLOGY

Association rule mining techniques are briefly given for classify the relationship and the association is among within the values of category variables using large data sets. The tasks of many data mining projects in data mining of sub group text mining things. In such a way we use the Association rule mining techniques in the way of mining the items from the databases, in new way to solve the existing problem that, mining of hidden data of the medical document-data, such a way it works in efficient manner to work with in it. These Enhanced techniques have a wide range of applications in many areas of business practice and also research from the analysis of document data preferences or resource management, to the history of medical data. These techniques enable analysts and researchers to uncover hidden patterns in large data sets, such as "medical document to search how to cure the disease" (those diseases are came already in the history year, in such a way we need to handle the disease by checking the databases which have been hidden that already stored in the databases) such that the association rule mining helps the algorithm very efficiently by example "client who order the item A often also order the item B or C" or already mined data history tells us the positive things about the searching history is X its frequently taken from the databases in that we have rapidly huge large data sets for the such associations based on predefined values for mining of document by "condition" || "threshold" || "value. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.

By proceeding of frequent individual items the databases is fully expanded of recognizing the individual frequent items that appear of sufficiently in databases. The frequent sets on item that is determined through Apriori it can be used in association rules of highlight trends in databases. Media, Primitive & Composite Events: A media event is modeled as a time stamped 4 -tuple, e = (u, s, r, a): t where u corresponds to a user (subject), s a region, r a resource, the category of the activity, and t designates the time the event occurred. The event e is interpreted as a user u performed activity a, in space s, involving resource r at time t. For Instance, a media event (BOB, LOADING DOCK,*, ENTRY):3:00pm represents that "Bob" entered the "loading dock" at "3:00pm". The resource here was unspecified (*) (could be a null value). The notation e.u will refer to the user associated with event e. In general, media events are domain dependent, with the specification of and mechanisms to detect them being implemented by the designer of the pervasive application. Using such mechanisms, a media stream can be converted into a stream of media events. The set of all media events that can be generated in this space is finite & enumerable they are a subset of the cartesian product of the set of all users, spatial regions, resources and activities that can be detected by the sensors.

The candidate generation algorithm illustrated, that except we substitute INPUT with the k-large item set. Then the output will be possible left hand body of the association rule. For example, if the 3-large item set is {1,2,3}, we will use the loop + recursion algorithm to generate {1,2}, {1,3}, {2,3}, {1,2,3} as the left-hand body, and the trivial one {1,2,3} → {1,2,3} will be discarded. Accordingly, we might mine out the following association rules (we need further prune, prune techniques will be discussed later)

{1,2} → {1,2,3} (equivalent to {1,2} → {3})
{1,3} → {1,2,3} (equivalent to {1,3} → {2})
{2,3} → {1,2,3} (equivalent to {2,3} → {1})

Input: database D, Mini Support ∈, Mini Confidence ∈
Output: Rt All association rules
Method:
1- L1 = large 1-itemsets;
2- for(k=2; Lk-1 ≠ ∅ ; k++) do begin
3- Ck =apriori-gen(Lk-1); //generate new candidates from Lk-1
4- for all transactions T ∈ D do begin
5- Ct=subset(Ck,T); //candidates contained in T.
6- for all candidates C ∈ Ct do
7- Count(C)=Count(C)+1; // increase support count of C by 1
8- end
9- Lk={C ∈ Ct | Count(C) ≥ ∈ × |D|}
10- end
11- Lf = ∪k Lk
12 Rt=GenerateRules(Lf , ∈)

The patterns are very big summary, it has two instances such locations are not fall in same cell such as two adjacent position appear in neighbor cells, the frequent pattern instances are separated between in different grid based patterns. The first problem could be alleviated by decreasing G, however, this would increase the chances of missing patterns due to the second problem. An alternative conversion technique adds the ids of cells that intersect with the line segments connecting consecutive locations to the transformed sequence. Thus, we need a better way to abstract the trajectory. Motivated by line simplification techniques, we represent segments of the spatio-temporal series by directed line segments. Line segment l summarizes the first three points in each of the three runs with little error. In this way, not only do we compress the original data, decreasing the mining effort, but also the derived line segments (which approximately describe movement) provide initial seeds for defining the spatial regions, which could be expanded later by merging similar and close segments.

V.CONCLUSION

In our research, we modeled the difficulty of mining sequential patterns from data by regard as both spatial and temporal information. Particular frequent patterns are found efficiently, by combining segments not only by similar shape, but also by closeness in space. In addition, we employed special properties of the trouble and a newly projected substring tree to accelerate search for longer patterns. The proposed hierarchical temporal association mining method offers a robust solution to explore and employ the feature temporal patterns with respect to the events of interest. This methodology efficiently addresses the issues of loose structure and skewed data distribution. Our proposed system emulates the efficient mining of event detection. In an general, the association rule mining produce a set of itemsets (retail transaction for item purchased), In here the algorithm gives to find the sub-sets in which the common of least minimum number of itemsets C.

The Apriori uses a "Bottom-up" approach it is also used to extend the frequent sub-sets were it is extended with one item at a time. The groups of candidates are tested with data and the algorithm is terminated with successful extensions. The common rule mining using the association rule mining is given a set of itemsets in our proposed research we have given that event detection through making list of items, in such a way find the subsets of each events which have been given from the datasets, it uses the approach to candidate generation and the algorithm is easily used for operate datasets, it contains of events such as collection of events one by one. Need to extract a pattern of sensors which is given in the information, through the extraction, the event detection is present of sensor pattern mining which is considering in actual data. The technique is followed for mining framework, the A-Priori algorithm, works with actual values read from the sensor that is given through datasets using binary decision for event detection, through this in future we enhance the event detection through advanced or hybrid A-priori with other algorithms to use more datasets.

VI.REFERENCES

[1] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna, "Semantic Event Detection via Temporal Analysis and Multimodal Data Mining," IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia, vol. 23, no.2, 2006, pp. 38-46. S.-C. Chen, M.-L. Shyu, and C. Zhang, "Innovative
[2] Shot Boundary Detection for Video Indexing," Edited by Sagarmay Deb, Video Data Management and Information Retrieval, Idea Group Publishing, 2005, pp. 217-236.
[3] S.-C. Chen, M.-L. Shyu, C. Zhang, and M. Chen, "A Multimodal Data Mining Framework for Soccer Goal Detection Based on Decision Tree Logic," International Journal of Computer Applications in Technology, vol. 27, no. 4, 2006, pp. 312-323.
[4] A. Ekin, A. M. Tekalp, and R. Mehrotra, "Automatic Soccer Video Analysis and Summarization," IEEE Transactions on Image Processing, vol. 12, no. 7, 2003, pp. 796-807.
[5] Y.-L. Kang, J.-H. Lim, Q. Tian, and M. S. Kankanhalli, "Soccer Video Event Detection with Visual Keywords," in Proceedings of IEEE Pacific-Rim Conference on Multimedia, vol. 3, 2003, pp. 1796-1800.
[6] R. Leonardi, P. Migliorati, and M. Prandini, "Semantic Indexing of Soccer Audio-visual Sequences: A Multimodal Approach based on Controlled Markov Chains," IEEE Transactions on

Circuits and Systems for Video Technology, vol. 14, no. 5, 2004, pp. 634-643.
[7] C. G. M. Snoek and M. Worring, "Multimedia Event-Based Video Indexing Using Time Intervals," IEEE Transactions on Multimedia, vol. 7, no. 4, 2005, pp. 638-647.
[8] P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Addison Wesley, ISBN: 0-321-32136-7.
[9] R. Vilalta and S. Ma, "Predicting Rare Events in Temporal Domains," in Proceedings of IEEE International Conference on Data Mining, 2002, pp.474-481.
[10] F. Wang, Y.-F. Ma, H-J. Zhang, and J.-T. Li, "Dynamic Bayesian network based event detection for soccer highlight extraction," in Proceedings of International Conference on Image Processing, vol. 1, 2004, pp. 633-636.
[11] T. He, et al., "VigilNet: An integrated sensor network system for energy-efficient surveillance," ACM Transactions on Sensor Networks (TOSN), vol. 2, pp. 1-38, 2006.
[12] E. Shih, et al., "Sensor Selection for Energy-Efficient Ambulatory Medical Monitoring," presented at the MobiSys '09 2009.
[13] M. Bahrepour, et al.,"Fast and Accurate Residential Fire Detection Using Wireless Sensor Networks," Environmental Engineering and Management Journal, vol. 9, pp. 215-221, 2010.
[14] M. Bahrepour, et al., "Sensor Fusion-based Activity Recognition for Parkinson Patients," in Sensor Fusion - Foundation and Applications, ed: InTech, 2011, pp. 171-190.
[15] M. Duarte and Y. Hu, "Vehicle Classification in Distributed Sensor Networks," Journal of Parallel and Distributed Computing, 2004.
[16] M. Z. A. Bhuiyan, J. Wu, G. Wang, , and J. Cao, "Sensing and decision-making in cyber-physical systems: The case of structural health monitoring," IEEE Transactions on Industrial Informatics , pp. 1–11,2016, http://dx.doi.org/10.1109/TII.2016.2518642.
[17] M. Z. A. Bhuiyan, G. Wang, and A. V. Vasilakos, "Local area prediction-based mobile target tracking in wireless sensor networks," IEEE Trans-action on Computers, vol. 64, no. 2, pp. 1968–1982, 2015.
[18] A. Mahmood, K. Shi, S. Khatoon, and M. Xiao, "Data mining techniques for wireless sensor networks: A survey," IEEE Transactions on Parallel and Distributed Systems , vol. 2013, pp. 1–24, 2013.
[19] H. J. Woo, S. J. Shin, K. H. Joo, and W. S. Lee, "Finding context association rules over sensor-actuator data streams," IEEE Transaction on Computers, vol. 62, no. 7, pp. 74–77, 2014.
[20] A. Boukerche and S. Samarah, "A novel algorithm for mining association rules in wireless ad hoc sensor networks," IEEE Transactions on Parallel and Distributed Systems, vol. 19, no. 7, pp. 865–877, 2008.
[21] C. Nawapornanan and V. Boonjing, "An efficient algorithm for mining complete share-frequent itemsets using bittable and heuristics," in Proc.of ICMLC , 2012, pp. 96–101.
[22] M. Rashid, I. Gondal, and J. Kamruzzaman, "Mining associated pat-terns from wireless sensor networks," IEEE Transaction on Computers, vol. 64, no. 7, pp. 1998–2011, 2014.
[23] K. Romer, "Distributed mining of spatio-temporal event patterns in sensor networks," in Proc. of DCOSS, 2006, pp. 103–116.
[24] S. Samarah, B. Azzedine, and S. Alexander, "Target association rules: A new behavioral patterns for point of coverage wireless sensor networks," IEEE Transaction on Computers, vol. 60, no. 6, pp. 879–889, 2011.
[25] S. Tanbeer, C. Ahmed, and B. Jeong, "An efficient single-pass algorithm for mining association rules from wireless sensor networks," IETE Technical Review , vol. 26, no. 4, pp. 280–289, 2009.