Journal of Computer Science 8 (9): 1501-1505, 2012
ISSN 1549-3636

Wireless Sensor Networks Fault Identification Using Data Association

[1]Abirami Kongu, T., [2]P. Thangaraj and [1]P. Priakanth Kongu
[1]Engineering College, Perundurai Erode-638052 Tamilnadu, India
[2]Bannari Amman Institute of Technology, Sathyamangalam Erode-638 401, India

Abstract: Problem statement: Wireless Sensor Networks (WSN) are formed by thousands of lightweight nodes equipped with transducers for capturing information. The captured data are transmitted using multi hop routes to a base station, also called a sink. They can be extensively deployed during emergency response, medical monitoring and missiion critical applications requiring extensive data capture from sensors**.**Wireless sensor network applications include finding out patterns from what has been observed though advance knowledge of such patterns is usually unavailable. Sensor which collect data hand them over to the sink which is followed by offline data analyses to extract patterns. The existence of a large communication overhead affects sensor network performance negatively. **Approach:** This large overhead becomes a hurdle for the deployment of long term large scale sensor networks. Association mining to discover frequent patterns which form part of date mining and study of spatial and temporal properties is thus the subject of this study. As the association mining is applied in-network, Patterns and not the raw data streams are forwarded to the sink when association mining is applied to the network which thereby reduces communication overhead significantly. In this study, it is proposed to investigate the association of data received in the sink from various nodes across the network. **Results and Conclusion:** Simulations show associations based on the received traffic can be effectively used to identify mote failures and link failures.The proposed method at accuracy levels greater than 75% was able to identify all associations among the motes. The proposed method was able to find all associations among the deployed nodes.

Key words: Wireless Sensor networks, data mining, association rules, spatio-temporal patterns

INTRODUCTION

Wireless Sensor Networks (WSN) are commonly utilized for monitoring of environmental and security scenarios and they can sense, compute, store and transmit data when they integrate with each other.. In WSN, data is transmitted to a equipped node named Sink after it monitors the immediate vicinity. But what holds it back are constraints including energy, memory, computational speed and communication bandwidth. Environmental sensing, industrial and structural monitoring and others are some of the successful applications of WSN (Sabri *et al*., 2011).

The Sink which receives raw data from nodes becomes a gateway for data to be forwarded to the server from the network. With the advances of wireless sensor networks and their ability to generate a large amount of data, data mining techniques to extract useful knowledge regarding the underlying network have recently received a great deal of attention (Yuan and He, 2005). However, the stream nature of the data, limited resources and distributed nature of sensor networks bring new challenges for the mining techniques that need to be addressed. A communication overhead follows when the data after offline pattern analysis is transmitted. As the overhead becomes a serious hurdle for the deployment of long-lived and large-scale sensor network, data mining techniques are applied in-network to locate data patterns from raw data streams so that only patterns mined at the sensor nodes are forwarded to the sink. What is transmitted to the Sink includes spatio-temporal patterns from raw data defined as a set of events by the user.

Data mining technique like Association rule mining is used to locate patterns or associations encoded in the data (Agrawal *et al*., 1993). Association rule states that A→B where *A* is the antecedent and B the consequent and A, B are sets of predicates. This rule is based on the idea of concept of support/ confidence. The support is the probability of a transaction/event in the database containing both the antecedent and consequent and confidence is the probability that a

Corresponding Author: Abirami Kongu, T., Engineering College, Perundurai Erode-638052 Tamilnadu, India

record with the antecedent will also have the consequent. If $I = \{i_1, i_2,\ldots,i_n\}$ is a set of items, a transaction T is a subset of I and dataset D is set of transaction. Association rule then means finding rules in the form shown in Eq. 1:

$$R \Rightarrow i\,[S, C] \quad (1)$$

where, $R \subseteq I$ and $i \in I$, S is the support and C is the confidence. The support, supportD(X) of an item X in the dataset can be defined by Eq. 2:

$$\text{SupportD(X)} = \frac{\text{countD(X)}}{|D|} \quad (2)$$

where, *countD(X)* is the number of transactions in *D* containing *X*. The user specifies a minimum support (*min_sup*) and confidence value (*min_conf*). An itemset is said to be frequent if its support is greater than the *min_sup* value specified. Numbers of algorithms are proposed for discovering association rules from large database (Agrawal and Srikant, 1994; Han *et al*., 2000; Berzal *et al*., 2001).

The *apriori algorithm* (Agrawal and Srikant, 1994) is on the most popularly used algorithms for discovering association rules. The algorithm first discovers all frequent itemsets $I_F \subseteq I$ which has a value of support equal to or greater than *min_sup*. The algorithm merges all the frequent itemsets until no more I_F are found. On generation of the frequent itemsets, it is split in any possible way into a rule antecedent $R \subseteq I$ and a rule consequent $i \in I$ such that and $R \cup i = I_F$ and $R \cap i = \phi$. The confidence is calculated for each rule candidate and the rule is output if the confidence is above *min_conf*.

Spatial data is data connected to objects which use space. Spatial data includes topological and/or distance information, organized by spatial indexing structures. In non-spatial association rule mining, it hopes to locate a grouping among transactions encoded specifically in database, while spatial association rule mining seeks patterns in spatial relationships not encoded in database but embedded in the spatial framework of georeferenced data (Koperski and Han, 1995). These spatial relationships should be got out from data before actual association rule mining.

The user defines *maxscope* i.e., the upper bounds of distance in which the event occurs and *maxhistory* which is a time frame. The sensors collect event notifications occurring within *maxscope* and keep a history with size *maxhistory* of these events. Association mining rule is applied to the data collected to discover patterns. Every node mines patterns in the form shown in Eq. 3:

$$A_1 \wedge \ldots \wedge_m \Rightarrow E[S, C] \quad (3)$$

where, event *E* occurs at node *n* with support *S* and confidence *C* given that antecedents A_i holds true. Antecedents are in the form of:

$$A_i = (E_i, D_i, T_i, N_i)$$

Every node sends a subset of discovered patterns to the sink, thus, reducing the communication overhead. The nodes share common time length called epoch. Distances are measured either as Euclidean distances or as number of hops. Spatial association rules describe the implication of one or a set of features by another set of features in spatial databases.

Literature review: Boukerche and Samarah (2007) introduced a new formulation of the association rules, a well known data mining technique that is able to generate the time relations between sensor devices in a particular sensor network. The proposed formulation will allow traditional data mining algorithms to solve the classical association rule mining problem to be applied to sensor data. The generated rules will give a clear picture about the correlations between sensors in the network and can be used to make decisions about the network performance, or it can be used to predict the sources of future events. Experimental results have shown that the distributed extraction solution is able to reduce the number of exchanging messages and the data size by 50% compared to a direct transmission of the data.

Sun and Bai (2008) introduced a new measure w-support, which does not require pre-assigned weights. First, the HITS model and algorithm are used to derive the weights of transactions from a database with only binary attributes. Based on these weights, a new measure w-support is defined to give the significance of item sets. It differs from the traditional support in taking the quality of transactions into consideration. Then, the w-support and w-confidence of association rules are defined in analogy to the definition of support and confidence. An Apriori-like algorithm is proposed to extract association rules whose w-support and w-confidence are above some given thresholds. Experimental results show that the computational cost of the link-based model is reasonable.

Boukerche and Samarah (2008) proposed a comprehensive framework for mining Wireless Ad Hoc Sensor Networks, which can extract patterns regarding the sensors' behaviors. The main goal of determining behavioral patterns is to use them to generate rules that will improve the WASN's Quality of Service by participating in the resource management process or compensating for the undesired side effects of wireless communication. The proposed framework consists of (1) a formal definition of sensor behavioral patterns and

sensor association rules, (2) a novel representation structure that we refer to as the Positional Lexicographic Tree (PLT) that is able to compress the data gathered for the mining process and thus allows the fast and efficient mining of sensor behavioral patterns and (3) a distributed data extraction mechanism to prepare the data required for mining sensor behavioral patterns. The experiments results show that the proposed distributed data extraction mechanism resulted in reduction of messages and amount of data received by 50-90% when compared to a direct reporting mechanism. The results have demonstrated that PLT outperforms FP-Growth in terms of CPU time and memory usage by 30-50%.

Romer (2007) explored the use of in-network data mining techniques to discover frequent event patterns and their spatial and temporal properties. Raw streams of sensor readings are collected at the sink for later offline analysis-resulting in a large communication overhead. With the proposed approach, compact event patterns rather than raw data streams are sent to the sink. Various issues with the implementation of the proposed method and preliminary experiments are also discussed.

Muyeba *et al*. (2009) extended the problem of mining weighted association rules. A classical model of boolean and fuzzy quantitative association rule mining is adopted to address the issue of invalidation of Downward Closure Property (DCP) in weighted association rule mining where each item is assigned a weight according to its significance w.r.t some user defined criteria. The problem of invalidation of the DCP was solved using an improved model of weighted support and confidence framework for classical and fuzzy association rule mining. The proposed methodology follows an A-priori algorithm approach and avoids pre and post processing.

Ren and Guo (2009) presented D-FIMA, a distributed frequent items mining algorithm. DFIMA, running at every sensor node, establishes items aggregation tree via forwarding mining request beforehand and each node maintains local approximate frequent items. However, a centralized algorithm brings severely data collision in WSNs and results in inaccurate mining results. The root of the aggregation tree outputs the final global approximate frequent items. Theoretical analysis and the simulation results show that energy consumption of D-FIMA is much less than the centralized algorithm and mining results of D-FIMA is more accurate than the centralized algorithm. Experimental results show that communication loads of D-FIMA is much less than that of C-FIMA in the case of same nodes' density and the quality of mining results under DFIMA is much better than under C-FIMA and average error of items' approximate frequency under D-FIMA is less than error upper bound.

Rub *et al*. (2007) proposed a novel and more flexible relevance feedback for association rules which are based on a fuzzy notion of relevance. This approach transforms association rules into a vector-based representation using some inspiration from document vectors in information retrieval. These vectors are used as the basis for a relevance feedback approach which builds a knowledge base of rules previously rated as (un) interesting by a user. Given an association rule the vector representation is used to obtain a fuzzy score of how much this rule contradicts a rule in the knowledge base. This yields a set of relevance scores for each assessed rule which still need to be aggregated. Using rule vectors as numerical representations of association rules the study derived a similarity-based notion of relevance which we aggregated to a final relevance score using an OWA operator. The study's relevance scoring approach can be used in a wide range of application scenarios where association rules are involved. In effect, the study has created a relevance feedback engine that adapts to each user as he explores the set of association rules.

In this study, it is proposed to investigate spatio-temporal relationship among the wireless sensor nodes using association rules based on prediction. The proposed method uses INTEL dataset consisting of 54 Mica2Dot motes with temperature, humidity and light sensors.

MATERIALS AND METHODS

The experiments were conducted using Intel Lab Sensor Data 2004. The dataset consists of sensor data collected for a month from 54 sensor node with epoch duration of 31 sec. The sensors collected timestamped topology information containing temperature, light, humidity and voltage readings. The dataset consist of 2.3 million epoch readings collected from all sensors. The data is represented as shown in Fig. 1.

In this study thirty minutes of data collected between first march 2004, 9:00 AM to first March 2004, 9:30AM consisting of over 9000 data messages received in the sink was studied to find the association between the sensor motes. The distribution of the data sent by each mote to the sink is shown in Fig. 2.

A predictive apriori approach (Scheffer, 1995) is used to determine the association among the motes in sending data to the sink. In predictive apriori, balance between confidence and support is found to maximize the accuracy of the predictions on unknown data. The predictive apriori algorithm does not have a fixed confidence and support thresholds, but try to find the n best rules. The prior π over all association rules with a specific length is estimated, followed by frequent item set generation with dynamic minsup threshold and all the association rules are generated and finally the redundant rules are removed in the predictive apriori algorithm.

Date	Time	Epoch	Moteid	Temp	Humidity	Light	Voltage
Yyyy-mm-dd	hh:mm:ss.	int	int	real	real	real	real

Fig. 1: The data captured in the sink

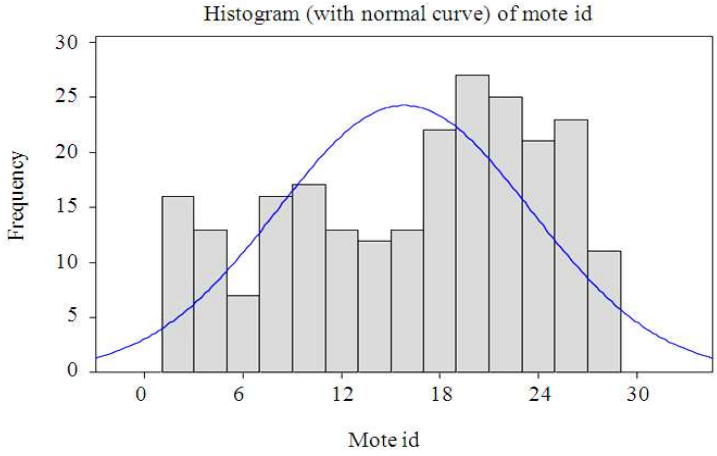

Fig. 2: The distribution of data sent by each mote during the 30 min interval

The algorithm of the predictive apriori (Scheffer, 1995) is given:

- Enter n the desired rules dataset with items a_1, a_2,…, a_k
- Draw a number of association rules [x⇒y] with i items at random. Measure their confidence at support greater than 0. Let $\pi_l(c)$ be the distribution of confidence
- For all c measure $\pi(c) = \dfrac{\sum_1^k \pi_i(c)\binom{k}{i}(2^i-1)}{\sum_1^k \binom{k}{i}(2^i-1)}$
- Let $X0 = \{\varnothing\}; LetX1 = \{\{a_1\},\ldots,\{a_k\}\}$ be all item sets with one single element
- For i = 1 ... k − 1 While (i = 1 or $X_i-1 \neq \varnothing$), determine the set of candidate item sets of length i as $X_i = \{x \cup x' | x, x' \in X_i\text{-}1 | x \cup x' | = i\}$. Generation of Xi can be optimized by considering only item sets x and x' ∈ Xi−1 that differ only in the element with highest item index. Eliminate double occurrences of item sets in Xi.
- Output best [1] . . . best[n], the list of then best association rules

RESULTS

The rules generated with proposed predictive apriori accuracy greater than 70% is shown in Table 1.

Table 1: Association identified in the proposed method

m3	m6	==>	m2	
m2	m4	==>	m3	m6
m2	m11	==>	m9	
m4	m6	==>	m2	m3
m6	m9	==>	m16	
m6	m10	==>	m7	
m8	m9	==>	m1	m11
m8	m11	==>	m1	m9
m10	m11	==>	m7	
m2	m6	==>	m3	
m3	m7	==>	m2	

Table 2: Rules Generated Using Apriori Algorithm

m3	m6	==>	m2		
m8	m9	==>	m1		
m8	m11	==>	m1		
m2	m4	==>	m3		
m4	m6	==>	m2		
m2	m4	==>	m6		
m2	m11	==>	m9		
m4	m6	==>	m3		
m4	m7=	=>	m3		
m3	m14	==>	m9		
m3	m16	==>	m9		
m6	m10	==>	m7		
m6	m16	==>	m9		
m8	m9	m11	==>	m1	
m1	m8	m11	==>	m9	
m1	m8	m9	==>	m11	
m8	m11		==>	m1	m9
m8	m9		==>	m1	m11
m3	m4	m6	==>	m2	
m2	m4	m6	==>	m3	
m2	m3	m4	==>	m6	
m4	m6		==>	m2	m3
m2	m4		==>	m3	m6
m3	m6	m7	==>	m2	

Table 2 shows the Apriori algorithm with a confidence level of 1.

Table 2 shows the rules generated using Apriori with a confidence level of 1.

Comparing Table 1 and 2 it can be seen that the proposed method is able to discover the associations to make predictive analysis such as node failure, asymmetric links. All associations with accuracy greater than 75% are identified in the Apriori algorithm also.

DISCUSSION

In this study it was proposed to find associations among motes in a wireless sensor network based on the packets received in the sink. Associations based on the received traffic can be effectively used to identify mote failures and link failures. A novel feature extraction method from the data stream was proposed and association among the motes were identified based on the type of data traffic sent was analyzed using predictive apriori algorithm.

CONCLUSION

The proposed method at accuracy levels greater than 75% were able to identify all associations among the motes. Further work needs to be done for larger duration of the streamed data as only 30 min of data were considered in this study.

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast Algorithms for Mining Association Rules in Large Databases. Comput. Inform. Sci.

Agrawal, R., T. Imieliński and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, (MD' 93), ACM Press, USA, pp: 207-216. DOI: 10.1145/170035.170072

Berzal, F., J.C. Cubero, N. Marín and J.M. Serrano, 2001. TBAR: An efficient method for association rule mining in relational databases. Data Knowledge Eng., 37: 47-64. DOI: 10.1016/S0169-023X(00)00055-0

Boukerche, A. and S. Samarah, 2007. An efficient data extraction mechanism for mining association rules from wireless sensor networks. Proceedings of the IEEE International Conference on Communications, Jun. 24-28, IEEE Xplore, Glasgow, pp: 3936-3941. DOI: 10.1109/ICC.2007.648

Boukerche, A. and S. Samarah, 2008. A novel algorithm for mining association rules in wireless ad hoc sensor networks. IEEE Trans. Parallel Distributed Syst., 19: 865-877. DOI: 10.1109/TPDS.2007.70789

Han, J., J. Pei and Y. Yin, 2000. Mining frequent patterns without candidate generation. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. (SIGMOD' 02), ACM Press, USA, DOI: 10.1145/342009.335372

Koperski, K. and J. Han, 1995. Discovery of spatial association rules in geographic information databases. Adv. Spatial Databases, 951: 47-66. DOI: 10.1007/3-540-60159-7_4

Muyeba, M., M. Sulaiman Khan and F. Coenen, 2009. Fuzzy weighted association rule mining with weighted support and confidence framework. New Frontiers Applied Data Mining, 5433: 49-61: DOI: 10.1007/978-3-642-00399-8_5

Ren, M. and L. Guo, 2009. Mining recent approximate frequent items in wireless sensor networks. Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery, Aug. 14-16, IEEE Xplore Press, Tianjin, pp: 463-467. DOI: 10.1109/FSKD.2009.607

Romer, K., 2007. Distributed mining of spatio-temporal event patterns in sensor networks. Institute Pervasive Computer.

Rub, G., M. Bottcher and R. Kruse, 2007. Relevance feedback for association rules using fuzzy score aggregation. Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society, Jun. 24-27, IEEE Xplore Press, pp: 54-59. DOI: 10.1109/NAFIPS.2007.383810

Sabri, N., S.A. Aljunid, R.B. Ahmad, M.F. Malek and A. Yahya *et al*., 2011. Performance Evaluation of Wireless Sensor Network Channel in Agricultural Application., 9: 141-151.

Scheffer, T., 1995. Finding association rules that trade support optimally against confidence. Intelligent Data Anal., 9: 381-395.

Sun, K. and F. Bai, 2008. Mining weighted association rules without preassigned weights. IEEE Transactions Kbnowledge Data Eng., 20: 489-495. DOI: 10.1109/TKDE.2007.190723

Yuan, Y., Z. Yang and J. He, 2005. An adaptive modulation scaling scheme for quality of services. Ensurance Wireless Sensor Netwoks, 2: 734-738.