Volume 5, Issue 6, June 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Distributed Datamining in Wireless Sensor Network Using Fuzzy Naive: A Survey

Nitika Malik, Pankaj Kumar

Computer Science, Uttar Pradesh Technical University

Uttar Pradesh, India

Abstract— ***Wireless sensor networks (wsns) consist of a collection of low cost and low powered sensor devices capable of communicating with each other via an ad-hoc wireless network. These sensors collect the data about the environment and this data can be mined for a variety of analysis. Unfortunately, post analysis of the data extracted from the WSN incurs high sensor communication cost for sending the raw data to the base station and at the same time runs the risk of delayed analysis. To overcome this, researchers have proposed several data mining technique which can deal with the data in situation – these data mining algorithms utilize the computing power at each node to first do some local computations and then exchange messages with its neighbours to come to a consensus regarding a global model. These algorithms reduce the communication cost vastly and also are extremely efficient in terms of model computation and event detection. In this paper we focus on such distributed data mining algorithms for data clustering, classification and outlier detection tasks.***

Keywords— Wireless Sensor network, classification, Naive Bayes classification

I. INTRODUCTION

A wireless sensor network (WSN) typically consists of a large number of small, low-cost sensor nodes distributed over a large area. A wireless sensor network (WSN) is a distributed platform that collects data over a broad area. It has a wide variety of practical military, medical, and industrial applications . The brain of a WSN is usually a decision-making algorithm that is capable of correctly mapping a set of newly collected observations from the sensors to one or more predefined categories. It uses a machine-learning algorithm to recall the classification of old data and classify the new data accordingly. There is no shortage of machine-learning algorithms available for decision making in wsns . However, the imbalanced classification is a common problem. This problem occurs when the classifier algorithm is trained with a dataset in which one class has only a few samples and there are a disproportionally large number of samples in the other classes. This kind of imbalanced data causes classifiers to be over fitted (i.e., produce redundant rules that describe duplicate or meaningless concepts) and as a result perform poorly, particularly in the identification of the minority class. In WSN applications, these rare minority classes are often critical. Some WSN examples include, but are not limited to, transaction fraud detection, machine fault monitoring, environmental anomalies, atypical medical conditions, and abnormal habitual behaviours—situations where the class of interest is a small sample of unusual readings. Studies have shown that using standard classification algorithms to analyse these imbalanced class distributions leads to poor performance. An imbalanced class problem may have another implication in WSN where it could be a symptom of producing traffic "hot-spot" in WSN. The energy consumption in the sensors may become imbalanced too, which leads to premature drain out for some local nodes. Some solution has been proposed to better cluster the nodes and traffics although it is aimed at the energy level. The sensor nodes are integrated with sensing, processing and wireless communication capabilities.

Analysis of data collected from sensor at timely manner is second important aspect . Raw data collected from the nodes often suffers from inaccuracy and incompleteness . Incomplete and inaccurate data measurement is known as WSN anomalies. Anomalies are defined as observation that do not correspond to a well defined notion of normal behaviour. Anomalies may be caused by errors , malfunctioning/failure of nodes and attacks. It is important to detect and respond to these anomalies.

II. LITERATURE REVIEW

Sensor data mining is emerging as a novel area of research and it offers wide application areas. The availability of sensors creates exciting new opportunities for data mining and data mining application.

Khushboo Sharma et. Al., [3], used Nearest Neighbors Classification technique to classify the Wireless Sensor Network data. Their experimental investigation yielded a significant output in terms of the correctly classified success rate being 92.3%.

Maria Muntean et. Al., [4], presented a wind energy monitoring with the help of classification in which sensor node monitors six attributes: speed, direction, temperature, pressure, humidity, and battery voltage. Every attribute value is set as four measures: average, instantaneous, minimum, and maximum. Authors present several data mining techniques applied on the wireless sensor network's data considered: Naïve Bayes, k-nearest neighbour, decision trees, IF-THEN rules, and neural networks. For solving wind energy monitoring problem authors have tested all these classifiers in order to conclude which is more suitable to their dataset. It is concluded that Decision Tree algorithm is more suitable to reduce the data transmission in WSN effectively and to implement classification of different types of parameters simply and practicably.

D-FLER [7] uses a distributed fuzzy engine for event detection. It combines the individual sensor readings with data from their neighbourhoods to produce a more accurate and robust classification result. A fault-tolerant event detection scheme is proposed in[8]. Since the event-data and noises are different with normal-data, a distributed Bayesian algorithm discriminates between sensor-node failures (noises)and event.

TECHNIQUES TO DETECT ANAMOLY DESIGNED FOR WSN

Anomaly detection techniques for wsns can be categorized into statistical-based, nearest neighbor-based, clustering-based, classification-based approaches.

A. Statistical - Based Anomaly Detection Techniques

These techniques build data reference model and evaluate each data pattern with respect to that reference model. Any deviation from the reference model is considered as anomaly. There are two types of statistical based techniques - parametric and non parametric techniques.In parametric techniques, known data distribution builds reference model against which parameters are evaluated. In nonparametric, as data distribution in not known a priori,some distribution estimation methods are used to build the reference model against which parameters are evaluated.

Limitations

- Dynamic nature of WSN makes it difficult to select appropriate threshold value for evaluation.
- Non-parametric statistical models are not that suitable for real time applications.
- Computational cost of handling multivariate data is more.

B. Nearest-Neighbourhoods Based Anomaly Detection Techniques

Methods based on data mining and machine learning are used for anomaly detection. These techniques use some data measurement methods to differentiate normal or anomalous data patterns.

Limitations

- Computational cost of handling multivariate data is more.
- Not scalable.

C. Clustering Based Anomaly Detection Techniques

A cluster is said to be anomalous if it is distant from other clusters in data set.Each node builds a local reference model (LRM) and sends it to its cluster head(CH). Upon receiving such lrms from its nodes, CH constructs a global reference model (GRM) from lrms. After that CH sends GRM to all its cluster members and summary of GRM to base station. Upon receiving GRM each cluster member uses same for detecting anomalies locally. Base station use GRM summary to differentiate normal or anomalous clusters.

Limitation

- Dynamic data streaming often outdate the LRM and it turn GRM, so there is need for both have to be updated continuously. Updating reference model involves lot of communication overhead and is also computationally expensive.
- Because of high computational complexity involved in measuring distance specially among multivariate data patterns, anomaly detection is expensive.

D. Classification-Based Approaches

Classification approaches are important systematic approaches in the data mining and machine learning community. They learn a classification model using the set of data instances (training) and classify an unseen instance into one of the learned (normal/anomalous) class (testing). The unsupervised classification-based techniques require no knowledge of available labeled training data and learn the classification model which fits the majority of the data instance during training. The one-class unsupervised techniques learn the boundary around the normal instances while some anomalous instance may exist and declare any new instance falling outside this boundary as an outlier. The classifier may need to update itself to accommodate the new instance that belongs to the normal class. In existing anomaly detection techniques for wsns, classification-based approaches are categorized into support vector machines (SVM)-based and Bayesian network-based approaches based on type of classification model they use.

I) Support Vector Machine-Based(SVM) Approaches:

SVM techniques separate the data belonging to different classes by fitting a hyperplane between them, which maximizes the separation. The data is mapped into a higher dimensional feature space where it can be easily separated by a hyperplane. Furthermore, a kernel function is used to approximate the dot products between the mapped vectors in the feature space to find the hyperplane.

II) Bayesian Network-Based Approach

Bayesian network-based approaches use a probabilistic graphical model to represent a set of variables and their probabilistic independencies. They aggregate information from different variables and provide an estimate on the expectancy of an event to belong to the learned class. They are categorized as naive Bayesian network, Bayesian belief network, and dynamic Bayesian network approaches based on degree of probabilistic independencies among variables. Naïve Bayesian networks techniques capture spatio-temporal correlations among sensor nodes. Bayesian belief network techniques consider the correlations among the attributes of the sensor data. Dynamic Bayesian networks techniques consider the dynamic network topology that evolves over time, adding new state variables to represent the system state at the current time instance.

Limitation

- Techniques are computationally expensive. Not suitable for online anomaly detection.
- Some techniques are not adaptive.
- Non scalable to handle multivariate data.

Concept Of Naive Bayes classification

The naive Bayes classifier, or simple Bayes classifier, works as follows :

(i) Each data sample is represented by an m+1 dimensional feature vector (a1; a2; . . . ; am; c), depicting m+1 measurements made on the sample from m+1 attributes, respectively, A1;A2; . . . ;Am;C, where C is the class attribute and c is the class label.

(ii) Suppose that the domain of C is (C1; C2; . . . ; Cl) where Ci≠Cj for i≠j, and thus there exist ƛ classes. Given an unknown data sample, X =(a1; a2; . . . ; am) (i.e., having no class label),the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naive Bayes classifier assigns an unknown sample X to class Ci if and only if

P(Ci|X) >P(Cj|X) (1)

For 1≤j≤ ƛ, j≠i.

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis.

(iii) By Bayes theorem, we have

P(Ci|X) = P(X|Ci)P(Ci)/ P(X) (2)

As P(X) is constant for all classes, only P(X│Ci)P(Ci) need to be maximized. The class prior probabilities may be estimated by

P(Ci) =Si/s (3)

Where Si is the number of training samples of class Ci and s is the total number of training samples.

(iv) In order to reduce computation in evaluating P(X│Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another , given the class label of the sample, that is, there are no dependence relationships among the attributes. Thus

P(X│Ci)= $\prod_{k=2}^{\pi} F(ak|Ci)$ (4)

The probabilities P(a1│Ci), P(a2│Ci),...., P(am│Ci) can be estimated from the training sample, namely,

P(ak│Ci) = ak/sik (5)

Where sik is the number of training samples of class Ci having the value ak for Ak.

(v) Inorder to classify an unknown sample X, P(X│Ci)P(Ci) is evaluated for each class Ci . Sample X is then assigned to the class Ci if and only if

P(X│Ci)P(Ci) > P(X│Cj)P(Cj) (6)

For 1≤j≤ ƛ , j≠i. In other word, it is assigned to the class Ci for which P(X│Ci)P(Ci) is maximum.

III. CONCLUSION

In this survey paper , we address the problems of anomaly detection in wireless sensor network . We also provided information about anomalies in wsns, desirable properties of any anomaly detection technique designed for wsns. Naïve bayes approach for classification is one of the existing technique which we discussed in this paper and found that in many applications it is the most suitable approach to detect anamoly.The short comings of existing technique for wsns lead to the requirement of new technique for the same , which take into account multivariate data and dependencies of attributes of sensor nodes , provides reliable , real-time adaptive technique for anomaly detection with unique characteristics of wsns.

REFERENCES

[1] Jiawei Han and michelinekamber 2006 Data Mining Concepts and Techniques. San Francisco, CA: Elsevier Inc,.

[2] Dr. Laxman Sahoo, Alok Ranjan Prusty and Sanjaya Kumar Sarangi,An outlier detection and rectification method in cluster based Wireless sensor network. International Journal of Computer Science and Telecommunications [Volume 3, Issue 4, April 2012]

[3] Khushboo Sharma, Manisha Rajpoot, Lokesh Kumar Sharma, “Nearest Neighbour Classification for Wireless Sensor Network Data”, International Journal of Computer Trends and Technology- volume2issue2- 2011, pp. 41-43.

[4] Maria Muntean, honoriuvălean, Adrian Tulbure, ioanileană, Manuella Kadar, “Data mining algorithms for wireless sensor network’s data”, Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies V, edited by Paul Schiopu, George Caruntu, Proc. Of SPIE Vol. 7821, 78212G © 2010 SPIE · CCC code: 0277-786X/10/$18 · doi: 10.1117/12.882215, Proc. Of SPIE Vol. 7821 78212G-1.

[5] Bhawana parbat and R.K.Dhuware,Comparative Study of Classification Technique with Labeled Data In Wireless Sensor Network, International Journal of computer applications(0975-8887)Volume69-No. 11, May 2013.

[6] Majid Bahrepour,Nirvana Meratnia and Paul J.M. Havinga, Sensor Fusion Based Event Detection in Wireless Sensor Networks, Published in Mobile Ubiquitous system: Networking and Services(mobiquitous ‘ 2009), 6th annual International.

[7] Marin-Perianu, M. And P. Havinga, *D-FLER –A Distributed Fuzzy Logic Engine for Rule-Based Wireless Sensor Networks* Lecture Notes in Computer Science. Vol. 4836. 2008, Heidelberg: Springer Berlin.

[8] Krishnamachari, B. And S. Iyengar, *Distributed Bayesian Algorithms for Faulttolerant Event Region Detection in Wireless Sensor Networks.* Computers, IEEE Transactions on, 2004. **53**(3): p. 241- 250.

[9] Satish S. Bhojannawar, Chetan M Bulla , Vishal M Danawade, Anomaly Detection Technique For Wireless Sensor Networks – A Survey, International Journal of Advanced Research in Computer And Communication Engineering, Vol.2,Issue 10,October 2013.

[10] MOHAMMAD ALWADI, GIRIJA CHETTY, "Energy Efficiency Data Mining for Wireless Sensor Networks Based on Random Forests", IJMIA: International Journal on Data Mining and Intelligent Information Technology Applications, Vol. 4, No. 1, pp. 1 ~ 8, 2014.

[11] Bill C.P. Lau, Eden W.M. Ma, Tommy W.S.Chow, Probabilistic fault detector for Wireless Sensor Network,Expert System With Applications , Volume 41,Issue 8,15 june 2014,Pages 3703-3711,ELSEVIER.

[12] Lambodar Jena,Narendra Kumar Kamila, Distributed Data Mining Privacy By Decompostion(DDMPD) with Naïve Bayes Classifier and Genetic Algorithm, International Journal of application Or Innovation in Engineering & Management,Volume2,Issue7,July2013.

[13] T.Kavitha,A.Chandra,Wireless Networks:A Comparison and Classification Based Outlier Detection Methods, Publications of problems and applicationin Engineering Research – Paper,Volume -04,Special Issue01,2013.