International Journal of Pure and Applied Mathematics
Volume 118 No. 24 2018
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

Machine Learning for Data Aggregation in WSN:A Survey

Vasundhara V.Ghate,
Dr.Vaidehi Vijayakumar
School of Computing Science Engineering,
VIT University,
Chennai,India

May 23, 2018

Abstract

Wireless Sensor Networks consist of several low-cost, low-energy sensor nodes that sense data about their corresponding environment and transfer it towards the destination through a sink or coordinator node. As the WSNs are power-constrained, thus efficient mechanisms should be incorporated for achieving energy conservation to increase the overall network lifetime by minimizing the network load. Clustering can be applied to divide the network into clusters with each cluster having a cluster head where data collection is performed. In many cases, the temporal or spatial data collected by the sensor nodes within the same cluster might provide redundant values and thus increase the overall size of data packets to be sent towards the Cluster-head. Thus, there is a need to aggregate the data at the Cluster-head such that redundancy can be removed and data can be compressed with less number of packets to be transmitted within the network. We provide a survey of various aggregation models incorporated with the help of Machine Learning techniques and also propose a Priority-based Data Aggregation (PbDA) scheme using machine learning in WSN.

***Key Words*:**WSN,Clustering,MachineLearning,PbDA.

1 Introduction

Wireless Sensor Network is an application-dependent network which comprises of multiple low-cost and low-power devices which are also called as MEMS (Micro-electro Mechanical Systems) deployed in an ad-hoc manner for sensing and transmitting application-specific data. The devices can have different types of sensors like thermal, temperature, moisture, pressure etc. based on the type of application. With the advent of IoT , Wireless Sensor networks are being considered as one of the building blocks of IoT as they contain these devices with embedded sensors which can not only sense the data but also act in an intelligent manner with the help of Embedded Computing, Machine Learning and AI Techniques. The basic architecture of Wireless Sensor Network is shown in Figure 1.

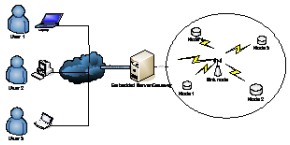


Fig 1.Wireless Sensor Network Architecture

As the devices are low-powered and in most cases with less storage capacity they make the WSNs energy-constrained. Thus, there is a need to incorporate energy-efficient routing mechanisms within the network by making use of In-network aggregation and efficient routing mechanisms. Thus, various routing mechanisms were proposed like Address-centric routing, data-centric routing, geographical location-based routing and hierarchical routing [1]. Among these strategies, the Hierarchical routing strategy is an efficient one which includes formation of clusters such that the network load towards the sink would be reduced and also energy can be conserved by minimizing the communication range of the transmission. Various clustering techniques have been in use based on the types of routing mechanism, but protocols like LEACH with its different variants has proven to be beneficial and successful in achieving energy-efficiency and load balancing even among the clusters[2]. The basic clustering technique is shown in Figure 2.

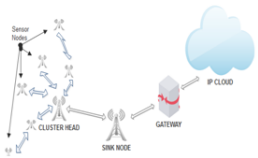


Fig 2.Clustering in Wireless Sensor Networks

For in-network aggregation simple aggregation measures can be utilized like sum, min, max, average etc. on raw data either at the nodes itself or at the cluster-head to transmit aggregated data towards the sink node and thus reduce the data packets and network load Data aggregation at cluster-heads is shown in Figure 3.

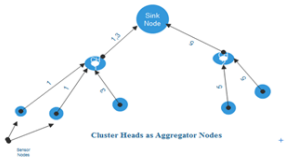


Fig 3.Data aggregation at Cluster-head

This technique works well for applications where real-time and accurate data delivery is not essential. But for applications such as emergency healthcare, disaster management the degree of aggregation to be done might differ. Also the devices should have the ability to learn from the raw data about the instant change or variation in data values below a certain threshold for data like health parameters. Thus, we can make use of Machine Learning Techniques for increasing the efficiency of predicting accurate values, classifying the data or estimate the time series based data such that alerts can be sent towards the sink node for any emergency condition based on the learning performed on the data either at the sensor node end or cluster-head end.

The remaining part of the paper is organized as given below:

Section II provides the need of Machine Learning in Wireless Sensor Networks. It also provides the overview of basic machine learning algorithms which have been implemented for improving the overall performance of WSN.

Section III provides the survey of the related research work in the functional domain of Data Aggregation in WSN using Ma-

chine Learning Techniques with underlying advantages and disadvantages.

In Section IV we introduce our novel scheme and provide brief overview of the need of such scheme in WSN and also provide the introduction of the proposed work with the basic algorithm and flowchart representation.

Section V outlines the major functional challenges in implementing the proposed scheme. In Section VI we conclude with the benefits of the proposed schemed and present the future research scope in section VII.

2 MACHINE LEARNING IN WSN

Various machine learning approaches are proven to be beneficial when implemented within a Wireless Sensor Network for energy-efficiency and real time accurate data transfer [3][4]. Machine learning is divided into following strategies:

A. Supervised Learning

Here, the input data is predefined and comes with known output data samples. It represents the relation learned between the input, system parameters and output. Supervised learning algorithms have been used in solving many problems in WSN like Node localization [5] and high dimensionality reduction [6] using neural networks, investigating spatial and temporal correlation of data collected using Support Vector Machines etc.[7]

B. Unsupervised Learning

Here there are no output class labels. The aim of this type of learning is to investigate similarity between the input samples to classify them in different classes as per the similarity. Unsupervised machine learning techniques have proven to be beneficial for data aggregation by data compression and dimensionality reduction in WSNs. Examples of algorithms include PCA(Principal Component Analysis) for dimensionality reduction of multivariate data with high dimensions by identifying important information from given set of input samples and represent it using minimum set of principal components[8].Other algorithms include k means clustering that form clusters of given samples which is mostly used for solving clustering problems in WSN.

C. Reinforcement Learning

Here, the nodes interact with their environment for learning purpose. Based on its own experience the node takes the best suitable actions which can maximize its life-term rewards. The most commonly used reinforcement learning technique in WSN routing is Q learning [9].

3 RELATED WORK

SomasekharKandukuri et al. in [10] have proposed an algorithm considering two terms: relative variation (RV) and DAWF (Data Aggregative Window Function) which are used to perform data compression by exploiting spatio- temporal redundancies in data through inspecting correlation. It has been proposed to perform all the pre-filtration required at the sensor nodes itself such that the load on the cluster-head would be reduced by sending only uncorrelated values from the sensor nodes to the cluster head. However, how the data can further be aggregated has not been addressed. Also, for emergency applications, there might be need of minimizing the degree of aggregation for important readings. Data loss wont be acceptable in such applications.

In [11], Lee et al. proposed novel network architecture named Cluster-based self-Organizing Data Aggregation. In this architecture, the nodes have ability to classify the aggregated data using a self-organizing algorithm. This technique helps mapping high dimensional space to a low dimensional one. This technique enhances the quality of data with reduced network load and improved network lifetime.

In [12], S. Lin et al. proposed a methodology which does not take into consideration the knowledge about the underlying network topology called Adaptive learner Vector quantization for performing online data compression. Here, the methodology helps in accurately retrieving compressed versions of sensor reading with the help of analysing the correlation between the readings and historical data. The limitation of this methodology is that it does not take into consideration the faulty or dead neurons whose values are far away from the values of training samples. This affects the robustness of this methodology.

In [13], Chang Liu, Juan Luo et al. proposed a Correlation-based Data Aggregation Model where a correlation of every pair of adjacent nodes and draws a correlation graph of network for constructing the aggregation tree by following the shortest path algorithm. The problem with the methodology is that it has taken into consideration the correlation based on only spatial data, temporal correlation has not been taken into account which is a major requirement of multiple emergency applications of WSN.

Macua et al. introduced a methodology called Consensus-based method for data compression using PCA and also the maximum likelihood of the observed data [14]. These two methods depended on the eigenvectors of local covariance matrices. The limitation with this methodology is the amount of trade-off between the approximation quality and the communication cost in adjusting the consensus round parameter.

Recently in [15], the researchers have made use of PCA algorithm for mapping a high dimensional space to low dimensional space. The compression is performed at cluster head which ignores the principal components with least variance values.

4 PROPOSED WORK

Emergency applications like healthcare require accurate predictions on real-time basis for decision making. Thus, in case of Body Area Networks, there is a need of efficient machine learning techniques to be applied at the data collector node with optimal data aggregation such that important readings are not skipped. Thus the aim here is to aggregate the data such that important data is extracted from input data and it is sent on a priority-basis by the cluster-head towards the sink based on its severity level The severity level can be decided based on the baseline values acceptable for health parameters and threshold levels can be depending upon various factors like patients age, sex, disease, previous history of health complications by consulting the healthcare experts. We propose a priority-based data aggregation scheme which has 2 variants based on the input data values.

The 2 aggregation schemes are:

a. Priority-based data aggregation for input data with class

labels.

b. Priority-based data aggregation for input data without class labels.

a.Priority-based data aggregation for input data with class labels.

In certain cases the health complication or the disease is known previously and can be provided as a class label in output vector. It might be a case that various body sensors might be used for such patients for sensing various health parameters like Heart-rate, Breathing rate, ECG, blood pressure systolic, blood pressure diastolic etc. If the class label is already known it may be needed that based on the health complication we can extract only necessary or related health parameter values. We can extract these values using Machine Learning Techniques or filtering techniques. Based on the extracted value we can monitor the variations from threshold values decide weights, rank such attributes and set their priority levels based on rankings. The cluster-head can then transmit the data further based on the priority on immediate basis or aggregate the data more if priority is less. Thus, Offline learning can be performed for filtering and deciding features and their weights. Based on this learner model the incoming live data can be tested and remaining steps can be performed at the cluster-head.

The block-diagram for the proposed scheme is shown in Figure 4.

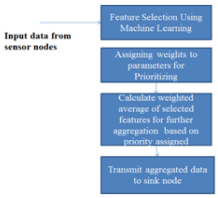


Fig 4.Block Diagram

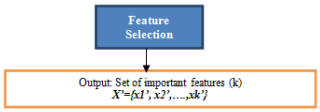
The algorithm for proposed scheme for data with class labels is given below:

• Let the Input vector be X and output vector be Y

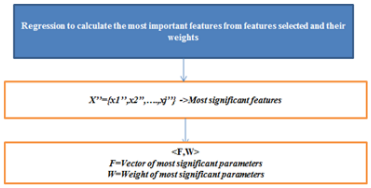
X= x1, x2, x3., xn //Input feature vector

Y= y1, y2 ,y3.,yn //Output feature vector

Step 1 : Based on ¡X,Y¿ select the important features using any Machine learning technique(SVM, Random Forest ,Wrapper methods etc.)



Step 2 : Calculate the most important features and assign weights to them using Machine Learning algorithms(Linear/Logistic Regression, Boosting techniques etc.)



Step 3: Perform classification and classify the feature vector into 2 classes most serious and less serious / set priority levels(0-5)(0-highest and 3 lowest priority)

Classification techniques like SVM can be used to decide priority and classify the data accordingly.

Step 4: Based on the seriousness or priority levels, perform data aggregation and transmit only required data instead of all data to the sink node for further transmission to doctors for taking necessary actions in case of most serious data values of patients.

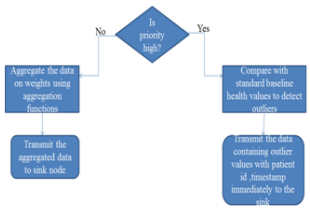


Fig 5.Algorithm Description with flowchart

The steps 1 and 2 can be performed using Offline learning for creating a learner model on cloud based on threshold values and historical data for deciding the important features and their weights. Once cluster head start receiving data it can perform steps 3 and 4.the tasks of setting priorities based on weights and transmit the

high-priority data immediately towards the sink node with data including the patient id, outlier data values with timestamp. In case of data with less priority it can be aggregated based on weights and sent in the output queue for further transmission to sink.

b. Priority-based data aggregation without class labels.

In certain cases, the output class labels might not be known or provided. In such cases, unsupervised machine learning algorithms like clustering algorithms can be implemented to form clusters of data values based on similarities. Also offline learning has to be done rigorously based on the threshold values and sample data values in similar situations. Remaining steps 2 and 3 will be performed similarly. For step 4 there might be cases of misclassification, which can be added further to the online learning performed at Step 3 so that wrong predictions in future can be avoided.

The further work focuses on surveying all algorithms to be used at each step of algorithm implementation such that the error factor is less and processing time is also less. Once the machine learning algorithms to be implemented are finalized, each step can be implemented on existing data samples to check efficiency and further the algorithm can be implemented in Body Area Network which can have a mobile device as a cluster-head and cloud as the sink where doctors can easily access the data through Internet and proceed with necessary treatment in emergency cases or in cases of symptoms found in case of less priority data accordingly.

5 FUNCTIONAL CHALLENGES

• Learning by examples (Offline learning requires learning from huge set of data samples to achieve the intended generalization capabilities (with smaller error bounds).

• For learning spatial as well as temporal correlations are essential in emergency applications so both correlations have to be taken into consideration.

• Implementing the algorithm in Wireless Body Area Networks and selecting the best medical instruments for less error-prone readings are to be used for accurate prediction. Also, setting threshold values for various health parameters requires consultation with domain experts.

6 CONCLUSION

Based on the survey of existing machine learning algorithms used in WSN, We propose a novel aggregation scheme which can be beneficial in emergency applications where the trade-off between the degree of aggregation and the prediction accuracy has to be decided. The cluster-head works as the aggregator node. Offline learning on cloud will further reduce the complexity expected at the cluster-head and overall implementation of algorithm will be faster. Due to aggregation the Energy-efficiency can be achieved as the load is balanced among the network equally. Same algorithm can also be used in traditional WSN applications like Precision Agriculture, Environment Monitoring, where certain huge variations in data values can be detectedand reported on an emergency basis to the sink node or application-side. The efficiency of algorithm depends on the machine learning algorithms we will select to use at each step of implementation.

7 FUTURE SCOPE

The proposed schemes provide lossy data aggregation schemes and require offline learning. The research focus will be to cater the need of online learning from continuous data stream for setting priority using selective sampling and also focus on providing lossless data aggregation schemes using compressive sensing and In-network aggregation where the original data can be reconstructed at the receiver side easily. Also, focus is on implementing time series analysis for real time data as input streams using Deep Learning mechanisms.

References

[1] J. Al-Karaki and A. Kamal, Routing techniques in wireless sensor networks: A survey, IEEE Wireless Communications, vol. 11, no. 6, pp. 628, 2004.

[2] Imane Horiya Brahmi , Soufiene Djahel , Damien Magoni and John Murphy, A Spatial Correlation Aware Scheme for Ef-

ficient Data Aggregation in Wireless Sensor Networks, Local Computer Networks Conference Workshops (LCN Workshops), 2015 IEEE 40th

[3] Mohammad Abu Alsheikh, Shaowei Lin, Dusit Niyato and Hwee-Pink Tan, Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications, IEEE Communications Surveys & Tutorials,Volume: 16, Issue: 4, Fourthquarter 2014

[4] T. O. Ayodele, Introduction to machine learning, New Advances in Machine Learning..InTech, 2010.K. Elissa,

[5] W. Dargie and C. Poellabauer, Localization. John Wiley & Sons, Ltd, 2010, pp. 249266.

[6] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science, vol. 313, no. 5786, pp. 504507, 2006.

[7] I. Steinwart and A. Christmann, Support vector machines. Springer, 2008.

[8] I. T. Jolliffe, Principal component analysis. Springer verlag, 2002.

[9] C. Watkins and P. Dayan, Q-learning, Machine Learning, vol. 8, no.3-4, pp. 279292, 1992.

[10] Somasekhar Kandukuri, , Richard Lorion, Jean Lebreton, Nour Murad, and Jean Daniel Lan-Sun-Luk, Energy-Efficient Data Aggregation Techniques for Exploiting Spatio-Temporal Correlations in Wireless Sensor Networks,2016 IEEE

[11] Lee S., Chung T. (2005) Data Aggregation for Wireless Sensor Networks Using Self-organizing Map. In: Kim T.G. (eds) Artificial Intelligence and Simulation. AIS 2004. Lecture Notes in Computer Science, vol 3397. Springer, Berlin, Heidelberg

[12] S. Lin, V. Kalogeraki, D. Gunopulos, and S. Lonardi, Online information compression in sensor networks,, IEEE International Conference on Communications, vol. 7. IEEE, 2006, pp. 33713376

[13] Chang Liu, Juan Luo and Yanchao Song Correlation-Model-Based Data Aggregation in Wireless Sensor Networks, 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)

[14] S. Macua, P. Belanovic, and S. Zazo, Consensus-based distributed principal component analysis in wireless sensor networks, in 11th International Workshop on Signal Processing Advances in Wireless Communications, 2010, pp. 15.

[15] C. Fenxiong, L. Mingming, W. Dianhong, and T. Bo, Data compression through principal component analysis over wireless sensor networks, Journal of Computational Information Systems, vol. 9, no. 5, pp. 18091816, 2013.