Feature Selection and Energy Management for Wireless Sensor Networks

Moh'd ALWADI, and Girija CHETTY

Faculty of Information Sciences and Engineering, The University of Canberra, Australia

Summary
Energy efficiency is a key issue in wireless sensor networks where the energy sources and battery capacity are very limited. In this paper we propose a novel pattern recognition based formulation for minimizing the energy consumption in wireless sensor networks. The proposed scheme involves an algorithm to rank and select the sensors from the most significant to the least, and followed by a naïve Bayes classification. Assuming that each feature represents a sensor in the wireless sensor network, various data sets with multiple features are considered to show that feature ranking and selection could play a key role for the energy management. We have examined Isolet, ionosphere and forest cover type datasets from the UCI repository to emulate the wireless sensor network scenario. From our simulation results, we show that it is possible to achieve two important objectives using the proposed scheme: (1) Increase the lifetime of the wireless sensor network, by using optimal number of sensors, and (2) Manage sensor failures with optimal number of sensors without compromising the accuracy.

Key words:
Wireless sensor networks, feature ranking, feature selection, data sets, accuracy, life time extension factor, WEKA machine learning framework.

1. Introduction

The field of wireless sensor networks (WSN) has be-come a focus of intensive research in recent years and various theoretical and practical questions have been addressed. WSNs can be used to monitor environmental or physical conditions such as temperature, wind and humidity [1]. Energy management in WSN is a key issue caused by a limited battery capacity and large number of sensors distributed along wide area. In sensor networks, there is no power support with constant power rate. The life time of a sensor is very restricted based on very limited power source. Therefore keeping the energy consumption in the lowest level is always a key issue. Though some approaches have been developed to address this issue, they have met with limited success, in terms of dynamically managing the energy requirements without compromising the accuracy in the event of sensor failures. In this paper we propose a novel pattern recognition based formulation of energy efficient WSNs during sensor failures without compromising the accuracy requirements. In this scheme, we model sensors with the features extracted from the data sets corresponding to different WSN application scenarios, including acoustic data (Isolet), ionosphere data and forest cover type data. In our formulation, minimizing the number of sensors for energy efficient management becomes equivalent to minimizing the number of features [2]. For minimizing, we use a feature ranking approach, where the features are ranked according to their significance of use in the wireless sensor network. That means we first rank the sensors from the most significant to the least significant, and then select optimal number of sensors to meet a specified accuracy.

For validating the proposed scheme, we used different publicly available datasets corresponding to wireless sensor networks in UCI Machine Learning repository [3]. We have studied Isolet, ionosphere and forest cover type datasets. Each data set consists of different number of sensors (features).

2. Background

Various approaches have been proposed to maximize energy efficiency and management in wireless sensor networks. Nakamura and Loureiro [4] proposed a scheme with four main contributions - an information fusion frame work for WSNs, a novel algorithm that applies information fusion to detect when a routing tree to be rebuilt, a novel routing strategy, based on role assignment which maximizes the gains of an information-fusion application and a critical survey about information fusion in WSNs. Bashyal and Venayagamoorthy [5] proposed a collaborative routing algorithm for WSN longevity, and this approach was based on four different possible node distribution in uniform or non-uniform distribution. Initial network with all surviving nodes, uneven distribution of surviving sensor nodes, a uniform distribution scheme of surviving sensor nodes and an optimal distribution for the last four surviving nodes for area coverage. Richter [6] introduced a scheme with five common steps for general pattern recognition process: signal recording, pre-processing, feature extraction, feature reduction and classification. Narasimhan and Cox [7] proposed a Handoff algorithm for wireless systems where it is necessary to switch or hand off the communication link from one base station to another for two main reasons: to maintain the signal quality and minimize interference

Manuscript received June 5, 2012
Manuscript revised June 20, 2012

caused to other radio links. Song and Allison [8] developed algorithms to break the frequency hopping spread spectrum patterns. In frequency hopping spread spec-trum the transmitter broadcasts on one frequency for small amount of time then switches to another frequen-cy using a known switching algorithm called a hopping or hopping pattern. Walchi and Braun [9] proposed an office monitoring system which is able to distinguish abnormal office access from normal access due to severe battery restrictions on the system. Therefore, office access pattern need to be classified. The node-level decision unit of self-learning anomaly detection mechanism for office monitoring with wireless sensor nodes is presented. Yu and He [10] developed the algorithm of resource reservation based on neural networks which is easy to implement and adaptable for different situations. It offers accurate classification about the user's random movement in small size cells and improved resource efficiency when resources are limited in wireless systems. Dziengel, Wittenburg and Schiller [11] presented ongoing work on distributed event detection system for WSNs. In contrast to other approaches, their system is self-contained for example it operates without a central component for co-ordination or processing, and makes active use of the redundantly placed sensor nodes in the network to improve detection accuracy. The experimental results in this paper show that distributed event detection yields higher accuracy than local detection on a single node. Wittenburg et al. [12] presented a system for distributed even detection in WSNs that allows number of sensor nodes to collaborate in order to identify which application- specific even has occurred.

3. Simulation Tools

In the current paper, MATLAB [13] and WEKA [14] have been used to develop the algorithm to rank and classify the sensors. The algorithm ranks the sensors based on the significance of use, from the most significant to the least. The following script is used on MATLAB to rank sensors in a descending order:

```
%feature selection
clear
M = csvread('covtype.csv');
[r,c] = size(M);
X = M(:, 1:c-1);
Y=M(:,c);
Sf =IndFeat(X,Y);
[SfSorted,indx] = sort(Sf,'descend');

figure
subplot(211), stem(Sf);
subplot(212), stem(SfSorted);
 grid on
 zoom on
csvwrite('selFeatures.csv',indx);
```

Fig. 1: Algorithm for feature selection [15]

4. Experiments

For the simulation work, we have studied four different data sets. We can summaries the data sets we used for this work from the UCI repository in the following table.

Table 1: Data sets.

Data set	#of instances	#of Attributes	Missing Values?	Associated tasks
ISOLET	7797	617	No	Classification
Ionosphere	351	34	No	Classification
Covertype	581012	54	No	Classification

The purpose of ISOLET dataset is to predict which letter or name was spoken. From the table above ISOLET is a large data set with 7797 instances and 617 attributes (features). It is divided into isolet 1+2+3+4 and isolet5. In this paper we used isolet5 part with only 1559 instances and 617 features because of limitations of memory size in the simulation.

Ionosphere data set that contains radar data was collected by system in Goose Bay, Labrador. The targets were free electrons in the Ionosphere. "Good" radar returns are those showing evidence of some type structure in the Ionosphere. "Bad" returns are those that do not let their signals pass through the Ionosphere [14]. In experiment 3 we used all 34 attributes in addition to the class "good" and "bad" has been replaced with "1" and "0" to be able to classify the data set, as our script and WEKA are not compatible in classifying characters.

Forest Covertype is a huge data set with very large number of 581000 attributes. This date set used to predict the forest cover type from cartographic variables [3]. In experiment 4 we used all the attributes and instances to find out the accuracy level in the classification.

The main aim of our experiments to show that to what level the number of features selected may affect the accuracy and the life time extension factor (life time of the sensor network before the sensor becomes unavailable). In the following experiments, we will show the accuracy and the life time of a sensor network based on the number of feature used in all sensor networks for ISOLET, Ionosphere and covertype datasets.

4.1. Experiment 1

The first experiment is on ISOLET dataset. The actual size of data we used consists of 1559 instances with 617 features to remove the data redundancy, whereas the original size of the dataset is 7797 instances and 617 features. After applying our Isolet5 dataset to our feature ranking algorithm, the most significant features are as shown in the following table:

Table 2: features selected on Isolet5

	1	2	3	4	5	6	7	8	9	10
1	455	453	454	456	457	458	459	460	461	462
2	69	6	101	38	37	70	39	5	262	261
3	7	102	40	71	72	103	43	104	8	44
4	76	73	42	2	41	133	74	75	230	9
5	106	11	110	108	109	78	77	263	105	45
:	107	10	12	293	3	46	264	134	229	135
:	34	111	66	290	98	226	79	47	137	140
:	227	258	294	231	139	136	225	165	332	166
:	138	265	130	112	80	486	259	142	48	232
10	233	141	295	13	81	545	266	167	481	113
:	..	..	…	…	…	…	….	…	…	…
:	….	…	…	…	…	…	…	…	…	…
20	236	467	157	177	329	485	94	147	270	239

The above features have been imported to WEKA and classification experiments performed on most significant 10 features, 20, 30, 40, 50, 100 and 200 features. We used the classify option on WEKA [14], and selected the NaiveBayes classification algorithm. We performed several tests to show the accuracy for each selection and the results as the following:

Table 3: experiment 1 accuracy

Features	Accuracy	Life time extension factor
10	9.62%	617/10= 61.7
20	11.80%	617/20 = 30.85
30	13.79%	20.56
40	14.62%	15.42
50	16.10%	12.34
100	23.92%	6.17
200	41.05%	3.08

$$\text{Life time Extension factor} = \frac{\text{Total number of features}}{\text{Number of features used}}$$

From the table above it seems to be clear that the accuracy is increased based on the number of features selected. However, this will be at the cost of the life time extension factor. Life time extension factor is increased if the number of features used is less, and redundant features are eliminated. So an appropriate feature ranking and selection algorithm can determine most influential sensors or most significant features, and allow redundant features to be eliminated. We propose a measure the "life time extension factor" as the total number of sensors in network or the features divided by the number of sensors or features used as per the ranking algorithm [2]. In our experiments the life time extension factor shown in table 3.

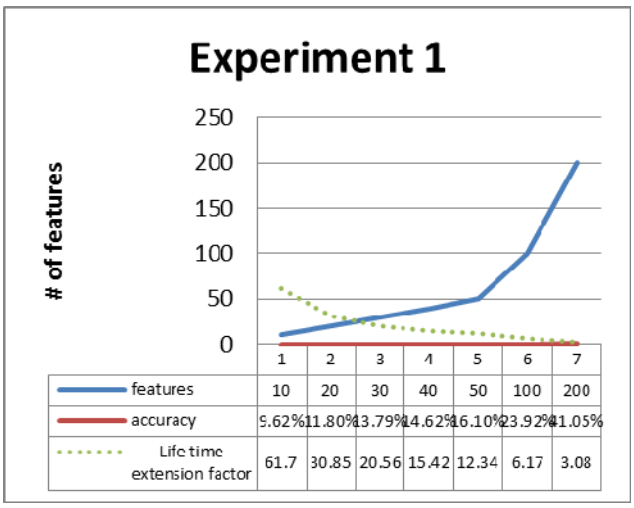

Fig. 2: Accuracy and life time extension factor. (ISOLET5)

From Figure 2 above, it can be seen that the life time extension factor increases with lesser sensors at the cost of accuracy. And the accuracy of a network could be increased at the cost of decreased life time extension factor. In the event Of a sensor failure or unavailability, it is possible to maintain the accuracy by increasing the number of features used. To emulate the sensor failure, we assign a probability. We are assuming that our sensor Si is not available with probability p= 0 , 0.01 , 0.05 , 0.10 , 0.50 [2]. In this experiment, we have multiplied our Isolet5 data set with all probability values above. We have selected 10 features and applied feature classification on WEKA and the results as the following table.

Table 4: Experiment 1 accuracy with Probability.

Features	Accuracy Without P	Accuracy P= 0.01	Accuracy P=0.05	Accuracy P=0.10	Accuracy P=0.5
10	9.62%	9.55%	9.42%	9.56%	9.56%

We can see from the table above that the system is quite stable with respect to occasional sensor faults. In case of using 20, 30, 40, 50, 100 and 200 features with probability the accuracy will still quite stable. In experiment 2 and 3 we are not going to repeat the probability multiplication in all datasets because of lack of space and memory size on WEKA.

4.2. Experiment 2

This experiment was based on Ionosphere data set. We used all 34 attributes in addition to the class “good" and "bad" have been replaced with "1" and "0" to be able to classify the data set, as our script and WEKA are not compatible in classifying characters in the data. After applying ionosphere data set into our feature ranking algorithm the most significance to the least significant features are as shown in Table 5 below:

Table 5: Experiment 2 features selected & ranked on Ionosphere dataset

	1	2	3	4	5	6	7	8	9	10
1	2	3	5	7	1	9	31	33	29	21
2	15	23	8	13	25	14	11	12	16	6
3	19	10	18	22	27	4	17	34	28	32
4	20	24	30	26						

The accuracy and lifetime extension factor achieved for this dataset is as shown in Table 6 and Figure 3 below:

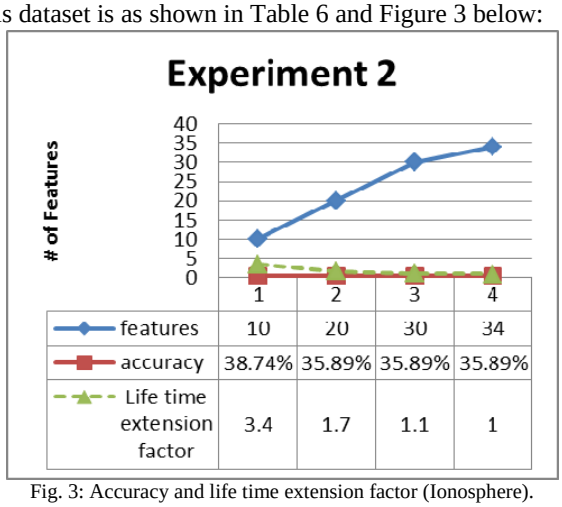

Fig. 3: Accuracy and life time extension factor (Ionosphere).

We can notice that the accuracy of Ionosphere data set is larger than the accuracy of Isolet in experiment 1. The accuracy when using more than 10 features is getting constant this may affected by the class type of 0 and 1 in the dataset. However, we can conclude that using more features is costing more resources and less life time of the sensor network.

4.3. Experiment 3

The data set used in experiment 3 is forest covertype dataset. This dataset is a large data set with size 581012 instances and 54 attributes. After applying feature selection algorithm to the forest cover type data, features are ranked and selected in the following table from the most significance to the least significance of use.

Table 6: Experiment 2 Accuracy.

Features	Accuracy	Life time extension factor
10	38.74%	34/10 = 3.4
20	35.89%	34/20 = 1.7
30	35.89%	34/30 = 1.1
34	35.89%	34/34 = 1

Table 7: Experiment 3 features ranked & selected on cover type dataset

Feat#	1	2	3	4	5	6	7	8	9	10
1	15	19	28	29	51	1	26	36	37	52
2	24	53	12	25	27	54	44	14	18	43
3	10	6	32	8	40	17	48	38	20	49
4	16	35	42	7	33	5	23	3	13	31
5	30	4	45	2	11	21	41	9	39	22
6	47	46	50	34						

The accuracy and life time extension factor for this dataset is as shown in Table 8.

Table 8: Experiment 3 Accuracy.

Features	Accuracy	Life time extension factor
10	68.00%	54/10 = 5.4
20	68.16%	54/20 = 2.7
30	68.27%	54/30 = 1.8
40	68.37%	54/40 = 1.3
54	68.49%	54/54 = 1

The results of this experiment are given in Table 8 and Figure 4. We can see that this dataset is more accurate compared to the previous datasets because covertype data set is very large contains large number of instances and large number of features.

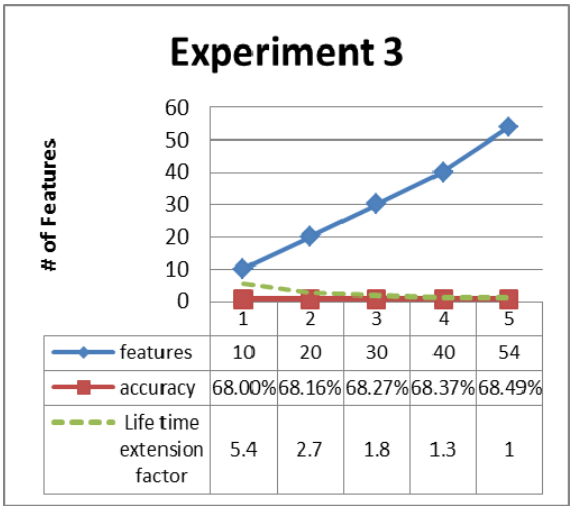

Fig.4: Accuracy and life time extension factor (Covertype)

A similar trend can be observed in terms of accuracy and life time extension factor. We can draw similar conclusions from the previous experiments that using more sensors is costing resources and reducing life time of the sensor network. A feature ranking and selection algorithm can increase the life time of the sensor network at the cost of deterioration in accuracy, and it case of sensor failures it is possible to maintain the accuracy to a specified level by employing more sensors.

5. Conclusions and Further Plan

Energy sources are very limited in wireless sensor networks. In this paper we propose a feature selection and ranking approach to manage energy in wireless sensor network. Using lesser sensors that are most significant can increase the life time of the network. Further, the proposed feature ranking and selection scheme allows graceful management of sensor network in the event of sensor failures, by increasing the number of sensors to meet the specified accuracy requirements. The proposed scheme was validated by extensive experimental evaluation for different datasets corresponding to wireless sensor networks used in different application scenarios. As a future plan, we are going to investigate more data sets to achieve higher accuracy.

References

[1] Ping, S., *Delay measurement time synchronization for wireless sensor networks.* Intel Research Berkeley Lab, 2003.

[2]. Csirik, J., P. Bertholet, and H. Bunke. *Pattern recognition in wireless sensor networks in presence of sensor failures*. 2011.

[3]. Frank, A. and A. Asuncion, *UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California.* School of Information and Computer Science, 2010. 213.

[4]. Nakamura, E.F. and A.A.F. Loureiro, *Information fusion in wireless sensor networks*, in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*2008, ACM: Vancouver, Canada. p. 1365-1372.

[5]. Bashyal, S. and G.K. Venayagamoorthy. *Collaborative routing algorithm for wireless sensor network longevity*. 2007. IEEE.

[6]. Richter, R., *Distributed Pattern Recognition in Wireless Sensor Networks.*

[7]. Narasimhan, R. and D.C. Cox. *A handoff algorithm for wireless systems using pattern recognition*. 1998. IEEE.

[8]. Song, M. and T. Allison, *Frequency Hopping Pattern Recognition Algorithms for Wireless Sensor Networks.*

[9]. Wälchli, M. and T. Braun, *Efficient signal processing and anomaly detection in wireless sensor networks.* Applications of Evolutionary Computing, 2009: p. 81-86.

[10]. Yu, W. and C. He. *Resource reservation in wireless networks based on pattern recognition.* 2001. IEEE.

[11]. Dziengel, N., G. Wittenburg, and J. Schiller. *Towards distributed event detection in wireless sensor networks*. 2008.

[12]. Wittenburg, G., et al. *A system for distributed event detection in wireless sensor networks*. 2010. ACM.

[13].MATLAB. 20/01/2012]; Available from: http://www.mathworks.com.au/.

[14]. Hall, M., et al., *The WEKA data mining software: an update*. ACM SIGKDD Explorations Newsletter, 2009. 11(1): p. 10-18.

[15]. 15/02/2012]; Available from: http://dwinnell.com/IndFeat.m.

[16]. Cortez, P. and A.J.R. Morais, *A data mining approach to predict forest fires using meteorological data.* 2007.

Moh'd Alwadi was born in UAE in 1981. He completed his Bachelor degree in computer science from Irbid National University in Jordan in 2003, He finished his master's degree in Information and Technology and Network computing from the University of Canberra, Australia in 2009 and currently studying PhD degree in University of Canberra. Moh'd research interests are in the area of computer networking and wireless sensor networks.

Dr. Girija Chetty has a Bachelors and Masters degree in Electrical Engineering and Computer Science from India and PhD in Information Sciences and Engineering from Australia. Currently she is an Assistant Professor and Head of Software Engineering in University of Canberra, Australia, and her research interests are in the area of wireless sensor networks, pattern recognition, and multimodal systems.