Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 483 – 490

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Energy Efficient Data Mining Scheme for High Dimensional Data

Mohammad Alwadi[a,*] ,Girija Chetty[b]

[a]The University of Canberra, Bruce, Canberra, ACT,2617, Australia
[b]The University of Canberra, Faculty of ESTM, Bruce, Canberra, ACT, 2617, Australia

Abstract

In this paper, we propose energy efficient big data mining scheme for forest cover type and gas drift classification. Efficient machine learning and data mining techniques provide unprecedented opportunity to monitor and characterize physical environments, such as forest cover type, using low cost wireless sensor networks. The experimental validation on two different sensor network datasets, forest cover type and gas sensor array drift dataset from publicly available UCI machine learning repository. Coupled with an appropriate feature selection, the complete scheme leads towards an energy efficient protocol for intelligent monitoring of large physical environments instrumented with wireless sensor networks.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Wireless sensor networks; physical environment monitoring; machine learning; data mining; feature selection;

1. Introduction

Wireless and wired sensor networks (WSNs/SNs) has become a focus of intensive research today, especially for monitoring and characterizing of large physical environments, and for tracking environmental or physical conditions such as temperature, pressure, wind and humidity. A wireless or a wired sensor network (WSN/SN) consists of a number of sensor nodes (few tens to thousands) storing, processing and relaying the sensed data, often to a base station for further computation [1, 2]. Sensor networks can be used in many applications, such as wildlife monitoring [3], military target tracking and surveillance [4], hazardous environment exploration [5], and natural disaster relief [6]. Given

* Corresponding author. Tel.: +61262693166.
E-mail address: u3019769@uni.canberra.edu.au

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.047

the huge amount of sensed data, automatically classifying them becomes a critical task in many of these applications.

Usually, the life time of a sensor in WSN is very poor due to limited battery life, and to keep the energy consumption to the lowest level is always a key issue. Therefore some ad hoc approaches have been developed to address this issue, based on communication theory principles, they have met with limited success, in terms of dynamically managing the energy requirements without compromising the accuracy with which a WSN can perform monitoring or classification or characterization in the event of sensor failures. Hence, there is an urgent need for intelligent and energy efficient schemes, especially for monitoring and characterization of large physical environments, monitored by WSNs and SNs with hundreds of heterogeneous sensor nodes.

In this paper, we introduce energy efficient data mining approach to characterize and classify some similar large scale physical environments based on a combination of feature selection and single mode/ensemble classifier techniques. The proposed machine learning and data mining formulation for energy efficiency, provides better implementation mechanism in terms of tradeoff between accuracy and energy efficiency, due to an optimal combination of feature selection and classifier techniques. By approaching the complexity of WSN/SN with a data mining formulation, where each sensor node is equivalent to an attribute or a feature of a data set, and all the sensor nodes together forming the WSN/SN set up equivalent to a multiple features or attributes of the data set, it is possible to use powerful feature selection, dimensionality reduction and learning classifier algorithms from data mining field, and come up with an energy efficient automatic classification system. In other words, by employing a good feature selection algorithm along with a good classification algorithm, for example, it is possible to obtain an energy efficient solution with acceptable characterization or classification accuracy (where the WSN/SN set up is emulated with a data set acquired from the physical environment). Here, minimizing the number of sensors for energy efficient control is very similar to minimizing the number of features with an optimal feature selection scheme for the data mining problem. Two different data sets, the forest cover type set [7] and gas sensor drift data set [7] have been used in our experiments to show that minimizing the number of sensors for energy efficiency is equivalent to minimizing the number of features.

2. Description of Datasets

Accurate natural resource inventory information is vital to any private, state, or federal land management agency. Forest cover type dataset provides such important information and is made available publicly through UCI machine learning repository [8]. The original Cover type data set is very large, and contains 581012 instances and 54 attributes. There are seven forest cover type classes (Class 1 to Class 7), such as spruce/fire, lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir and Krummholz. We used smaller subsets of this data, with each subset containing around 500 instances from each class (Class 1 to 7), with total number of instances 500 * 7 (3500) instances. Table 1 describes the Forest cover type data set.

Table 1. Cover type Data set description

Forest Cover Type data set	Number of Attributes	Number of Instances
Forest Cover Type (Original dataset)	54	581012
Subsets (Batches of Forest Cover type data set used for experiments	54	3500

The second data set used in this work is the Gas Sensor Array drift dataset, which consists of 13,910 measurements from 16 chemical sensors to 6 gases at different concentration level. The purpose of this dataset is to provide information about the concentration level at which the sensors were exposed for each measurement.
This data set, used in a classification task, consists of 6 classes for 6 measurements and gas type. The data set is divided into 10 batches collected over 36 months , each containing the number of measurements per class and month

(indicated in table 2), is details of the data set description [7, 9].

Table 2. Gas sensor Array drift data set description

Number of Attributes	Number of Instances	Number of Classes	Class 1	Class 2	Class 3	Class4	Class 5	Class 6
129	13910	6	Ethanol	Ethylene	Ammonia	Acetal Deyhde	Acetone	Toulene

In our experiments we have re-organized the data set to fit our classification requirements by dividing the data set with128 features (sensors) and 1 class field.

3. Classification Algorithms

For single mode classification schemes, five different classification algorithms have been examined in this work, including Naive Bayes, J48, MLP, Random Forests and Random trees. Each of these classification algorithms are used in our experimental work. In this Paper we discuss the experimental results of applying ensemble learning method into our data sets in section 5.

4. Experiments and Results

Different sets of experiments were performed to examine the relative performance of single mode and ensemble classifiers proposed here. We used k-fold stratified cross validation technique for performing experiments, with k=10, and for baseline benchmark performance measures, full training set was used. Further, we performed experiments without feature selection and with different feature selection techniques. Feature selection algorithm allows to find out an optimal number of features or sensor nodes needed to characterize or to classify the environment into one of the classes (which in turn leads to an energy efficient scheme). As can be seen in the experimental validation of the scheme in this Section, both the Forest cover type data set and Gas Sensor Array drift dataset show the consistent results. For Gas sensor Array drift data set has been re-organized, with large number of experiments have been done to each batch. After applying Naive Bayes, Random forest, J48 (Decision Trees), Random tree and Random committee classification for each batch, the average of the 10 batches have been taken. The details are shown in Figure 1 to Figure 4 for Forest Cover dataset and in Table 3 and Figure 5 for Gas sensor dataset.

5. Discussion

The comparative performance of different classifiers has been examined for both data sets in Section 5. As can be seen, the proposed data mining scheme based on random forest and random tree perform significantly better than conventional classifier approaches based on Naïve Bayes and decision tree (J48). With 10 fold cross validation, it was possible to achieve 86.45% with random forest, and 78.14% with random tree, as compared to 71.08% with Naïve Bayes, and 86.05% with decision trees (J48). However, with full training set mode, random forest results in 99.94% and random tree results in 100% accuracy. To clarify, for full training set mode, we use the entire training data for building the model with each classifier, and use the same data for testing it. However, when we use k fold cross validation (k = 10 here), we partition the data into 10 equal sized subsets. For the first fold, the first nine subsets (90% labeled data) is used for training, and last subset (10% data) is used for testing. For next fold, the training data consists of subset 2 to 10, and test set consists of subset 1. Likewise for each fold, the training data rotates to next 9 folds, so for each fold, the test data is unseen 10% data, as compared to 90% of training data. As can be expected, testing with unseen data (i.e. 10 fold cross validation), results in a marginal improvement for proposed random forest (86.45%)/random tree(78.14%) as compared to conventional Naïve Bayes(71.08%) and J48 classifiers(86.05%).

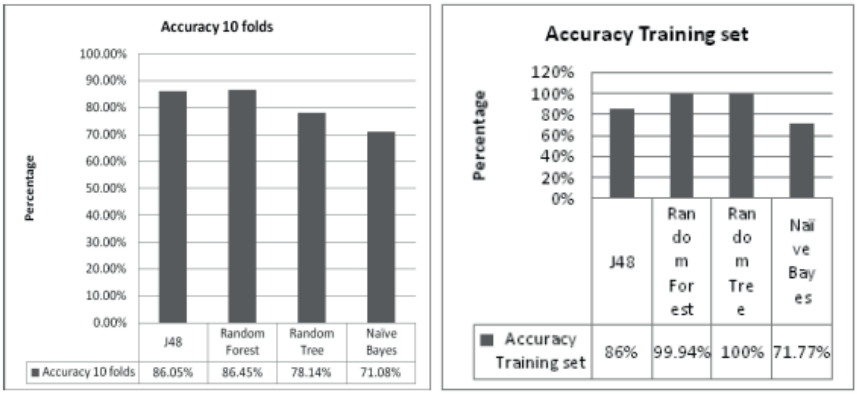


Fig. 1 and 2. (a) Performance of classifiers with 10 folds cross validation; (b) training set

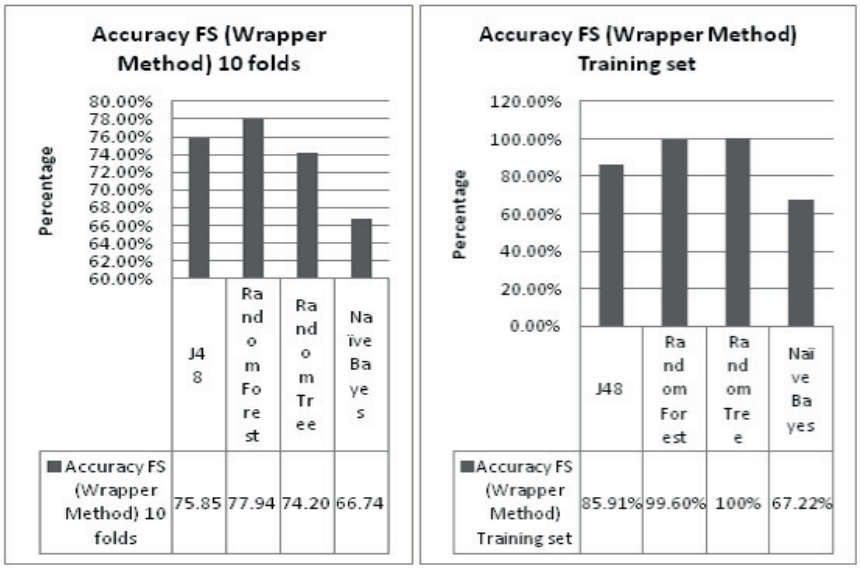
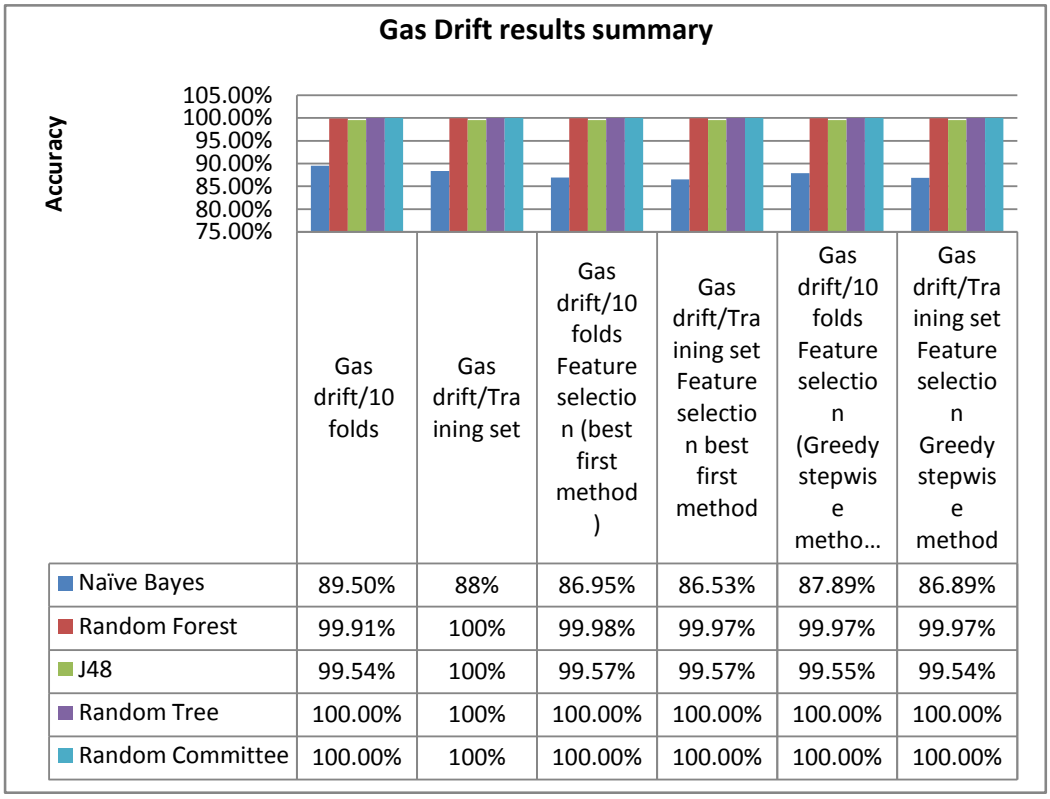


Fig. 3 and 4. (a) Performance of classifiers with feature Selection 10 folds cross validation; (b) training set

Table 3. Performance of Gas drifts sensor dataset

Forest Cover Type data set	Naive Bayes	Random Forest	J48	Random Tree	Random committee
Gas drift/10 folds	89.50%	99.91%	99.54%	100.00%	100.00%
Gas drift/Training set	88%	100%	100%	100%	100%
Gas drift/ folds Feature selection (best first method)	86.95%	99.98%	99.57%	100%	100%
Gas drift/Training set Feature selection best first method	86.53%	99.97%	99.57%	100.00%	100.00%
Gas drift/10 folds Feature selection (Greedy stepwise method)	87.89%	99.97%	99.55%	100.00%	100.00%
Gas drift/Training set Feature selection Greedy stepwise method	86.89%	99.97%	99.54%	100.00%	100.00%



	Gas drift/10 folds	Gas drift/Training set	Gas drift/10 folds Feature selection (best first method)	Gas drift/Training set Feature selection best first method	Gas drift/10 folds Feature selection (Greedy stepwise metho…	Gas drift/Training set Feature selection Greedy stepwise method
Naïve Bayes	89.50%	88%	86.95%	86.53%	87.89%	86.89%
Random Forest	99.91%	100%	99.98%	99.97%	99.97%	99.97%
J48	99.54%	100%	99.57%	99.57%	99.55%	99.54%
Random Tree	100.00%	100%	100.00%	100.00%	100.00%	100.00%
Random Committee	100.00%	100%	100.00%	100.00%	100.00%	100.00%

Fig.5. Gas drifts summary of experimental results

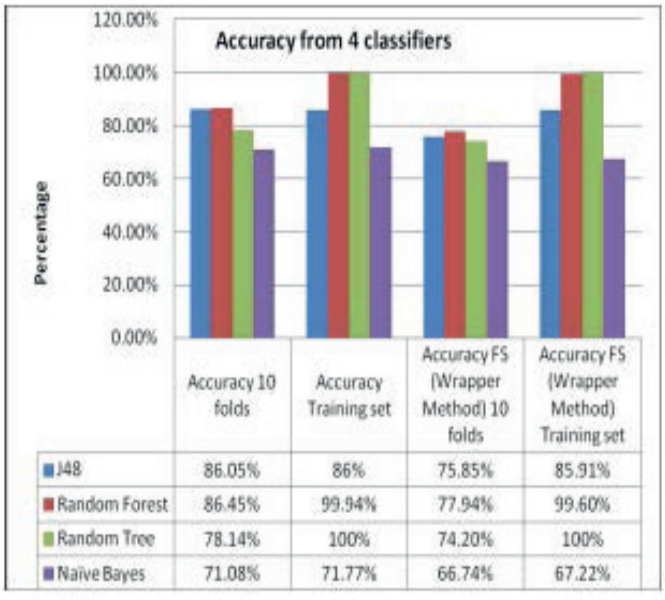
However, the improvement is significantly higher with feature selection algorithm involved, which is a wrapper type feature selection method used here. With 10 fold cross validation and feature selection, the accuracy achieved is 77.94% (random forest) and 74.20% (random tree), as compared to 66.74% (Naïve Bayes) and 75.85% (J48). With full training set (testing done on same data as training data), the improvement achieved is much higher, as is evident from figure 5. It must be noted that use of feature selection method denotes improvement in energy efficiency, as lesser number of features results in lesser computational power and storage requirements. So, a trade-off between accuracy and energy efficiency can be achieved with appropriate choice of feature selection and classification techniques. With such an intelligent monitoring protocol involving least number of sensor nodes in active mode (and insignificant sensors in sleep mode), we can potentially monitor large complex physical environments deployed with WSNs and SNs.

For the feature selection method reported here, we selected 8 features (sensors) using wrapper method for feature selection. Wrapper method searches for the best subset of features, where the feature subset assesses the quality of a set of features using a specific classification algorithm by internal cross validation. Here, the wrapper type feature selection method allows selection of most significant 8 features, instead of full feature set (54 features), resulting in reduced energy consumption in terms of sensor computation and storage requirements. As each feature represents a sensor in WSN, use of reduced 8 features here, implies 8 sensors in active mode and 46 sensors in sleep mode for classifying the forest cover type environment. This can lead to increased life for sensors, which we measure with a metric called as life time extension factor. The life time extension factor can be obtained as ratio of total number of features to number of features in active mode. In this case, the life time extension factor achieved is 54/8 = 6.75, that is around 6 times increase in life of sensors or improvement in energy efficiency. The results for the Gas Sensor

Array drift dataset were very similar to the forests cover type results. Navie Bayes classifier results from 86.89% to 89.50% . Random Forest, J48 ,Random Tree and Random committee achieved very high accuracy from 99.57% to 100%. Using feature selection the life time extension factor has been achieved up to 25 times higher efficiency 128/5 = 25.6 for the 10 folds and the same results for the full training set. With only 5 features selected instead of 128 number of features for this dataset, the energy efficiency has been improved significantly (1).

$$LifetimeExtensionFactor = \frac{TotalNumberofFeatures}{NumberofFeaturesused} \quad (1)$$

Next set of experiments involved use of ensemble learning to examine the combination of weak classifiers, which led to weak performance, such as Naïve Bayes and J48 (decision trees). As can be seen in Figure 7, the performance of weak classifiers is improved. With 10 fold cross validation mode, for Naïve Bayes classifier, due to bagging, the classification accuracy improves from 67.14% to 71%, and accuracy with J48 classifier improves from 81.94% to 88 % due to bagging. The improvement in performance due to ensemble earning is similar with full training set as is shown in Figure 7. This validates that for bagging method of ensemble learning, the performance is improved for both previously seen data (full training data – a benchmark performance), and unseen data (10 fold cross-validation). For rest of the experiments, we just used the benchmark case, i.e. full training set. The other ensemble learning techniques, including Adaboost and Stacking classifiers also result in improvement in accuracy, as can be seen in Table 4.



	Accuracy 10 folds	Accuracy Training set	Accuracy FS (Wrapper Method) 10 folds	Accuracy FS (Wrapper Method) Training set
J48	86.05%	86%	75.85%	85.91%
Random Forest	86.45%	99.94%	77.94%	99.60%
Random Tree	78.14%	100%	74.20%	100%
Naïve Bayes	71.08%	71.77%	66.74%	67.22%

Fig. 6. Comparative classifier performance

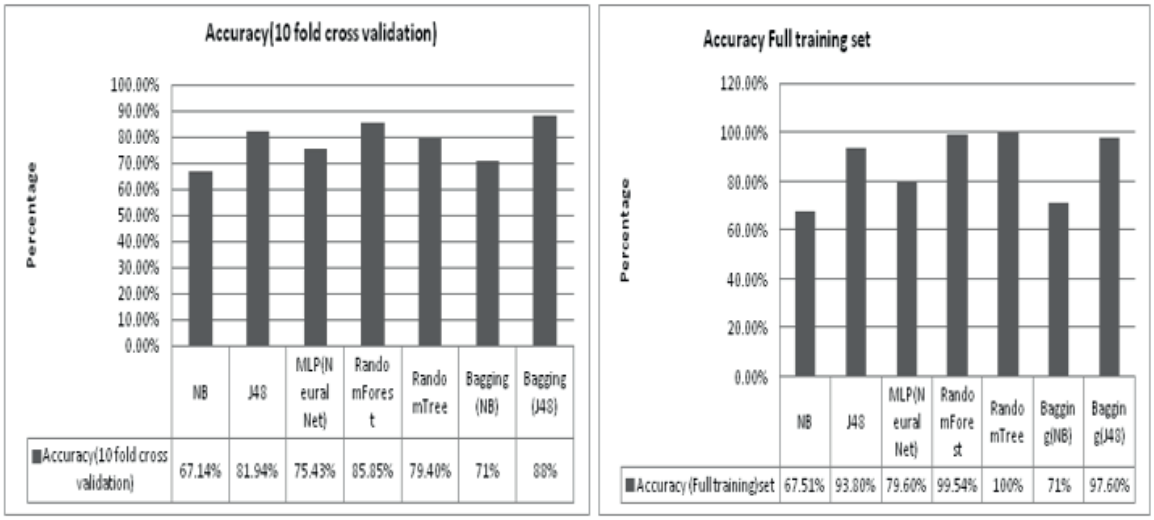


Fig. 7 and 8 : (a) Ensemble Learning 10 folds (b) ensemble learning training set

It can also be seen in Table 4, other performance measures normally used for examining the data mining schemes such as true positive rate (TPR) and false positive rate (FPR), precision and recall are satisfactory for both the single mode and ensemble mode classifiers. Table 5 here also shows the performance of different variations of ensemble learning experiments for Gas drift sensor array data.

Table 4. Single mode and Ensemble learning on forest cover type dataset

Classifier	Accuracy	Avg.TPR	Avg FPR	Precision	Recall
NB	71.77	0.72	0.05	0.74	0.72
J48	96.11	0.96	0.01	0.96	0.96
Random Forest	99.60	1.00	0.00	1.00	1.00
Random Tree	100.00	1.00	0.00	1.00	1.00
Random committe	100.00	1.00	0.00	1.00	1.00
Bagging-NB	71.50	0.71	0.05	0.74	0.71
Bagging-J48	97.85	0.98	0.00	0.99	0.93
Bagging-MLP	31.60	0.32	0.11	0.32	0.11
Adaboost-NB	71.77	0.72	0.05	0.74	0.72
Adaboost-J48	100.00	0.00	1.00	1.00	1.00
Adaboost-MLP	89.90	0.90	0.02	0.90	0.90
Stacking-NB+J48	79.80	0.80	0.03	0.86	0.80
Stacking-J48+NB	93.85	0.94	0.01	0.94	0.94
Stacking-NB+MLP	86.40	0.86	0.02	0.87	0.86

Table 5. Ensemble Learning on Gas drift sensor Array data

Ensemble Learning Methods	Use cross Validation 10 folds	Use Training set
MLP Multilayer Perceptron	97.94%	99.58%
Meta- Bagging- NB	59.25%	59.50%
Meta- Bagging-j48	98.61%	99.72%
Meta-Adaboost-NB	59.16%	59.38%
Meta-Adaboost-J48	99.38%	100%
Meta-stacking-NB	16.66%	16.66%
Meta-Stacking-J48	16.66%	16.66%

6. Conclusion

Energy sources are very limited in sensor networks, in particular wireless sensor networks. For monitoring large physical environments using SNs and WSNs, it is important that appropriate intelligent monitoring protocol is used to achieve energy efficiency and increase the lifetime of sensor nodes. In this paper, we proposed a novel data mining scheme based on random forests, ensemble learning and feature selection for characterizing two different types of publicly available sensor network data sets, the forest cover type and the gas sensor array drift dataset. These two datasets model large physical environments instrumented with SN/WSN, with each attribute/feature from the data set depicting the node/sensor of a SN/WSN set up to monitor or characterize a complex and large physical environment. Both data sets perform consistently when applying the proposed energy efficient data mining scheme. As each feature in the dataset represents a sensor, the use of an appropriate data mining scheme, with a right combination of feature selection and classification approach can facilitate an energy efficient monitoring protocol. Our further research involves extending this work with adapting these classifiers for big data stream data mining schemes , for real time dynamic protocol that can monitor complex and large physical environments in an energy efficient manner.

References

1. Ping, S., Delay measurement time synchronization for wireless sensor networks. Intel Research Berkeley Lab, 2003.
2. Hall, M., et al., The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter, 2009. 11(1): p. 10-18.
3. Csirik, J., P. Bertholet, and H. Bunke. Pattern recognition in wireless sensor networks in presence of sensor failures. 2011.
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 data2008, 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. Asuncion, A. and D. Newman, UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2007. URL:< http://www. ics. uci. edu/mlearn/MLRepository. html, 2010.
8. Hastie, T., et al., The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer, 2005. 27(2): p. 83-85.
9. Vergara, A., et al., Chemical gas sensor drift compensation using classifier ensembles. Sensors and Actuators B: Chemical, 2012. 166: p. 320-329.