Nearest Neighbour Classification for Wireless Sensor Network Data

Khushboo Sharma[#1], Manisha Rajpoot[#2], Lokesh Kumar Sharma[#3]

[#]*Department of Computer Science and Engineering, Rungta College of Engineering and Technology Kohka Road, Kurud, Bhilai, India*

***Abstract*— Advances in wireless technologies have led to the development of sensor nodes that are capable of sensing, processing, and transmitting. They collect large amounts of sensor data in a highly decentralized manner. Classification is an important task in data mining. In this paper a Nearest Neighbour Classification technique is used to classify the Wireless Sensor Network data. Our experimental investigation yields a significant output in terms of the correctly classified success rate being 92.3%.**

***Keywords*— Sensor Data Mining, Sensor Wireless Network, Classification.**

I. INTRODUCTION

Advanced computer technology leads to the emergence of computation and wireless enabled sensor devices which can be deployed to collect data from the physical world. Sensor devices currently used are computer like devices. They have a CPU, Main memory, Operating system and a suite of sensors. A typical device is the sensor nodes which has measurement, communication and computation capabilities and is powered by a small battery. There are many types of sensor which collect data from different resources. These sensors include GPS sensors, vision sensor (i.e., cameras), audio sensor (i.e., microphones), light sensors, temperature sensor, direction sensor (i.e., magnetic compass) and acceleration sensor (i.e., accelerometers). The availability of these sensors in mass-marketed communication devices creates exiting new opportunities for data mining application. A large number of sensor devices deployed in a region can communicate with each other there by forming a network [1] [2] [3] [5].

A sensor network is composed of a large number of sensor nodes, which are densely deployed either inside the phenomenon or very close to it. The position of sensor nodes need not be engineered or pre-determined. This allows random deployment in inaccessible terrains or disaster relief operations. On the other hand, this also means that sensor network protocols and algorithms must possess self-organizing capabilities. Another unique feature of sensor networks is the cooperative effort of sensor nodes. Sensor nodes are fitted with an on-board processor. Instead of sending the raw data to the nodes responsible for the fusion, sensor nodes use their processing abilities to locally carry out simple computations and transmit only the required and partially processed data. The sensor network collects the massive amount of data. To manage these data the appropriate data analysis is required. Therefore the two disciple sensor network and data mining can be combined. Knowledge from sensor data (Sensor-KDD) is important due to many application of crucial important to our society and large scale sensor system need to process heterogeneous and multisource of information from diverse type of instruments. The raw data of sensor need to be efficiently manage and transform to usable information through data fusion, which in turn must be induced tactical decision or strategic policy.

A typical application of Wireless sensor network is environment monitoring, habitat monitoring, traffic control and battle field rescue. It is also used in several real life applications, especially for monitoring several physical phenomena such as climate, Building structure and response to earthquakes. Due to their low cost, these devices are expected to become very common and every object will afford to have a sensor on it [10]. In this paper we focused on building a Nearest Neighbour (NN) classification for wireless sensor data. The main issue of a Nearest Neighbour Classifier is measuring the distance between two items. A nearest-neighbour classifier is a 'lazy learner' that does not process patterns during training [4]. When a request to classify a query vector is made the closest training vectors, according to the distance metric are located.

The remainder of the paper is organized as follows: In Section 2, the related works on sensor data is presented. In Section 3, the problem definition is described and nearest neighbour classification algorithm for sensor data is presented. In Section 4, the experimental investigation is reported and our study is concluded in Section 5.

II. RELATED WORKS

Sensor data mining is emerging as a novel area of research and it offers wide application areas. The highly distributed infrastructure provided by sensor networks supports fundamentally new ways of designing many systems. The availability of these sensors creates exciting new opportunities for data mining and data mining application. R. Srinivasan et. al. [8] investigate the possibility of using Acti-graph watches to recognize to predict ADLs (Activities of Daily Living).They apply Machine Learning Algorithms to the Acti-graph data to predict the ADLs. Also a comparative study of activity prediction accuracy obtained from four machine learning algorithm is discussed. Machine learning algorithms have been used exclusively to learn and recognize complex patterns and classify objects based on sensor data. Kwapisz et. al.[6] describe and evaluate a system that uses phone based

accelerometer to perform activity recognition. In this paper for activity recognition task they use supervised learning. They first collect data from many users as they perform activities, and then aggregated this raw time series accelerometer data while that data was being collected. Then they built predictive models for activity recognition using three classification algorithms.

Boukerche et. al. [1] proposed a framework for mining wireless sensor networks. In this framework consists of a new formulation for the sensors' associations' rules problem, distributed extraction methodology, and a compressed representation structure for sensor data. The new formulation captures the temporal relations between sensors, these relations are able to generate the set of correlated sensors which can be used later to estimate the value of another sensor, to predict the future sources of events, or to identify faulty nodes. The distributed extraction tries to maximize the network lifetime through optimizing number of exchanged messages. The measurements used to evaluate the performance of the distributed extraction were the number of messages needed to extract the data from the sensor network and the amount of the data routed to the sink. The comparison is based on the simulator that has been built. In this simulator, they abstracted the underlying communication protocols and it has been assumed that events generation is uniformly distributed over the number of slots within the given historical period along with a certain degree of correlation between sensors that ranges from 0 to 1%.

Chikhaoui et. al.[9] proposed Decision Tree (DT) based classification technique for sensor data. DT algorithm builds pattern classifier from a labelled training data set using a divide- and-conquer approach. To build up a DT model, it recursively select the attribute that is used to partition the training data set into subsets until each leaf node in the tree has uniform class membership. At each partition node, one appropriate attribute is selected and the optimal threshold is determined based on the entropy measurement to produce the greatest information gain, which assures the training samples can be well separated. Each intermediate node of a DT can have multiple branches. In order to simplify the analysis the binary DT was taken. The characters in the rectangle represent which feature or attribute of samples, A_i, is used to classify the data. The number near the rectangle is the optimal threshold value for linear classification using the attribute A_i. The leaf nodes are the final classification results in predefined classes. Malhotra et. al.[7] present schemes to generate effective feature vectors of low dimension, and also present a cluster based algorithm, where sensors form clusters on-demand for the sake of running a classification task based on the produced feature vectors.

III. Nearest Neighbour Classification for Sensor Data

A. Problem Description

Classification is one of the fundamental problems in machine learning theory. Suppose we are given n classes of Wireless Sensor Data, and when we are faced with a new, previously unseen Wireless Sensor Data, we have to assign it to one of the classes. The problem can be formalized as follows:

$$(S_1, c_1)... (S_m, c_m) \in (SD \times C) \quad (1)$$

where SD is a non empty set of the Sensor Data samples list $\{(t_0, x_0, y_0), (t_1, x_1, y_1)\ldots (t_N, x_N, y_N)\}$, with $t_i, x_i, y_i \in \Re$ for $i = 0,\ldots, N$ and $t_0 < t_1 <\ldots< t_N$ and in the present context $C = \{1,...,n\}$; the $c_i \in C$ are called labels and contain information about which class a particular trajectory belongs to. Classification means *generalization* to unseen trajectory data (S, c), i.e. we want to predict the $c \in C$ given some new Sensor Data $S \in SD$. Formally, this amounts to the estimation of a function f: $S \rightarrow C$ using the input-output training data generated independently and identically distributed according to an unknown probability distribution

B. Algorithm

In this study, a methodology which classifies Sensor Data is proposed. A Nearest Neighbour Trajectory Classification (NNTC) is explored for such purpose. Distance similarity is an important issue for Nearest Neighbour Classification therefore an efficient trajectory similarity technique is used [4]. A Nearest Neighbour classifier is a `lazy learner' that does not process patterns during training. When a request to classify a query vector is made the closest training vector(s), according to a distance metric are located. The classes of these training vectors are used to assign a class to the query vector. The nearest-neighbour method predicts the class of a test example. The training phase is trivial: simply store every training example, with its label. To make a prediction for a test example, first compute its distance to every training example. Then, keep the k closest training examples, where $k \geq 1$ is a fixed integer; look for the label that is most common among these examples. This label is the prediction for this test example. To predicts the $c \in C$ given some new Sensor Data $S \in SD$; NNTC starts with training the Sensor Data, and build a model. Algorithm 1 represents nearest neighbour trajectory classification.

```
Algorithm 1: Nearest Neighbour Classification
for Sensor Data
Input: Train Data, Test Data
Output: Classified Test Data
Methods:
Compute number of training Data NTRAIN
Compute number of test Data NTEST
For i = 1: NTRAIN
  For j = 1: NTEST
Sim[i,j] = Similarity(xtrain(,i), xtest(:,j))
  End; End
Find train data xtrain which is closest to
xtest
Assign the class label c (xtest) = c (xtrain)
```

IV. Experimental Investigation

A. Data Pre-processing

In this section the data pre-processing is applied for mining wireless sensor data. The fact that Wireless Sensor Networks

(WSNs) have limited resources necessitates the creation of a critical, efficient procedure to prepare the data. Most sensor networks use the publish/subscribe paradigm to deliver sensor readings [11]. In this paradigm, the user injects a particular interest into the network via a well equipped device (the Sink) that diffuses that interest to all sensors in the network. Each sensor maintains the received interest in a special table; upon detecting an event, a sensor checks its interest table for any interest that matches the detected event. If so, this event will be stamped with a current timestamp, and will be sent to the application. The application interprets the received events and delivers useful information to the user; however, it may happen that the application receives the events out of order, thus entailing the need for an ordering mechanism to be invoked before the events are delivered to the application. Ordering can be done by either initiating a particular ordering algorithm, like those introduced in or by simply waiting a predefined length of time to confirm that all the events issued before the received events have arrived. We perform experiment on syntactic data of wireless sensor network. It is generated from wireless sensor network simulators and also we use Intel Lab Sensor Data. It was downloaded from the Intel Berkeley Research Lab [14]. The data set was collected from 54 sensors. The data collected from each sensor device consists of its location, along with humidity, temperature, light and voltage values. Further the interested objects are identified to apply the SenSCAN algorithm [12] [13].

B. Result Analysis

Our experiment is a two-step process. First, we build a classifier from the pre-processed training data. Second, we classify test trajectory data using class labels after building the classifier. Fig. 1 shows the training data set with three class label. Our experimental investigation yields a significant output in terms of the correctly classified success rate being 92.3%. The summaries of accuracy are given in Table I. To measure the agreement between predicted and observed categorization of a dataset, while correcting for agreement that occurs by chance, is carried out by Kappa statistic.

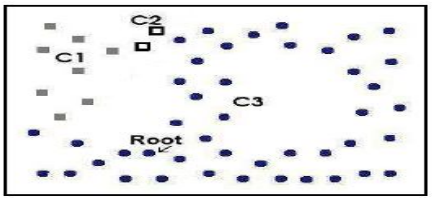

Fig. 1: Training data set of Intel Lab Sensor Data

TABLE I
FONT SIZES FOR PAPERS

Correctly Classified Data	60	92.3%
Incorrectly Classified Data	05	7.6%
Kappa statistic	81.19	
Mean absolute error	0.28	
Root mean squared error	0.18	
Relative absolute error	0.25	
Root relative squared error	0.23	
Total Number of Test Data	65	

V. CONCLUSIONS

A Nearest Neighbour classification method for sensor data has been proposed in this paper. Its primary advantage is the high classification accuracy. The classification results have demonstrated performing classification accuracy as well as classification efficiency. Overall, we have provided a paradigm in Wireless Sensor Network Data classification. Various real-world applications can benefit from our proposed framework. Though there are many challenging issues such as integration with other feature generation frameworks, and currently being investigated into detailed issues by us as a further and future study.

REFERENCES

[1] A. Boukerche A. and S. Samarah, “A Performance Evaluation of Distributed Framework for Mining Wireless Sensor Networks”, Proc. of the 40th Annual Simulation Symposium (ANSS'07), IEEE, pp. 239 - 246, 2007

[2] A. Boukerche and S. Samarah, “A Novel Algorithm for Mining Association Rules in Wireless Ad Hoc Sensor Networks”, IEEE Tran.on Parallel And Distributed Systems, Vol. 19, (7), Page 865-877, July 2008.

[3] A. Boukerche and S. Samarah “An Efficient Data Extraction Mechanism for Mining Association Rules from Wireless Sensor Networks”, Proc. of IEEE Int. Conf. on Communication (ICC’07), pp. 3936 – 3941, 2007.

[4] L. K. Sharma, O. P. Vyas, S. Schieder, and A. K. Akasapu,”A Nearest Neighbour Classification for Trajectory Data”, Springer CCIS, Volume 101, Part 1, pp. 180-185, 2010.

[5] S. K. Tanbeer, C. F. Ahmed, B. S. Jeong, Y. K. Lee,“Efficient Mining of Association Rules from Wireless Sensor Networks”, IEEE Int. Conf. Advanced Communication Technology(ICACT’09), Feb. 15-18, pp. 719 – 724, 2009.

[6] J. R. Kwapisz, G. M. Weiss, S. A. Moore, "Activity Recognition using Cell Phone Accelerometers", Proc. of the Fourth Int. Workshop on Knowledge Discovery form Sensor Data (ACM SensorKDD'10), Washington, DC, July 25-28, pp.10-18 2010.

[7] B. Malhotra, I. Nikolaidis, M. Nascimento, "Distributed and efficient classifiers for wireless audio-sensor networks", Proc. of 5th Int. Conf. on Networked Sensing Systems,IEEE, pp. 203-206, 2008

[8] R. Srinivasan, C. Chen, D. J. Cook, "Activity Recognition using Actigraph Sensor", Proc. of the Fourth Int. Workshop on Knowledge Discovery form Sensor Data (ACM SensorKDD'10), Washington, DC, July 25-28, pp. 29-34, 2010.

[9] B. Chikhaoui, S. Wang, H. Pigot, "A New Algorithm Based On Sequential Pattern Mining For Person Identification In Ubiquitous Environments", Proc. of the Fourth Int. Workshop on Knowledge Discovery form Sensor Data (ACM SensorKDD'10), Washington, DC, July 25-28 pp. 20-28, 2010.

[10] V. Jakkula and D. J. Cook, "Mining Sensor Data in Smart Environment for Temporal Activity Prediction", ACM KDD’07, August 12–15, 2007, San Jose, California, USA.

[11] A. Boukerche, S. Samarah, and H. Harbi, "Knowledge Discovery in Wireless Sensor Networks for Chronological Patterns", Proc. of 33rd IEEE Conf. on Local Computer Networks (LCN'08), pp.667 - 673, 2008.

[12] M. Ester, R. Ge, and W. Jin, "Location-Aware Clustering of Sensor Network Data", Technical Report TR 2005-13, Simon Fraser University, Burnaby, B.C., V5A 1S6, Canada, 2005.

[13] T. Wang and Z. Yang, "A Location-Aware-Based Data Clustering algorithm in Wireless Sensor Networks", 11th IEEE Singapore Int. Conf. on Communication Systems (ICCS'08), pp. 1-5, 2008.

[14] Intel Lab Data. http://db.lcs.mit.edu/labdata/labdata.html. Valid on 10 November 2011.