Support Vector Machines based on K Nearest Neighbor Algorithm for Outlier Detection in WSNs

Suya Xu *Caiping Hu* *Lisong Wang* *Guobin Zhang*

xsy048120023@163.com

College of Computer Science and Technology

Nanjing University of Aeronautics and Astronautics

Nanjing, 210016, China

Abstract —**Support vector machine approach is an effective technique to solve poly-dimensional outlier detection, which can avoid the curse of dimensionality problem and has higher accuracy. One-class support vector machine-based outlier detection techniques take advantage of spatial and temporal correlations that exist between sensor data to cooperatively identify outliers. However, for large scale training samples, SVM techniques take more spatial and temporal overhead to process and optimize training samples. In this paper, we propose KNN-SVM techniques (Support Vector Machines based on K-Nearest Neighbor Algorithm) for Outlier Detection in Wireless Sensor Networks. It utilizes KNN techniques to reduce training samples' scale which can shorten training time and optimize time. Then it maps the samples into feature space by kernel function. Experiments with data collected from the Intel Berkeley Research Laboratory show that our techniques are feasible and can effectively reduce spatial and temporal consumption with high accuracy.**

Keywords: SVM; KNN;outlier; wireless sensor network

I. INTRODUCTION

WSNs is widely used that the application scope involves personal spaces, scientific, industrial, business, and military domains. Examples of these applications include environmental and habitat monitoring, object and inventory tracking, health and medical monitoring, battlefield observation, industrial safety and controlling etc. [1]. However the inherent limitations of sensor nodes make the network more vulnerable to faults and malicious attacks [2]. A key application of WSNs is incident monitoring. For incident monitoring, normal data cannot show problem, the anomalous data that deviate so much from normal data may indicate unexpected happenings. Thus a challenge in WSNs is outlier detection. Outliers also known as anomalies are those measurements that do not conform to the normal behavioral pattern of the sensed data [3].

Support vector machines[4] (SVMs) is a new learning approach and first put forward by Vapnik et al. based on statistical learning theory. SVMs-based techniques have been widely used to detect outliers due to the following three main advantages [5]: (i) SVM-based techniques do not require an explicit statistical model, (ii) provide an optimum solution for classification by maximizing the margin of the decision boundary, and (iii) avoid the curse of dimensionality problem. Zhang et al. [1] has proposed one-class SVM-based outlier detection techniques which can update the normal behavioral model of the sensed data in an online manner.

The paper improves SVM-based outlier detection techniques by combining KNN and SVM techniques. This approach firstly applies the idea of K nearest neighbor to extract margin vectors according to training samples [6].

II. RELATED WORK

SVM-based outlier detection techniques mainly consist of three stages [7]: (i) The first stage is data processing stage. (ii) The second stage is training stage. Related parameters should be confirmed and data vectors are trained by SVMs. Then the support vectors of training samples can be found. (iii) The third stage is diagnosis

978-1-61284-683-5/12/$31.00 ©2012 IEEE

stage: put the new data vectors into SVM classification model which has been trained and then make decision if they are outliers.

III. KNN-SVM TECHNIQUES

K-nearest neighbor (KNN) techniques is a kind of distance-based approaches. They identify as outliers the objects lying in the sparsest regions or outside the model boundary of the feature space. Distance-based definitions [8, 9, 10] represent an useful tool for data analysis [11, 12, 13]. Given parameters k and R, an object is a distance-based outlier if less than k objects in the input data set lie within distance R from it [8].

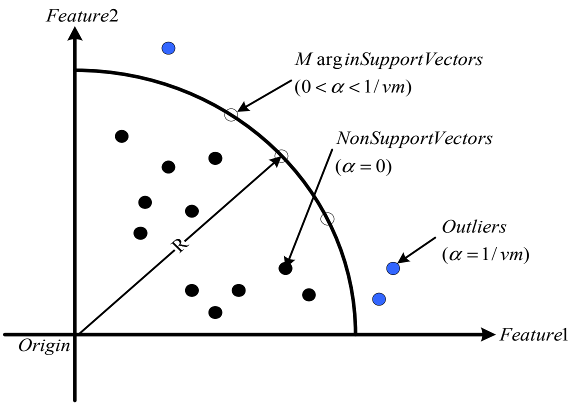

Figure1. Geometry of the quarter-sphere formulation of one-class SVM

KNN-SVM techniques combine both the idea of KNN algorithm and SVMs technique. The geometry of the one-class centered quarter-sphere SVM based approach is shown in Figure 1[6].

We assume a positive sample set, i.e., $\{x_i, i = 1,2,\ldots,l\}$, $x_i \in R^d$. The origin is presented by a and the radius is R which can contain most of the samples. We introduce the idea of kernel functions to make optimization region more tighten. First we map low dimension input space F into high dimension attribute space H via non-linear mapping function. Then we obtain the minimum hyper-spherical that contain most of the samples in high dimension feature space. Here we introduce slack variable ξ which represent the data error, and we can utilize kernel function that satisfy mercer condition to replace inner products in high dimension space, i.e., finding a kernel function $k(x,y)$ that satisfy the following condition: $k(x,y) =< \phi(x), \phi(y) >$, thus the optimization problem is as follows:

$$\min\ F(R,a,\xi_i) = R^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i \qquad (1)$$

$$s.t.\|\phi(x_i)-a\|^2 \le R^2+\xi_i, i=1,2,\ldots l \qquad \xi_i \ge 0, i=1,2,\ldots,l \qquad (2)$$

where l denotes the number of data vectors in the training set. The parameter $v \in (0,1)$ controls the number of outliers. The squared norm $\|\phi(x_i) - a\|^2$ indicates the vector length between $\phi(x_i)$ and the origin a in the feature space.

Lagrange multipliers $\alpha_i, \beta_i \ge 0, i = 1,2,\ldots,l$ was introduced to solve the problem above, we can get the following results:

$$L = R^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\beta_i\xi_i - \sum_{i=1}^{l}\alpha_i\Big(R^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \|\phi(x_i - \alpha)\|^2\Big)$$

$$s.t.\alpha_i, \beta_i \ge 0, i = 1,2,\ldots,l \qquad (3)$$

Taking the derivative of L with respect to R, ξ_i and a to zero result to:

$$\sum_{i=1}^{l}\alpha_i = 1, \alpha_i = \frac{1}{vl} - \beta_i, a = \sum_{i=1}^{l}\alpha_i\phi(x_i) \qquad (4)$$

Substituting (4) into (3) and substituting kernel function inner products in feature space produces:

$$\max \sum_{i=1}^{l}\alpha_i K(x_i,x_i) - \sum_{i=1}^{l}\sum_{i=1}^{l}\alpha_i\alpha_j K(x_i,x_j) \qquad (5)$$

$$s.t.\sum_{i=1}^{l}\alpha_i = 1, 0 \le \alpha_i \le \frac{1}{vl}, i = 1,2,\ldots,l \qquad (6)$$

Where $\{\alpha_i\}$ value can be easily obtained using some effective optimization techniques. The data vectors with $\alpha_i = 0$ are Non Support Vectors which are regarded as normal data vectors. The data vectors with $\alpha_i = 1/(vl)$ are classified as outliers. The data vectors with $0 \prec \alpha_i \prec 1/(vl)$ are Margin Support Vectors that their distances to the origin indicate the minimal radius of the hyper-sphere and can be used to classify the new unseen data vectors.

IV. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

The sensor nodes are organized into clusters that each cluster includes m sensor nodes N0, N1, N2, …, Nm-1 based on spatial correlation. Each cluster has a centered node which is recorded as N0. The network topology is modeled as an undirected graph G where G = (N, E). N represents nodes in the network and E represents an edge which connects two nodes within a cluster. The geometry is as follow in Figure 2.

We consider the time interval is Δ_t , every node measures data vectors in the time interval. We can make an example with the neighborhood in Figure 2. Let the data vectors measured by Node N_0, N_1, N_2,……, N_{m-1} are denoted by X_0 , X_1 , X_2 , …… , X_{m-1} respectively. Each data vector collected by wireless sensor node includes d attributes,i.e., $X_i^j = \{i = 0,\cdots,m-1; j = 1,\cdots,d\}$ and $X_i^j \in \Re^d$.

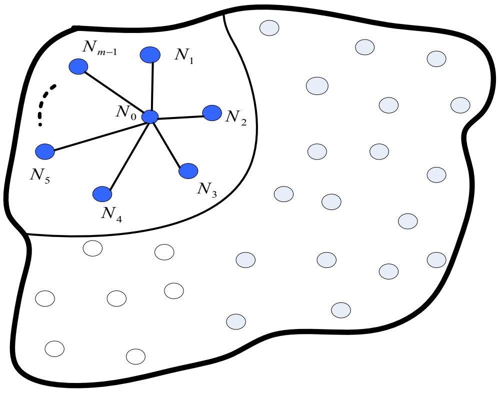

Figure2. A closed neighborhood in WSNs

In our experiment, the data is gathered from a deployment of WSN in the Intel Berkeley Research Laboratory [14]. The closed neighborhood we select is show in Figure 3 which is composed of seven nodes and we simulate our program in matlab 2007.

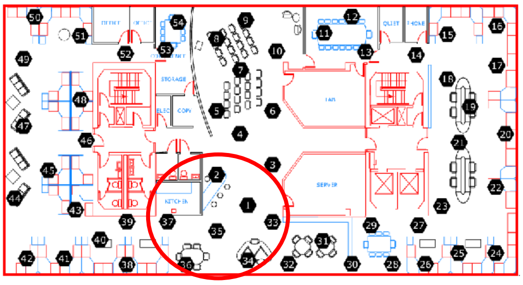

Figure3. The deployment of WSNs in Intel Berkeley Reseach Laboratory

The data vector collected from the laboratory has 8 attributes, i.e., date, time, epoch,moteid, temperature, humidity, light and voltage. This experiment chooses the latter four attributes. The closed neighborhood is marked by red curve in Figure3. Node35 is the centered node, the rest nodes are neighbors. Experiments use all the data recorded on 5th March 2004 for each data measurement. Each node has 50 anomalous data which are randomly generated by matlab.

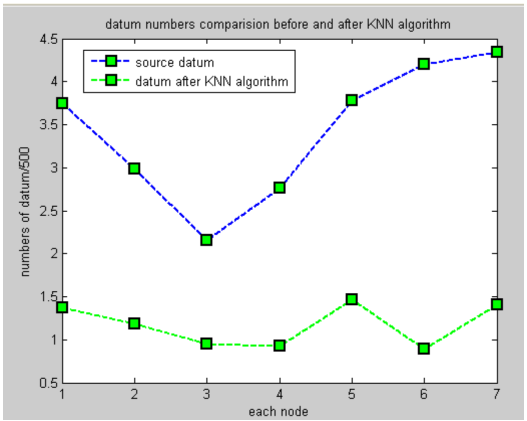

Figure4. The comparison between source datum and datum after KNN

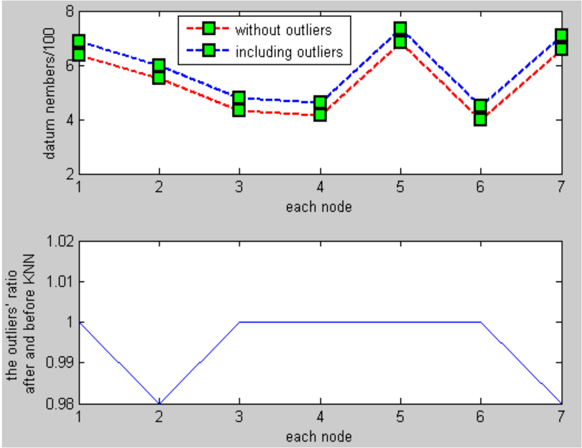

Figure 5. The ratio of outliers in sensor data vectors

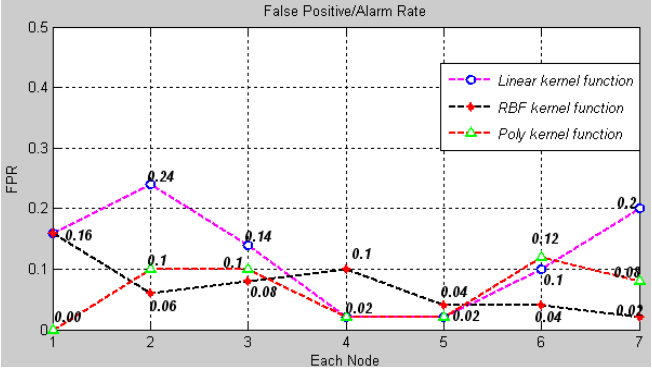

Figure 6. FPR

Here, a problem exists in KNN technique is that if KNN algorithms delete some outliers. Experiments demonstrate that nearly all the outliers are exist after KNN. Figure5 show the results of experiments. The paper calculates False Positive (Alarm) Rate of each node respectively which represents the ratio of normal data that are incorrectly classified as outliers. From figure6, the KNN-SVM has relatively high accuracy rate.

CONCLUSION

This paper proposes KNN-SVM techniques to detect outliers in the wireless sensor networks. Although reducing training time in feature space and achieve higher detection accuracy, KNN also needs substantial temporal and spatial consumption when the data samples are large scale. Excepting to find what the sample scale can make the KNN-SVM techniques achieve the best efficiency, there are many questions to be solved. Maybe compare with k-nearest neighborhood algorithm, there exist other techniques can cut the data samples better. This is also an important work to research for us.

ACKNOWLEDGEMENT

This research supported by NUAA Research Funding (No. NN2012014) and Aeronautical Science Foundation of China(No.20115552030).

REFERENCES

[1] Adaptive and Online One-Class Support Vector Machine-based Outlier Detection Techniques for Wireless Sensor Networks Yang Zhang, Nirvana Meratnia, Paul Havinga

[2] A.Perrig, J.Stankovic, and D.Wagner, “Security in wireless sensor networks,” in CACM, June 2004, vol. 47, pp. 53–57.

[3] V. Chandola, A. Banerjee, and V. Kumar. Outlier detection: A survey. Technical Report, University of Minnesota, 2007.

[4] Vapnik v. The Nature of Statistical Learning Theory. Spring-Verlag, New York, 1995.

[5] Yang Zhang, Nirvana Meratnia, Paul Havinga. An Online Outlier Detection Technique for Wireless Sensor Networks using Unsupervised Quarter-Sphere Support Vector Machine. International Conference On Intelligent Sensor, Sensor Networks and Information Processing; 2008:151-156.

[6] Shengfa Sun, Huaitie Xiao. A Fast Training Technique for Support Vector Machine based on K-nearest neighborhood. National University of Defense Technology, College of electron science and engineering， 2008.

[7] Ming Li, Anrong xue. Outliers Detection Techniques for Wireless Sensor Networks. [Master degree thesis], pages 17-21,2010.

[8] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In Proc. Int. Conf. on Very Large Databases (VLDB98), pages392-403, 1998.

[9] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In Proc. Int. Conf. on Managment of Data (SIGMOD'00), pages 427-438, 2000.

[10] F. Angiulli and C. Pizzuti. Fast outlier detection in large high-dimensional data sets. In Proc. Int. Conf. on Principles of Data Mining and Knowledge Discovery (PKDD'02), pages 15-26, 2002.

[11] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo. A geometric framework for unsupervised anomaly detection : Detecting intrusions in unlabeled data. In Applications of Data Mining in Computer Security, Kluwer, 2002.

[12] A. Lazarevic, L. ErtÄoz, V. Kumar, A. Ozgur, and J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In Proc. of the SIAM Int. Conf. on Data Mining, 2003.

[13] E. Knorr and R. Ng. Finding intensional knowledge of distance-based outliers. In Proc. Int. Conf. on Very Large Databases (VLDB99), pages 211-222, 1999.

[14] http://db.csail.mit.edu/labdata/labdata.html