Distance Based Method for Outlier Detection of Body Sensor Networks

Haibin Zhang[1], Jiajia Liu[1,*], Cheng Zhao[1]

[1]School of Cyber Engineering,hbzhang@mail.xidian.edu.cn, Xidian University, Xi'an 710071, P.R. China

Abstract

We propose a distance based method for the outlier detection of body sensor networks. Firstly, we use a Kernel Density Estimation (KDE) to calculate the probability of the distance to k nearest neighbors for diagnosed data. If the probability is less than a threshold, and the distance of this data to its left and right neighbors is greater than a pre-defined value, the diagnosed data is decided as an outlier. Further, we formalize a sliding window based method to improve the outlier detection performance. Finally, to estimate the KDE by training sensor readings with errors, we introduce a Hidden Markov Model (HMM) based method to estimate the most probable ground truth values which have the maximum probability to produce the training data. Simulation results show that the proposed method possesses a good detection accuracy with a low false alarm rate.

1. Introduction

The improvement of living standards, unreasonable diet, excess energy, environmental pollution and other factors enable chronic diseases developing more quickly. This leads to a shortage of qualified healthcare professionals and equipments to treat the sick and needy persons. The wireless body sensor network (BSN) is one of solutions to this problem. BSNs use wireless devices attached to or implanted in the body to collect various vital signs such as heart rate (HR), oxygen saturation (SpO2), blood pressure (BP), etc, and transmit collected data to a central device for processing. This allows real-time monitoring and early detection of clinical deterioration, and greater freedom and mobility while maintaining the quality of medical care [9].

Wireless devices are restricted by resources. In addition, they are frequently susceptible to environmental effects, vulnerable to the malicious, which lead to unreliability sensor data. However, medical applications have strict requirements for reliability to avoid false alarm, so outlier detection is extremely important to ensure the reliability and accuracy of sensor data before the decision-making process [9].

Outlier detections in wireless sensor networks have been studied for many years [1–4]. They estimated sensor readings or probabilities of sensor readings using spatial correlation in measurements at different sensors for the outlier detection of wireless sensor networks. These methods have an assumption that there are a large number of sensors used for the same events detection, a sensor value can be deduced from other sensors' values. However, they are not suitable for body sensor networks, because it is difficult to put too many sensors on body and it is usual that different sensors are used to collect different vital signs.

Kim *et al*. in [5] gave an approach for motion outlier detection in body sensor networks, they used history data to train Gaussian Mixture Model to generate clusters of data in similar motion groups, these cluster of data are used to estimate a Gaussian distribution to compute fault probability for new node reading. These methods can detect faults that are deviated largely to the normal data. For those faults whose readings are normal values without a low probability these methods may have a bad performance.

Chen *et al*. in [6] diagnosed abnormal data of time series of multivariate variables by two steps. Firstly, they suppose that there is a relationship between the variables represented by an expression. Then a data is diagnosed as faulty if its deviation to the estimated value by this expression is larger than a threshold.

*Corresponding author. Email: liujiajia@xidian.edu.cn

Similarly, Salem *et al.* in [8] tried to use a linear regression to estimate the reading of a sensor, e.g. HR, by values of its neighbour sensors. However, there may not be relationship between variables, and it can not decide which data is faulty when some fault occurs.

Salem *et al.* in [9] used a Kernel function to estimate the the distribution of the distance between a sensor reading to the mean of training data. A data instance is diagnosed as faulty if the probability of the calculated distance is very low. Rajasegarar *et al.* in [13],[12] used a naive Bayes based method for outlier detection. This method simply calculated the frequency of each attribute to estimate the probability. However, the Bayesian model is established on the basis of independence hypothesis, which assumed that the attributes of data are independent. This is not always true, and may affect the results of the classification.

In this paper, we propose a distance based outlier detection method for BSNs. Firstly, we calculate the average distance to k nearest neighbors of training data to estimate a KDE. For any diagnosed data, we calculate the probability of its distance to k nearest neighbors using the KDE, if this probability is less than a threshold, then the diagnosed data may be an outlier. Then we check the distances of the diagnosed data to its left and right neighbors, if both of these distances are great, then we decide the diagnosed data as an outlier.

In some conditions like that the sensor readings have successive outliers, then the previous method may have a poor performance. We introduced a sliding window to this issue. Similarly, we calculate the probability of the distance to k nearest neighbors for a diagnosed sliding window with the estimated KDE, if this probability is less than a threshold, then we decide that there are some outliers in this window. Then we check the distance to the left and right neighbors to locate the outlier.

Estimating the KDE with training data is the key issue of the proposed method for outlier detection. However, the history training data containing errors can disrupt the estimated values of the KDE. We use the Hidden Markov Model (HMM) to estimate the most probable ground truth values which have the maximum probability to produce the training sensor readings.

The rest of this paper is organized as follows. Section II introduces system models and some definitions. In section III, the distance based outlier detection method is formalized. In section IV, experiments are carried out to test the performance of the proposed method. Conclusions are drawn in Section V.

2. System Models and Definitions

Fig. 1 shows the network architecture of our considering medical deployment scenario. We use three sensors to monitor heart activities, blood pressure, respiration rate and saturation of oxygen in the arterial
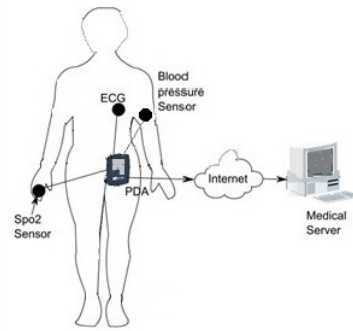

Figure 1. The Architecture of BSNs.

blood. These sensors monitor vital signs and transmit the collected data periodically at every discrete time instance to neighboring personal server devices, such as a smartphone. Then by wireless and wired connection, these data are streamed remotely to a medical doctor's site for real time diagnosis, to a medical database for record keeping, or to the corresponding equipment that issues an emergency alert.

Time series of vital sign data: It is a sequence of vital sign data arranged in time order $X = (X_a, X_{a+1}, \cdots, X_c)$, where $X_i = (x_{ih}, x_{ib}, x_{is})$ is the set of sensor readings of HR, BP and SpO2. The main purpose of time series analysis is to diagnose the current sensor readings based on the existing historical data. At any time t, suppose the ground truth values of HR, BP and SpO2 are $G_t = (g_{th}, g_{tb}, g_{ts})$, the measured values transmitted from sensors are $X_t = (x_{th}, x_{tb}, x_{ts})$, the outlier detection process decides whether x_t accords with G_t. However, for outlier detection process, the difficult is that we have no way to know the ground truth values G_t.

2.1. Kernel Density Estimation

Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable. Let $(y_1, y_2, \cdots, y_n)$ be an independent and identically distributed sample obtained from some distribution with an unknown density function f. The shape of function f can be estimated as

$$\widehat{f_h}(y) = \frac{1}{n} \sum_{i=1}^{n} K_h(y - y_i) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{y - y_i}{h}) \qquad (1)$$

where $K(\bullet)$ is a non-negative function called the kernel that integrates to one and has mean zero, $h > 0$ is a bandwidth. $K_h(x)$ is a kernel with subscript h is given as

$$K_h(x) = \frac{1}{h} K(\frac{x}{h}) \qquad (2)$$

2.2. Hidden Markov Model

A hidden Markov model is a 5-tuple $\lambda = (Q, V, \Pi, A, B)$, where $Q = \{q_1, \cdots, q_n\}$ is a set of hidden states with s_t denoting the state at time t, $V = \{v_1, \cdots, v_m\}$ is a set of observation symbols with o_t denoting the symbol at time t, $\Pi = \{\pi_1, \cdots, \pi_n\}$ is a vector of initial probabilities with $\pi_i = P(s_1 = q_i)$, A is a matrix $(a_{ij})_{(n \times n)}$ of transition probabilities with each $a_{ij} = P(s_{t+1} = q_j s_t = q_i)$, $1 \leq i, j \leq n$, B is matrix $(b_{ij})_{(n \times m)}$ of observation probabilities with each $b_{ij} = P(o_t = v_j s_t = q_i)$. We also use π_{s_1}, $a_{s_t s_{t+1}}$ and $b_{s_t o_t}$ to denote π_i, a_{ij} and b_{ij} respectively.

3. Outlier Detection Based on Distance

3.1. Simple Distance to Neighbors

We use the Euclidean distance to calculate the distance between two multivariate data. Let $X_i = (x_{ih}, x_{ib}, x_{is})$ and $X_j = (x_{jh}, x_{jb}, x_{js})$ be sensor readings on time i and j, the Euclidean distance between X_i and X_j is

$$d_{ij} = \sqrt{(x_{ih} - x_{jh})^2 + (x_{ib} - x_{jb})^2 + (x_{is} - x_{js})^2} \quad (3)$$

The k nearest distance of X_i is calculate as follows.

$$d_i^k = \sum_{i - \lfloor k/2 \rfloor \leq j < i + \lceil k/2 \rceil} w_j d_{ij} \quad (4)$$

where w_j is a weight. The distance of X_i to its left and right neighbors (*the nearest distance* for short) is

$$d_i^2 = d_{i(i-1)} + d_{i(i+1)}$$

Given a time series of history training sensor readings, we firstly calculate the k nearest distance of each data X_i, then the univariate KDE is used to estimate the probability distribution of these k nearest neighbors distance. For a recently produced sensor readings X_t, the k nearest neighbors distance d_t^k of X_t and the probability p of d_t^k is calculated by the KDE obtained by history training data. If p is less than a threshold, the the nearest distance d_t^2 of X_t is calculated, if d_t^2 is greater than a pre-defined value, then X_t is diagnosed as an outlier.

If some error occurs at time t, then the k nearest distance data at time closing to t, e.g. time $t - 1$ or $t + 1$, may have a small probability. The similar condition exists in the nearest distance. This leads to that the average rate of outliers newly introduced will be high using the k nearest distance or the nearest distance based method alone for outlier detection. Thus, we use the combination of these two method for outlier detection in this paper.

Fig. 4 shows a simulation result of 5000 HR data from a real medical dataset using the proposed outlier
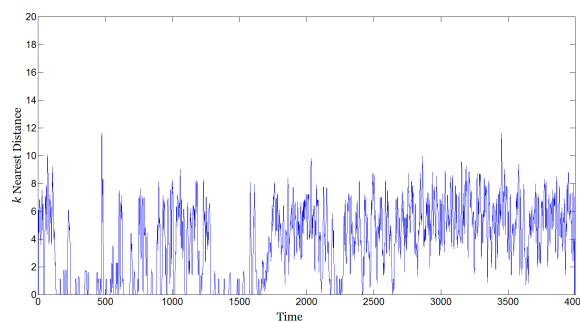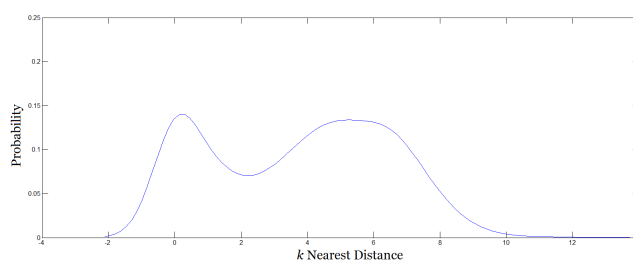

Figure 2. k Nearest Distance of Training Data.

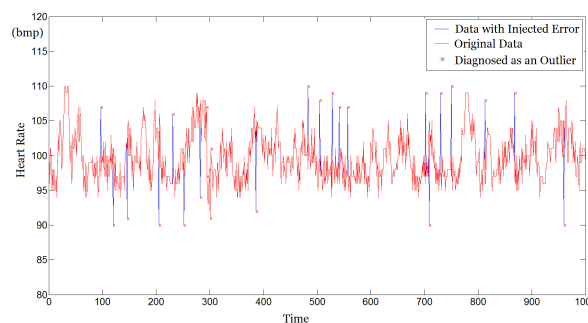**Figure 3.** Probability Density Distribution of The k Nearest Distance.

Figure 4. Outlier Detection.

detection method. We use the first 4000 data to estimate the KDE. The k nearest distance of the 4000 training data is shown in Fig. 2, where k is selected as 2, and the kernel density estimation of the distance in Fig. 2 is depicted in Fig. 3. Then we inject 5% errors to the rest of 1000 data. The original data, the data with injected errors, and the data diagnosed as outlier are all marked in Fig. 4.

3.2. Distance to Neighbors of Sliding Window

For some condition like that the time series of sensor readings has too much successive outliers, then the detection method in the above subsection may have a poor performance. The reason is that the k nearest

neighbors of an outlier X_t may be contained in the successive outliers. This leads to the k nearest distance of X_t having a normal probability, and X_t is diagnosed as a normal data. To this problem, we improve the distance based outlier detection method on a sliding window.

A sliding window with width m at time t contains X_t and its left $m-1$ nearest neighbors $X_{t-m+1}, X_{t-m+2}, \cdots, X_t$, that is

$$B_t = \begin{pmatrix} x_{(t-m+1)h}, & x_{(t-m+1)b}, & x_{(t-m+1)s} \\ x_{(t-m+2)h}, & x_{(t-m+2)b}, & x_{(t-m+2)s} \\ & \cdots & \\ x_{th}, & x_{tb}, & x_{ts} \end{pmatrix}$$

We define a distance D_{ij} between sliding window B_i and B_j as

$$D_{ij} = \frac{1}{m} \sum_{l=0}^{m-1} d_{(i-l)(j-l)} \qquad (5)$$

where $d_{(i-l)(j-l)}$ is the Euclidean distance between X_{i-l} and X_{j-l}, and the weighted average distance D_t^k of sliding window B_t to its k nearest neighbors:

$$D_t^k = \sum_{t-\lfloor k/2 \rfloor \leq l < i + \lceil k/2 \rceil} w_l D_{tl} \qquad (6)$$

where w_j is a weight. We can see that the weighted average distance D_t^k of sliding window B_t to its k nearest neighbors is the average of the k nearest distance for all X_i in B_t.

Similarly, given a time series of sensor readings of history training data, we firstly calculate the k nearest distance of each sliding window to estimate a univariate KDE. For a recently produced sensor reading widow B_t, the k nearest distance D_t^k of B_t is calculated. If the probability of D_t^k calculated by the KDE is less than a pre-defined threshold, then we decide that there are some outliers in the window B_t. The next thing is to locate the outlier sensor reading. Since the sliding window B_t contains three time series–B_{th} for HR, B_{tb} for BP and B_{ts} for SpO2, for each sensor reading y_i in B_{th}, B_{tb} or B_{ts}, if the nearest neighbor distance of y_i is greater than a pre-defined M, then y_i is diagnosed as an outlier.

3.3. Handling Error Training Data With HMMs

It is impossible to ensure that the history training data are all correct. The outliers in the training data can disrupt the estimated value of the KDE, and influence the performance of the proposed outlier detection method. To this issue, we use the HMM to estimate the most probable ground truth values which have the maximum probability to produce the training sensor readings.

Given a time series $\{y_1, \cdots, y_n\}$ of training sensor readings and a sensor error probability p, we select the first half data $O = \{y_1, \cdots, y_{T=\lfloor n/2 \rfloor}\}$ to estimate the parameters of an HMM λ by improving the Baum-Welch algorithm [?]. For HMM λ and the rest training sensor readings $O' = \{y_{T+1}, \cdots, y_n\}$, we can use the Viterbi algorithm [?] to find the most likely ground truth vital sign values G of all possible G' that can produce O'.

Given the sensor reading sequence O, the Baum-Welch algorithm finds a local maximum

$$\zeta = (\Pi, A, B) = \max_{\zeta'} P(O|\zeta') \qquad (7)$$

for sequence O with random initial conditions. However, through our experiments, we find that the performance using the HMM with parameters being estimated by the Baum-Welch algorithm directly is poor. We improve Baum-Welch algorithm as the following steps.

Forward procedure: let

$$\alpha_t(i) = P(o_1 = y_1, \cdots, o_t = y_t, s_t = q_i | \zeta) \qquad (8)$$

then

$$\alpha_1(i) = \pi_{q_i} b_{q_i o_1} \qquad (9)$$

$$\alpha_{t+1}(i) = b_{q_i o_{t+1}} \sum_{j=1}^{n} \alpha_t(j) a_{ji} \qquad (10)$$

Backward procedure: let

$$\beta_t(i) = P(o_{t+1} = y_{t+1}, \cdots, o_T = y_T, s_t = q_i | \zeta) \qquad (11)$$

then

$$\beta_T(i) = 1 \qquad (12)$$

$$\beta_t(i) = \sum_{j=1}^{n} \beta_{t+1}(j) a_{ij} b_{q_j o_{t+1}} \qquad (13)$$

Update: we can now calculate the temporary variables:

$$\gamma_t(i) = P(s_t = q_i | O, \zeta) = \frac{\alpha_t(i)\beta_t(i)}{\sum\limits_{j=1}^{n} \alpha_t(j)\beta_t(j)} \qquad (14)$$

$$\begin{aligned} \xi_t(ij) &= P(s_t = q_i, s_{t+1} = q_j | O, \zeta) \\ &= \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) b_{q_j o_{t+1}}}{\sum\limits_{k=1}^{n} \beta_t(k)\alpha_t(k)} \end{aligned} \qquad (15)$$

ζ can now be updated:

$$\pi_i^* = \gamma_1(i) \qquad (16)$$

let $a'_{ij} = \dfrac{\sum\limits_{t=1}^{T-1} \xi_t(ij)}{\sum\limits_{t=1}^{T-1} \gamma_t(i)}$, then

$$a_{ij}^* = \begin{cases} a_{ij}^* = \varepsilon \times a'_{ij} + (1-\varepsilon)/n & if\ P(q_i) < \Theta \\ a_{ij}^* = a'_{ij} & otherwise \end{cases} \qquad (17)$$

where $0 < \varepsilon \leq 1$ is a weighting, and Θ is a preselected threshold.

let $b'_{ik} = \dfrac{\sum\limits_{t=1, O_t = v_k}^{T} \gamma_t(i)}{\sum\limits_{t=1}^{T} \gamma_t(i)}$, then

$$b^*_{ik} = \begin{cases} 1 - p & if \ q_i = v_k \\ p * \dfrac{b'_{ik}}{\sum\limits_{v_k \neq q_i} b'_{ik}} & otherwise \end{cases} \qquad (18)$$

These steps are now repeated iteratively until a desired level of convergence.

4. Simulation Results

In order to examine the performance of the proposed outlier detection method, we carry out some experiments on medical datasets from the PhysioNet database [**?**]. The dataset contains 7 attributes: BPmean, systolic BP, diastolic BP, HR, pulse, respiration rate, and SpO2. We only focus on three attributes: BPmean, HR, and SpO2. We use a data sequence of 5000 data, in which the first 4000 data are selected as the training data, and the last 1000 data with injecting faults as diagnosed data.

4.1. Performance Without Sliding Window

Since the 4000 training data are not injected faults, so we can estimate the KDE of the k nearest distance for each sensor reading. Given sensor error probabilities, we inject faults into the 1000 diagnosed data with the position and the value of the injected error all selected by random numbers. The simulation results are the average performance of 300 times randomized experiments.

Table I shows the performance of the k nearest distance based outlier detection method. In this simulation, if the probability of the k nearest distance calculated by the estimated KDE is less than a threshold $\delta = 0.001$, then the diagnosed data is decided as an outlier. From Table I, we can see that the k nearest distance based method has a good outlier detection, but the false alarm rate is high, which leads to that the error rate after executing this method is enormous greater than the original error probability.

Table II shows the performance of the nearest distance based outlier detection method. If the nearest distance is greater than 4, then the diagnosed data is determined as an outlier. From Table II, we can see that the number of errors can be reduced by 50% approximately using this method.

Table III shows the performance of the combination of the k nearest distance and the nearest distance based outlier detection method. From this table, we can see the the performance is better than that of using k nearest distance and the nearest distance based method alone.

Table 1. k Nearest Distance Based Simulation Result (p Indicates Prior Error Probability), OD denotes Outlier Detection, FA denotes False Alarm and EP denotes Error Probability.)

p (%)	5	10	15	20	30	40
OD (%)	84	86	98	97	97	97
FA (%)	9.5	21.4	30.9	38	53.7	67.8
EP (%)	9.4	19.7	26.5	31	38.3	41.6

Table 2. The Nearest Distance Based Simulation Result (p Indicates Prior Error Probability), OD denotes Outlier Detection, FA denotes False Alarm and EP denotes Error Probability.)

p (%)	5	10	15	20	30	40
OD (%)	80	70	70	53	37	31
FA (%)	1.3	1.2	1.2	1.3	3.1	1.9
EP (%)	2.2	4.0	5.6	10.5	21.0	28.7

Table 3. The Combination Method Simulation Result (p Indicates Prior Error Probability, OD denotes Outlier Detection, FA denotes False Alarm and EP denotes Error Probability.)

p (%)	5	10	15	20	30	40
OD (%)	78	69	70	53	37	31
FA (%)	0.3	0.4	0.2	0.7	2.7	1.8
EP (%)	1.4	3.5	4.6	10.1	20.0	28.7

Table 4. The MD In [9] Based Simulation Result (p Indicates Prior Error Probability), OD denotes Outlier Detection, FA denotes False Alarm and EP denotes Error Probability.)

p (%)	5	10	15	20	30	40
OD (%)	72	71	69	72	70	71
FA (%)	0.95	0.93	0.98	0.96	0.83	0.92
EP (%)	2.3	3.8	5.5	6.4	9.6	12.2

As a comparison, Table IV gives the performance of the Mahalanobis distance (MD) and KDE based approach in [9] on the same dataset. In this method,
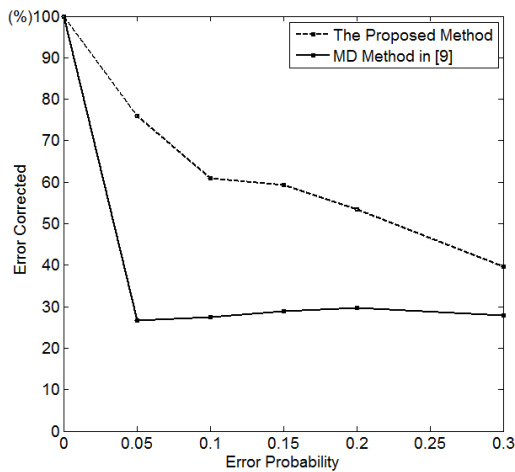
Figure 5. Comparison of Proposed Method and The Method in [9] On Probability of Errors Corrected.
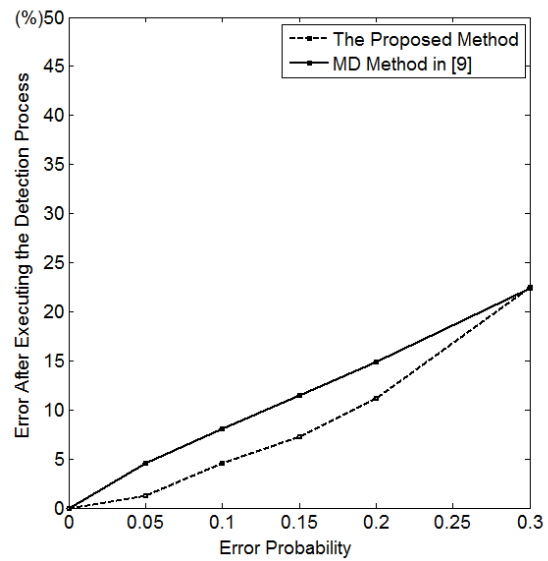

Figure 6. Comparison of Proposed Method and The Method in [9] On Probability of Errors Introduced.

Figure 7. Comparison of Proposed Method and The Method in [9] On Probability of Errors after Executing the Detection Algorithm.
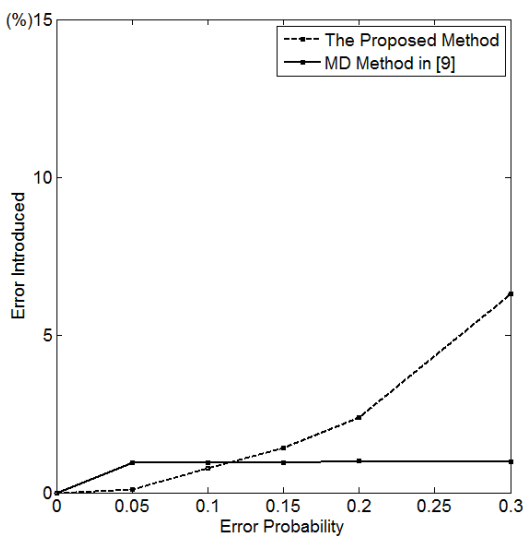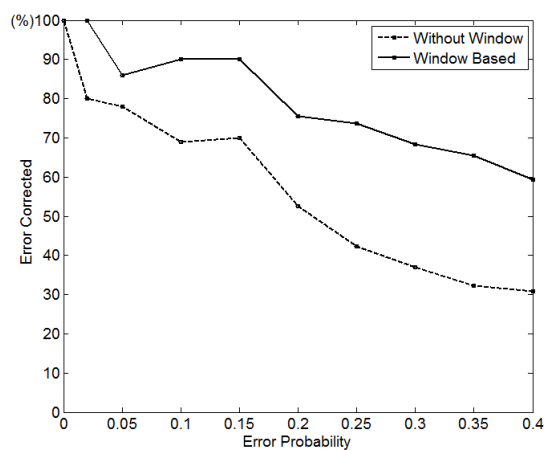

Figure 8. Comparison of With and Without Sliding Window On Probability of Errors Corrected.

when a new data is obtained, the Mahalanobis distance is calculated between the current arrival data and the mean of training data, then KDE is used to estimate the probability of this distance, if it is less than threshold , then the current arrival data is diagnosed as an outlier. From table IV, we can see that the performance of the proposed method is not better than the method in [9]. The reason is that the simulation performance of the outlier detection process depends on the sensor error rate, besides, the range in which the injected errors must lie is another factor influences the outlier detection performance, the closer the outlier to the ground true value, the harder it is to be detected.

In previous experiments, the range of injected errors is wider than the normal value of vital signs. Fig 5-7 give the comparison of the performance of the proposed method and the method in [9], in which range of injected errors is set as the same to the normal value range. From these figure, we can see that the proposed method has a better performance when the faults appear in the range of most normal vital sign data occurs.

4.2. Performance With Sliding Window

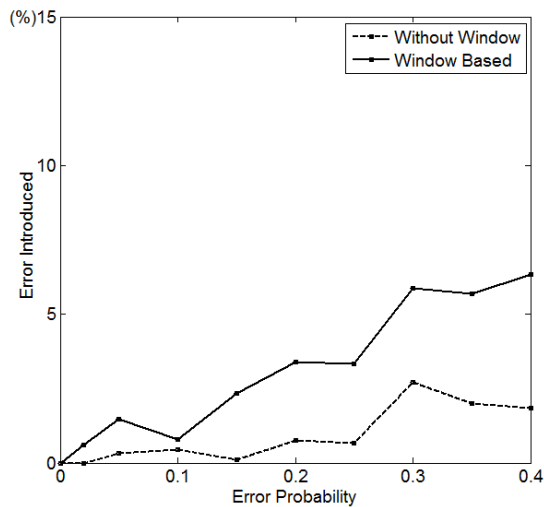Fig. 8-10 show the comparison of the distance based method with and without sliding window. We can see

Figure 9. Comparison of With and Without Sliding Window On Probability of Errors Introduced.
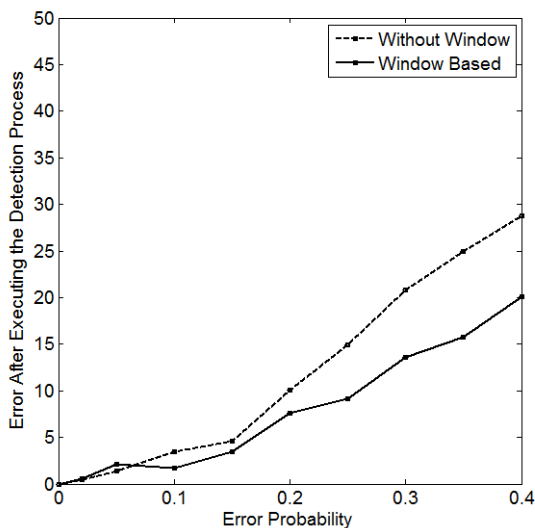

Figure 10. Comparison of With and Without Sliding Window On Probability of Errors after Executing the Detection Algorithm.

that, the method with sliding window has a high outlier detection rate. Although its false alarm rate is a little higher, but the error rate after executing the outlier detection algorithm is obvious lower.

5. Conclusion

Outlier detection is very important for BSNs to avoid false medical diagnosis and false alarms. In this paper, we formalize a distance based method for outlier detection of BSNs. This method consider both distance to k nearest neighbors and to the left and right neighbors. To deal with the condition like successive

errors, we formalize a sliding widow based method to improve the performance of the outlier detection method. To handle errors in the training data, we introduce a Hidden Markov Model based method to estimate the most probable ground truth values which have the maximum probability to produce the training data. Simulation results show that the proposed method possesses a good performance.

6. Acknowledgment

References

[1] B. Krishnamachari and S. Iyengar, Distributed Bayesian Algorithms for Fault-Tolerant Event Region Dectection in Wireless Sensor Networks. IEEE Transactions on Computers, 53(3):241-250, 2004.

[2] X. Luo, M. Dong, and Y. Huang. On distributed fault-tolerant detection in wireless sensor networks. IEEE Transactions on Computers, 55(1): 58-70, 2006.

[3] W. Wu, X. Cheng, M. Ding, K. Xing, F. Liu, and P. Deng. Localized Outlying and Boundary Data Detection in Sensor Networks, IEEE Transactions on Knowledge and Data Engineering, 19(8): 1145-1157, 2007.

[4] A. Annichini, E. Asarin, A. Bouajjani. Symbolic Techniques for Parametric Reasoning about Counter and Clock Systems. The Proceeding of 12th International Conference on Computer Aided Verification, LNCS 1855, pp. 419ÍC434, 2000.

[5] D.J. Kim and B. Prabhakaran. Motion fault detection and isolation in body sensor networks. IEEE International Conference on Pervasive Computing and Communications, pp. 147-155, 2011.

[6] Y. Chen, Y. Zi, H. Cao, Z. He, H. Sun. A data-driven threshold for wavelet sliding window denoising in mechanical fault detection. Science China (Technological Sciences),03: 589-597, 2014.

[7] Jie Ying, T. Kirubarajan, Krishna R. Pattipati. A Hidden Markov Model-Based Algorithm for Fault Diagnosis with Partial and Imperfect Tests. IEEE Transactions on Systems, Man and Cybernetics, 30(4): 463-473, 2000.

[8] O. Salem, A. Guerassimov, A. Mehaoua, A. Marcus and B. Furht. Sensor Fault and Patient Anomaly Detection and Classification in Medical Wireless Sensor Networks. IEEE International Conference on Communications, pp. 4373-4378, 2013.

[9] O. Salem, Y. Liu and A. Mehaoua. Anomaly Detection in Medical Wireless Sensor Networks. Journal of Computing Science and Engineering, 7(4): 272-284, 2013.

[10] Physionet, http://www.physionet.org/cgi-bin/atm/ATM.

[11] M.M. Breunig, H.P. Kriegel, R.T. Ng, *et al.* LOF: Identifying Density-Based Local Outliers. ACM Sigmod Record. 29(2): 93-104, 2000.

[12] K. Niu, F. Zhao and X. Qiao. An Outlier Detection Algorithm in Wireless Sensor Network Based on Clustering. Proceedings of 15th IEEE International Conference on Communication Technology, 2013

[13] S. Rajasegarar, C. Leckie, M. Palaniswami, and J.C. Bezdek, Distributed Anomaly Detection in Wireless Sensor Networks, IEEE ICCS, 2006.