Data Classification in Wireless Sensor Networks

Chen Yilei (Student ID:5080309871)

July 1, 2011

Abstract

Enerny consumption is an important issue in wireless sensor networks. In sensor networks, sensors collect data and transmit them to the sink. But in some cases, there might be unimportant data which is not necessary to be transmited. Some algorithms of classifying data are introduced and analysised through emulation in this paper. By using certain classification and transmission strategy, we optimize the energy consumption problem.

Index words: data classifying, wireless sensor networks

1 Introduction

In some applications of sensor networks, data collected by sensor could be classified as important and normal ones. Supposed a wireless sensor network is distributed in the mountain to detect if there was a fire in the forest. In most cases no fire exists, the status of 'normal' would be collected by sensor, which contains comparativly 'little' information. These 'normal' information could endure more delays. If the fire breaks out, higher tempreture would be detected by the sensor. Here the 'important' information need fast and direct transmission.

The whole strategy solving this problem was proposed in [1] by Luo Zhe and Qian Chunhua. In this paper, I would do a further analysis on the data classification problem with certain test and emulation.

The rest of the artical would be organized as follows: In Section II we discuss the data classification algorithms given in [1]. In Section III, several simulations will be introduced. More discussion would be proposed in Section IV.

2 Basic models of data classification

The probability of data collected by sensor is denoted as $P(X = x_i)$, so the effective information collected by sensor is the entropy of X, [1]

$$H(X) = -\sum_{x_i} P(X = x_i) log_2(P(X = x_i)) \tag{1}$$

Intuitively, the classification of data would be accomplished according to the character of its distribution, as shown in figure 1 and figure 2:

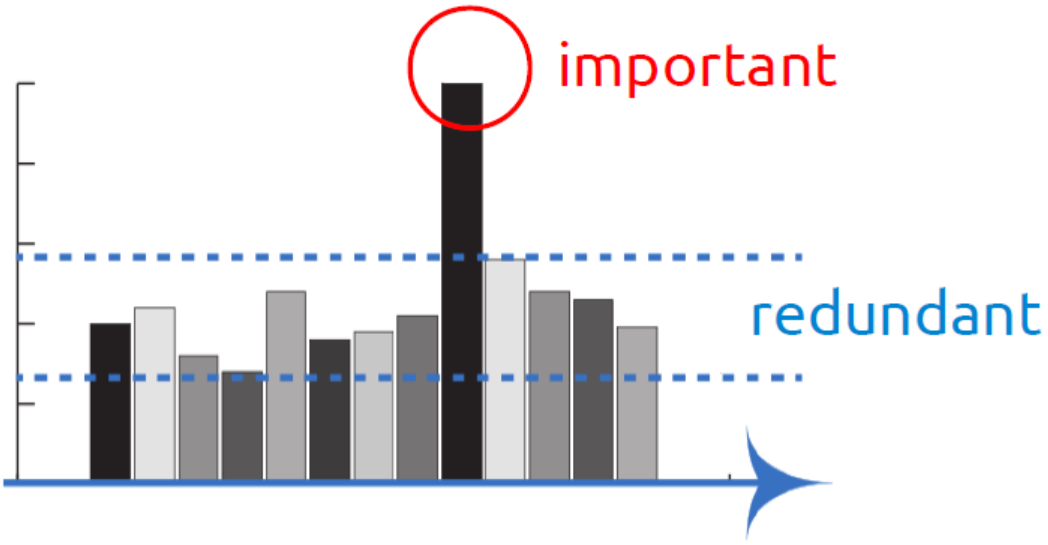


figure 1: classify data as normal or important

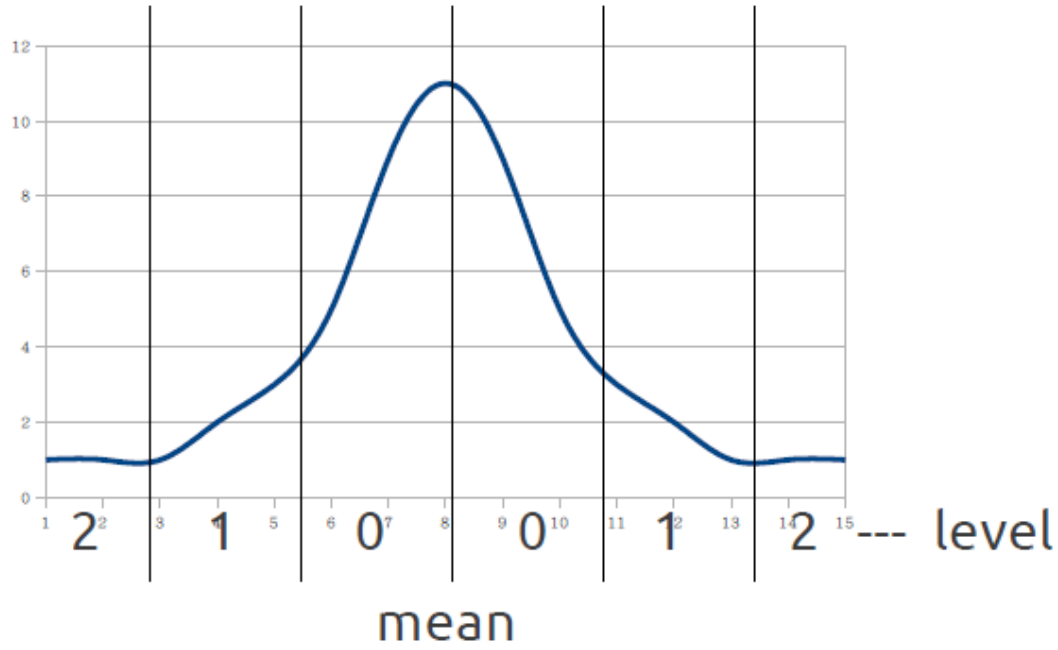


figure 2: the relationship between data distribution and classification

Mean value and deviation are the common parameters used in classification,

$$level_{importance}(x) = floor(|x - m| / d) \tag{2}$$

Concerning these two parameters, there are 4 algorithms which make use of the distribution in 4 different ways, as was discussed below.

2.1 naïve classification

Using default value of system design is the simplest way to classify data. Designers of system estimate the characteristic of data and give default values of classification. Setting a default mean and deviation in each sensor is the simplest and most memory-saving way.

2.2 Static a priori distribution

This method could be identifyed as the one between the preceding and the following. Similar to naïve classification, we judge data acoording to static parameters. The difference was lying on the format of data. We store a priori distribution instead of pure mean and deviation. The spacial cost was raised, but its the gateway to further improvement.

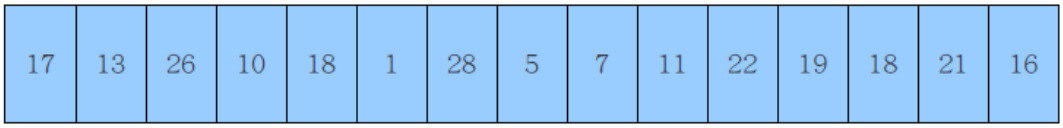


figure 3: static memory of a sensor

2.3 Dynamic distribution

In some cases, the distribution of data changes with time. Take the example of forest into concern, the tempreture on the mountain changes with sun rising and sunset, day after day. So a static scale cannot handle the change. A dynamic classification system can solve this problem and improve performance of classification system.

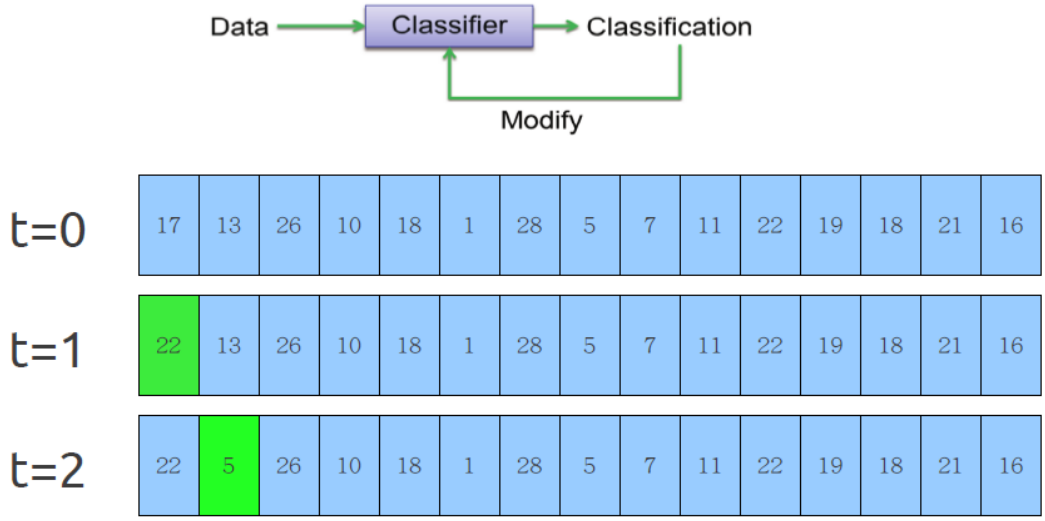


figure 4: Dynamic distribution of a sensor

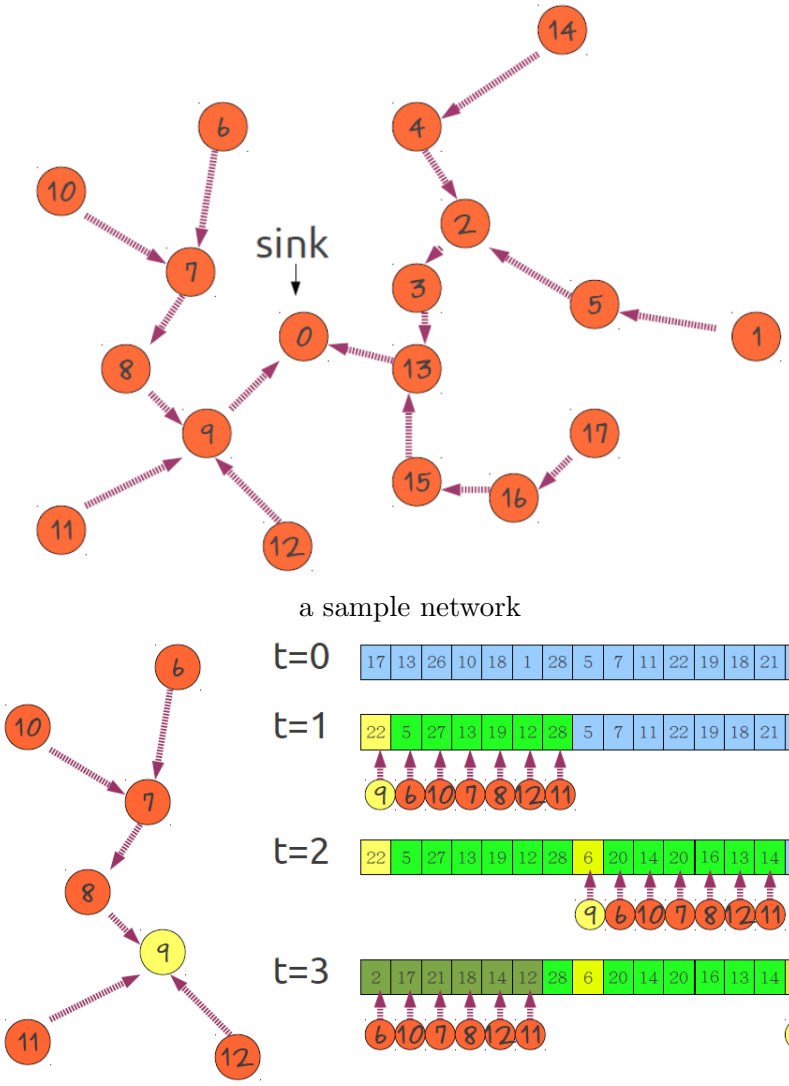
A dynamic classification system means that the standard of classification must be adjusted with time. Only sensors adjust their standards according to their own collected

data was an easy way to realize dynamic classification system. Figure 4 demonstrates the specific steps of this dynamic classification system.

2.4 Dynamic Classification by Sensors and Relays

Considering not only the dynamic redundancy in time but also that in network, node requires to recalculate classification of data when it receives and combines data of same type from other nodes for relay.

This dynamic classification method classifies data when data is collected or data is transmitted in network both. In this method, class of data is remodified when new nodes catch it.



a sample network

figure 5: Dynamic Classification by Sensors and Relays

Figure 5 is a sample wireless sensor network, now we illustrates the process of dynamic classification of node 9.

The main difference between these two dynamic classification systems is that one only uses collected data to modify probability distribution function but the other uses all data which nodes can get to do it.

3 Simulations and analysis

In the following paragraphs, several tests are designed to check the effectiveness of the three algorithms.(the two static ones are combined)

Assumption: All nodes shares the same distribution each time, with gauss distribution N(m(t),d), in which d always equal to 10 and m change with t respectively in four ways:

1. constant condition: $m(t) = 50$
2. first order: $m(t) = 50 + t/100$
3. second order: $m(t) = 50 + t(1000 - t)/1000$
4. sinusoidal: $m(t) = 50 + 5 * sin(t/100)$

The output of each condition would be the error rate of classification, which is defined as:

$$Err(t) = (\sum_{i=0}^{t} |level_{i.classified} - level_{i.acture}|)/t \tag{3}$$

In all the experimental figures, blue line stands for the error rate of static algorithms, while the oringe and yellow ones represent the dynamic and enhenced dynamics ones.

Brief comment would be given at each testband while the compound comparison will be made in section 4.

3.1 constant condition

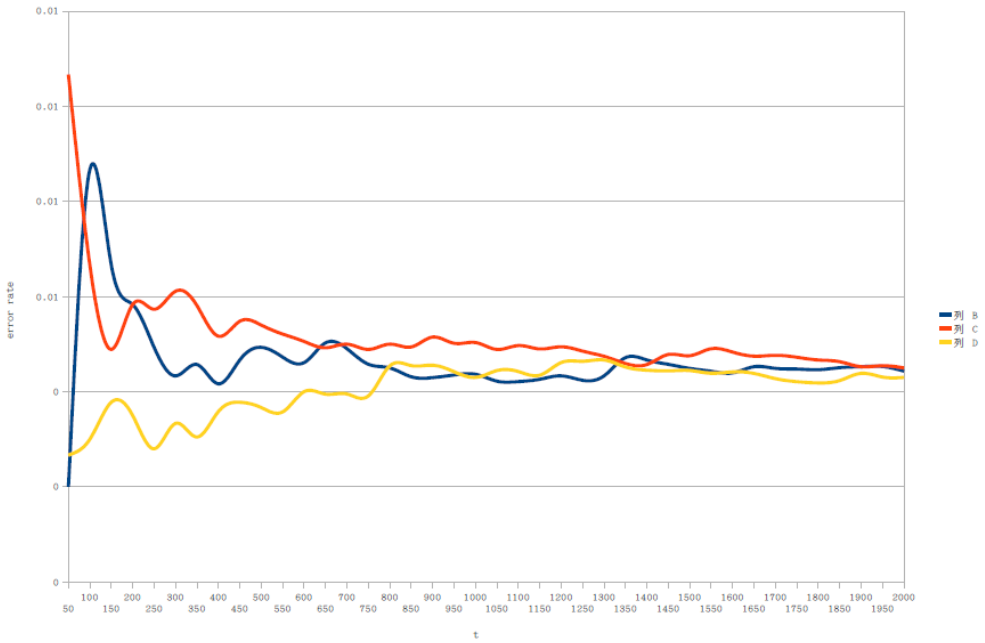


figure 6: error rate of constant condition: $m(t) = 50$

In the constant condition, 3 methods perform approximately the same as time flows.

3.2 first order condition

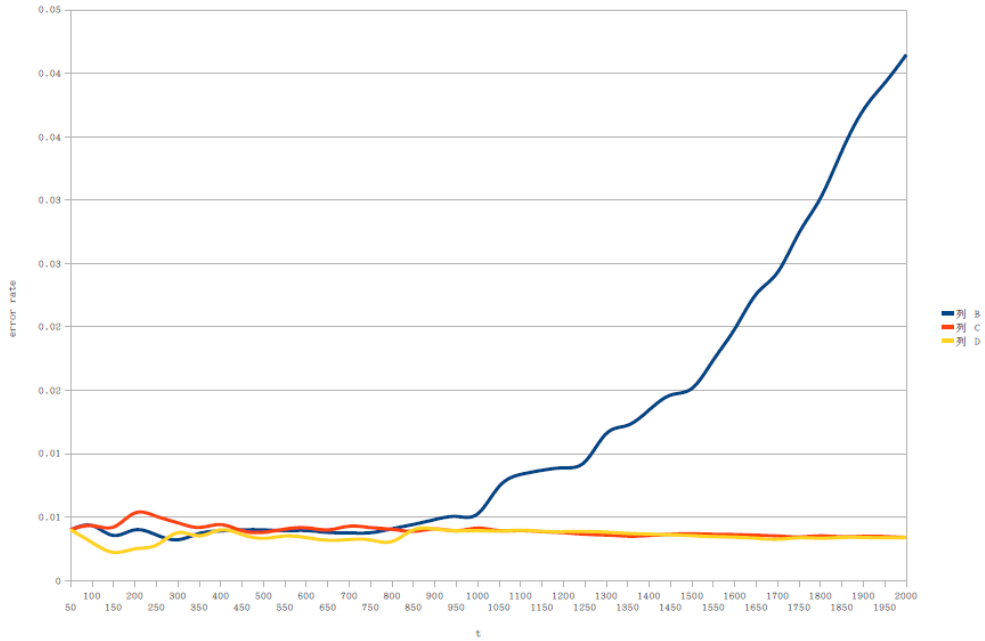


figure 7: error rate of first order condition:$m(t) = 50 + t/100$

Default value cannot evalue the importance of speed model, while dynamic models play well as same.

3.3 second order condition

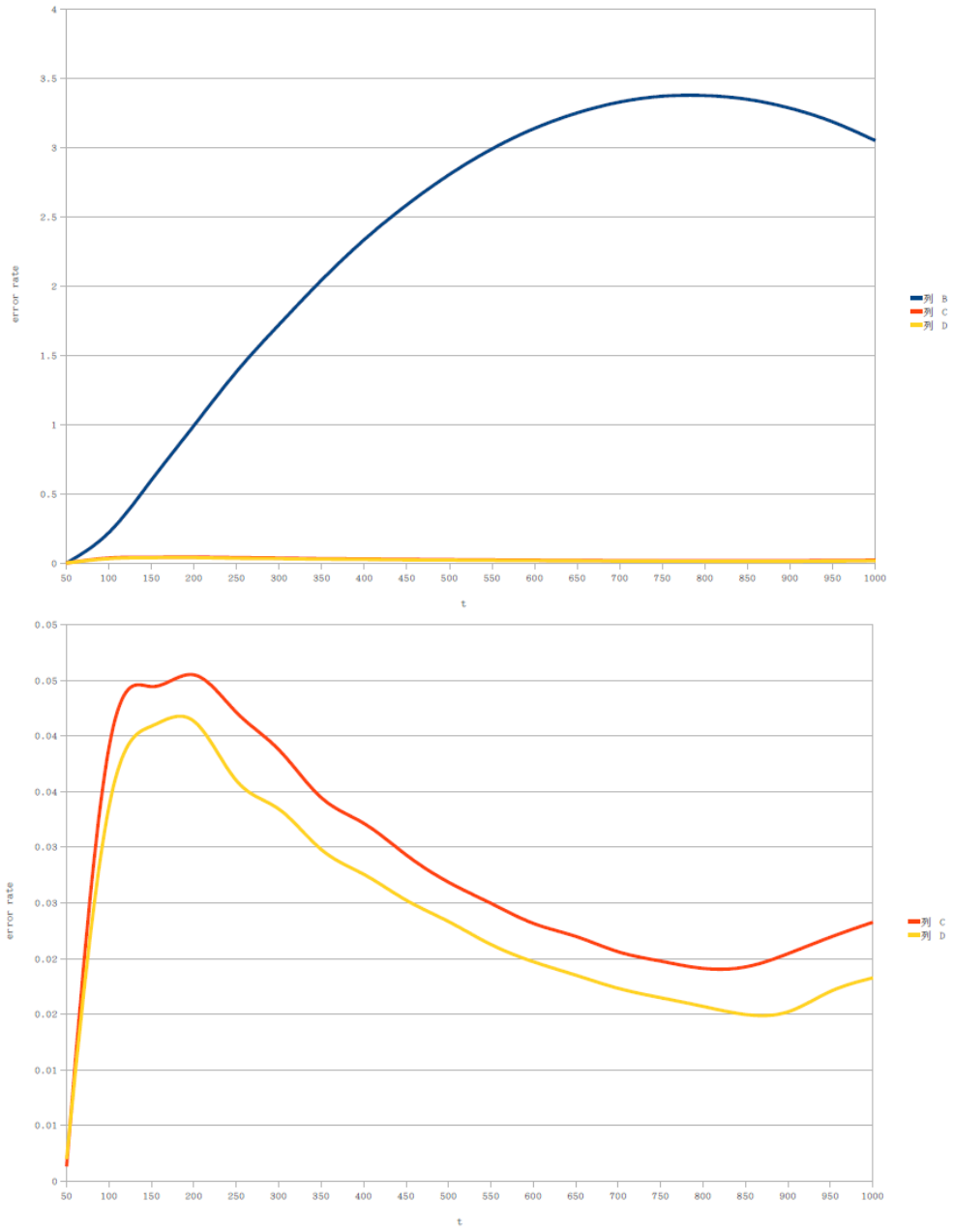


figure 8: error rate of second order condition: $m(t) = 50 + t(1000 - t)/1000$
(with the upper figure include blue and the lower figure not including static algorithm)

Constant model goes far beyong, but what about the dynamic models? If the ex-

periment continued with time, we would have seen that the dynamic models were also divergent. After all they are "first order dynamic models".

3.4 sinusoidal condition

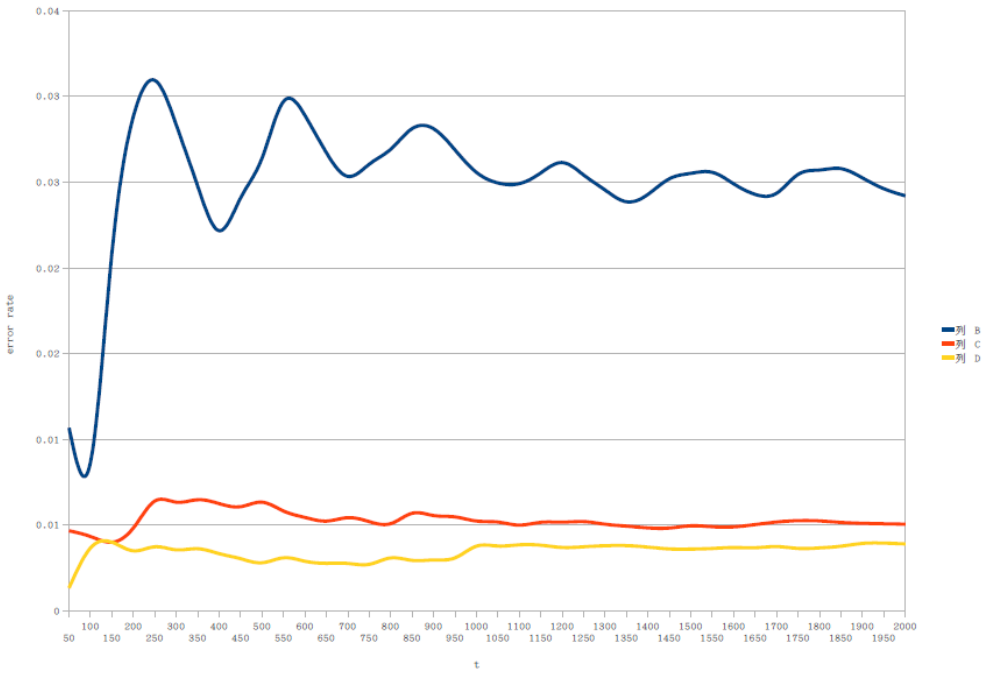


figure 9: error rate of sinusoidal condition: $m(t) = 50 + 5 * sin(t/100)$

All algorithms converge in this situation, the only diference is the constant.

4 Comparison and conclusion

In constant and sinusoidal cases, error rates of all the four algorithms are convergent, while in the other cases, dynamic feedback mechanism must be introduced. In the second order condition, even the dynamic methods could not make sure of the convergence, more limits are necssary in application.

Compare these two dynamic methods, Dynamic Classification by Sensors and Relays performs better, but only by constant. The order of error rate of these two methods are same.

By the way, the better algorithm considers more details in data transmission, which is the key issue in energy comsumption. So the lower error rate might cost more energy. That should be considered in the designing progress.

Reference:
[1] Cognitive Sensor Network for Smart Grid, Zhe Luo, Chunhua Qian, Xinbing Wang