Computer Science and Information Systems 19(1):25–45 https://doi.org/10.2298/CSIS201218041L

Link quality estimation based on over-sampling and weighted random forest

Linlan Liu[1], Yi Feng[3, 1], Shengrong Gao[1], and Jian Shu[2]

[1] School of Information Engineering, Nanchang Hangkong University,
330063 Nanchang, China
765693987@qq.com
1322415547@qq.com
[2] School of Software, Nanchang Hangkong University,
330063 Nanchang, China
shujian@nchu.edu.cn
[3] School of Engineering, Zhejiang Normal University Xingzhi College,
321000 Jinhua, China
458018002@qq.com

Abstract. Aiming at the imbalance problem of wireless link samples, we propose the link quality estimation method which combines the K-means synthetic minority over-sampling technique (K-means SMOTE) and weighted random forest. The method adopts the mean, variance and asymmetry metrics of the physical layer parameters as the link quality parameters. The link quality is measured by link quality level which is determined by the packet receiving rate. K-means is used to cluster link quality samples. SMOTE is employed to synthesize samples for minority link quality samples, so as to make link quality samples of different link quality levels reach balance. Based on the weighted random forest, the link quality estimation model is constructed. In the link quality estimation model, the decision trees with worse classification performance are assigned smaller weight, and the decision trees with better classification performance are assigned bigger weight. The experimental results show that the proposed link quality estimation method has better performance with samples processed by K-means SMOTE. Furthermore, it has better estimation performance than the ones of Naive Bayesian, Logistic Regression and K-nearest Neighbour estimation methods.

Keywords: Wireless Sensor Network, Link Quality Estimation, Weighted Random Forest, Oversampling.

1. Introduction

The Wireless Sensor Network [1] (WSN) is a self-organizing network formed by wireless communication through various inexpensive micro sensor nodes with sensing capabilities, computing power, and communication capabilities. In recent years, WSN has been widely used in different fields, such as military target tracking, natural disaster rescue, biomedical health monitoring, hazardous environment detection, and so on.

WSN is different from traditional network because of its design and limited resources. Sensor nodes are usually deployed in the wild environment with few people. The network is built by self-organization between nodes, and the information in monitoring areas is sent to the sink node through multi hops. The ability of processing, storage and communication of sink nodes are relatively great. They process the received information such as fusion, calculation and storage, and transmit the data to the remote terminal equipment through the external networks.

Wireless communication technology is used to transmit sensing data between sensor nodes. Due to the small communication range and low bandwidth of sensor nodes, a large number of sensor nodes are needed to monitor one area together, so as to obtain relevant data.

Due to the different deployment environments of sensor nodes, there are great differences in the interference received by nodes [2]. In some harsh field environments, interference will seriously affect the communication, resulting in unstable communication links, and data loss in the process of transmission. In the existing routing protocols [3], although there is a packet loss retransmission mechanism, which allows sending nodes to retransmit data packets, for worse quality links, this will cause nodes to repeatedly send data packets, result in seriously consuming node energy. Selecting high quality links for communication through link quality estimation will improve the efficiency of data transmission and reduce the number of retransmissions. For nodes with limited energy storage, this will further reduce the energy consumption of the nodes, thus extending the life cycle of the entire network.

The existing link quality estimation methods mainly use physical layer parameters and link layer parameters to construct a link quality estimation model. However, in the existing methods, the distribution problem of sample data is not considered, so that it will affect the accuracy of link quality estimation, especially when imbalanced samples occur. The existing estimation models tend to be more inclined to the majority of sample data, thus affecting the classification performance of the model on the minority of sample data. Therefore, this paper employs K-means SMOTE to deal with imbalanced sample data. K-means is used to cluster samples. In order to balance samples, SMOTE is employed to increase minority class samples. Make the distribution of link quality samples achieve balance through K-means SMOTE. Benefit from the great effect on the imbalanced data processing of weighted random forest, weighted random forest is adopted to construct the estimation model, which lean the interest towards to the correct classification of rare class, and improve the prediction performance.

Considering the fluctuation and asymmetry existing in the links are fully considered. The mean, variance and asymmetry metrics of the physical layer parameters are selected as the link quality parameters, and a link quality estimation model based on weighted random forest is constructed.

The main contributions of this paper are as follows:

(1) A method of processing imbalanced samples is proposed during the link quality estimation process. Aiming at the problem of link quality imbalanced samples, this paper uses synthetic minority over-sampling technique to deal with imbalanced samples. K-means clustering is used to divide the samples into different clusters, and minority samples is increased by stochastic linear interpolation in each cluster, so that samples of each link quality level is balanced.

(2) A link quality estimation model based on weighted random forest is proposed, which set decision trees with poor classification performance smaller weight and good classification performance bigger weight.

(3) The evaluation metrics suitable for imbalanced samples are employed to propose evaluate models. This paper comprehensively evaluates the performance of the proposed evaluate model by accuracy, recall and F1 value.

The rest of the paper is organized as follows. The state-of-the–art is discussed in Section 2. The selection of link quality parameters is addressed in Section 3. The division of link quality level is described in Section 4. The processing of imbalanced data is presented in Section 5. The link quality estimation model is constructed in Section 6. The experiments and analysis are presented in Section 7. The conclusion is made in section 8.

2. Related Work

Link quality estimation has attracted many scholars to carry out in-depth research, the existing methods fall mainly into the following categories: link quality estimation method based on link characteristics [4], link quality estimation method based on probability estimation theory, and link quality estimation method based on intelligent learning [5].

2.1. Link quality estimation method based on link characteristics

Such methods mainly use the received signal strength indicator (RSSI), link quality indicator (LQI), and signal to noise ratio (SNR) to estimate link quality. In order to solve the link quality problem of the upper communication network in the transmission detection system, the literature [6] analyzes the network characteristics of WSN and selects the optimal next hop node in the routing establishment stage according to the hop count and network environment. Literature [7] proposes a simple, accurate and low-cost link quality estimation technique, which is suitable for WSN scenarios with limited resources. Kalman filtering and fuzzy logic are used to optimize the influence of RSSI and LQI on link quality at low cost. Experimental results show that the method realizes error-free transmission at the cost of less delay.

Literature [8] employs several link property indicators in the fuzzy algorithm, without specific calculation methods and formulas. Literature [9] proposes a new link delay aware energy efficient routing metric, namely, predicted remaining deliveries (PRD), for routing path selection of wireless sensor networks deployed in harsh environments. Literature [10] proposes a link quality metric method based on triangle metric. The link quality can be evaluated quickly and reliably with fewer link detection packets by using geometric methods combined with packet reception rate (PRR), LQI and SNR information.

Literature [11] establishes a generalized model that connects PLR to link quality indicator, a physical layer link quality measure, and packet length under diverse environmental conditions. By rich observations on the spatio-temporal characteristics of the dependency of packet loss rate (PLR) on LQI and packet length, propose a packet loss rate model as a function of LQI and packet length, that is applicable in all experimented scenarios. A comparison with a literature LQI-only based PLR model shows that the proposed model has higher accuracy for various packet lengths.

2.2. Link quality estimation method based on probability estimation theory

Such methods focus on estimating the packet reception rate at the receiving end in communication. In literature [12], based on the analysis of common lognormal path loss models, the wireless link quality characterized by SNR can be decomposed into two parts with different characteristics, a time-varying nonlinear part and a non-stationary random part. By processing the two parts separately, a new link quality estimation method WNN-LQE is proposed to obtain the confidence interval of link quality. In literature [13], the mathematical model of exponentially weighted moving average (EWMA) and link quality prediction are used to solve the problem of unstable packet transmission rate. The experimental results show that the correlation is established in the EWMA model. Literature [14] proposes a three-layer impulse response framework to analyze the time fading effect of fixed wireless links in industrial environment. The original observation of link quality is mapped to the distributed parameter space. In the new space, the measurement noise has a constant covariance, meeting the requirements of linear Kalman filtering. Based on this, an enhanced Kalman filter is designed, which can obtain better link quality prediction effect in industrial environment.

Estimating the quality of a link is a key primitive in WSNs, as upper layers use this piece of information in making performance-critical decisions. Literature [15] proposed Rep, a novel sampling scheme able to extract the link quality from the packet repetitions of low-power preamble sampling MACs. The experiments show that Rep reduces the energy and traffic used for link estimation by one order of magnitude, and increases the speed of the process by one order of magnitude, while maintaining state-of-the-art accuracy.

Considering the selection of high quality path in mobile ad-hoc network is a critical issue due to mobility of nodes and variable channel conditions during data transmission, Literature [16] proposed routing protocol named as modified expected transmission count enabled ad-hoc on-demand distance vector (MXAODV). Select the path with the smallest number of hops as the next hop of the mobile ad hoc network, and modify the expected transmission number at the same time. Extensive simulation has been done to analyze the performance of MXADOV. Significant improvements found in terms of throughput, packet delivery fraction, normalized routing load and end-to-end delay.

2.3. Link quality estimation method based on intelligent learning

Such methods are mainly modeled by intelligent learning methods such as machine learning and pattern matching.

Literature [17] proposed a lightweight, fluctuation insensitive multi-parameter fusion link quality estimator, which have characteristics of high flexibility and low overhead.

Signal-to-Noise Ratio and Link quality indicator are preprocessed by exponential weighted Kalman filtering. These two parameters are fused using lightweight weighted Euclidean distance. Link quality is estimated quantitatively with the mapping model of the fused parameter and packet reception ratio, which is constructed by logistic regression. Experimental results show that the proposed estimator could reflect link quality more realistically. Compared with some similar estimators, estimate error of the proposed one is reduced by 18.32% to 60.11%.

Literature [18] uses naive Bayes (NB), logistic regression (LR), artificial neural networks (ANN) to build a prediction model. By combining link layer parameter PRR and physical layer parameters RSSI, LQI and SNR, link quality is predicted. Literature [19] uses RSSI and LQI as estimation parameters. It divides link quality into five grades according to PRR, and establishes a multi-classification link quality estimation mechanism based on support vector machine. Literature [20] proposes a missing data estimation algorithm based on k-nearest neighbor (KNN), which uses a linear regression model to describe the spatial correlation of data from different sensor nodes and uses data information from multiple neighbor nodes to estimation missing data. Literature [21] proposes a link quality estimation algorithm based on stacked autoencoder. The zero-filling method is developed to process the original missing link information. The SAE model is used to extract the asymmetric characteristics of the uplink and downlink.

The estimation method based on the physical layer and link layer parameters only estimates the link quality from a single perspective and cannot fully reflect the link characteristics. The estimation method based on machine learning adopts a data-driven approach, which constructs a learning model to mine the relationship between link data and link quality by collecting a large amount of link quality data.

Considering the fluctuation and asymmetry of the link, this paper selects the mean, variance and asymmetry indicators of the physical layer parameters as the link quality parameters and determines the link quality level according to the PRR so as to estimate the link quality. Since sensor nodes are often deployed in harsh environments, they are subject to environmental noise during communication, which makes the link quality worse. Samples of good link quality and bad link quality account for a small proportion, and samples of medial link quality account for a big proportion. So, the samples show imbalance. If the samples are directly used for training, the model will produce deviation. Therefore, the samples need to be processed before training. In this paper, K-means SMOTE is used to preprocess the samples to reduce the imbalance. And a link quality estimation model based on weighted random forest classification (WRF) algorithm is constructed to estimate the link quality.

3. Selection of Link Quality Parameters

In the process of link quality estimation in this paper, the link quality level is estimated according to the physical layer parameters. The relationship between the physical layer parameters and link quality levels can be mined through the link quality estimation model. Reasonable link quality parameters can better characterize the status of links, so as to improve the effect of the estimation model.

WSN is often deployed in harsh environment. It is easy to be affected by the environment in the communication, causing instability of communication links and making link quality volatility and asymmetry [22]. The physical layer parameters RSSI, LQI, and SNR can quickly detect link changes, which can reflect the sensitivity of the link. So, we choose physical layer parameters as link quality parameters. The asymmetry level (ASL) of the link is reflected by difference between physical layer parameters of uplink and downlink, and the stability of the link is reflected by the variance of the physical layer parameters. There are defined as follows:

$$Input = [\overline{PHY}, \sigma^2(PHY), ASL(PHY)] \tag{1}$$

Where

$$\begin{aligned} ASL &= \left| PHY_{up} - PHY_{down} \right| \\ PHY &\subset (RSSI, LQI, SNR) \end{aligned} \tag{2}$$

4. Division of Link Quality Level

We take link quality level to measure link quality. Literature [23] divides the link quality into three levels according to the PRR value, which are good link, middle link and bad link. Experiments show that the average number of consecutive lost packets in good links is 1.6, the number of consecutive lost packets in middle links and poor links is 5.3 and 56.8, respectively. The link can be well distinguished by the PRR value. In this paper, the link quality level is divided by the same standard, and the link quality is divided into three levels by the PRR value. The specific division criteria are shown in Table 1.

Table 1. Divide Standard.

Link quality level	Link status	PRR
Level 1	Good Link	[80,100]
Level 2	Middle Link	[20,80)
Level 3	Bad Link	[0,20)

5. Processing of Imbalanced Data

The experiments are conducted in parking, indoor and forest scenarios. Link quality samples in different scenarios with different interferes have different distribution characteristics. In some real-world scenarios, the link quality samples of a certain level account for a small proportion of the total sample set is small, which leads to the imbalance of link quality samples distribution.

Due to the imbalance of link quality samples, if the samples are directly used for training, there will be a certain deviation to the classification model, which will make the model more inclined to the majority class samples and eventually lead to the decline of the learning performance of the classification model. For the processing of imbalanced samples [24], the distribution of imbalanced samples can be changed through data sampling to reduce the degree of data imbalance. Commonly used methods mainly include over-sampling [25] and under-sampling [26]. Over-sampling improves the classification performance of the model for minority class by adding the number of minority class samples, while under-sampling reduces the imbalance of data by reducing the number of classes of majority class samples. Since under-sampling will delete data, it may cause some important information of majority class samples to be lost. When there are few minority class samples, the distribution of samples will be balanced by under-sampling, which will cause a too small data set and further limit the performance of the classifier. The commonly used over-sampling method is random over-sampling, which replicates a small number of minority class samples randomly, thereby increasing the number of minority class samples and changing the degree of data imbalance. However, it does not add additional information to minority class samples, which increases the training time and easily leads to over-fitting of the model.

Chawla [27] et al. proposed SMOTE in 2002. This method not only replicates existing minority class samples, but also creates synthesizes samples by interpolation, which can avoid the over-fitting risk of random over-sampling. During the execution of SMOTE algorithm, a sample a is randomly selected from minority class samples, and a sample b is randomly selected from its nearest neighbor. And then, random linear interpolation is performed between samples a and b, namely new sample $x=a+w*(b - a)$, where is the random weight between (0, 1).

SMOTE algorithm still has some deficiencies in dealing with within-class imbalance and noise. There are mainly two problems. First, when randomly selecting minority class samples for uniform over-sampling, the problem of between-class imbalance can be effectively solved, but the problem of within-class imbalance is ignored. In areas where minority class samples are dense, the number of samples will further increase, while in areas where minority class samples are sparse, the samples are still sparse. Second, SMOTE algorithm may further amplify the noise existing in the data. When the selected sample a is noise in majority class samples, the linear interpolation is performed by selecting the nearest neighbor, and the resulting synthetic sample is likely to be noise, which will reduce the classification performance of the trained model.

K-means SMOTE algorithm [28] combines K-means clustering and SMOTE algorithm to avoid noise generated by carrying out over-sampling in the clustering area and solves the problem of between-class imbalance and within-class imbalance. Due to the adoption of K-means clustering, when the samples synthesis is performed in the cluster region, the sparse samples distribution region synthesizes more samples than the dense samples region, thus solving the within-class imbalance problem.

K-means SMOTE algorithm includes three steps: clustering, filtering and over-sampling. In the clustering step, the entire data is clustered into clusters by the K-means algorithm. The filtering step selects clusters with a high proportion of minority class samples, determines the number of synthetic samples assigned to each cluster, and assigns more synthetic samples to clusters with sparse sample distribution. In the over-sampling step, SMOTE is applied to the selected cluster to achieve a balance between the majority and minority samples.

K-means is a commonly used unsupervised clustering algorithm in data mining. For a given sample set $D=\{x_1, x_2, \mathrm{L}\ , x_m\}$, randomly selecting k samples from D as initial mean vectors $\{\mu_1, \mu_2, \cdots, \mu_k\}$, calculating the distance $d_{ji} = \left\| x_j - \mu_i \right\|_2$ between the sample $x_j\,(j = 1, 2, \cdots, m)$ and each mean vector $\mu_i\,(1 \le i \le k)$, and dividing the sample x_j into clusters with the smallest d_{ji} according to the size of d_{ji}. In each cluster, calculate new mean vector μ_i' and replace the original mean vector until current mean vector is not updated or the maximum number of iterations is reached. Finally, the clustered $C = \{c_1, c_2, \cdots, c_k\}$ is obtained.

The filtering step selects clusters to be over-sampling and determines the number of samples to be generated in each cluster. The selection of clusters is determined according to the proportion of minority class samples and majority class samples in each cluster, and clusters with at least 50% minority class samples are selected. The cluster with a higher proportion of minority class samples can also be selected through the hyperparameter imbalance rate threshold of K-means SMOTE algorithm. The default value is 1, and the imbalance rate threshold of cluster c_i is defined as:

$$irt = \frac{majority\ count(c_i) + 1}{minority\ count(c_i) + 1} \tag{3}$$

In order to determine the generating sample number of each cluster, it is necessary to assign sampling weights to the selected clusters and assign bigger weights to the clusters with sparse samples to generate more samples. The calculation steps of sampling weights are as follows:

1) For each filtered cluster f, the Euclidean distance matrix between minority class samples is calculated, ignoring majority samples.

2) The average distance of each cluster is obtained by calculating the average value of non-diagonal elements in the distance matrix.

3) Calculate the density of minority class samples in the cluster.

$$density(f) = \frac{minority\ count(f)}{average\ minority\ distance(f)^m} \tag{4}$$

Where m is the number of features.

4) Calculate the sparsity of minority class samples in a cluster.

$$sparsity(f) = \frac{1}{density(f)} \tag{5}$$

5) The sampling weight of each cluster is defined as the sparsity of the cluster divided by the sum of the sparsity of all clusters, and the sum of the sampling weights of all clusters is 1.

In the over-sampling step, SMOTE is used for over-sampling in the selected clusters, and the number of samples to be generated for each cluster is $\left\| sampling\ weight(f) \times n \right\|$, where n is the total number of samples to be generated. The process of imbalanced link quality samples processing based on k-means SMOTE is shown as Algorithm 1.

Algorithm 1 Imbalanced sample processing algorithm

```
  Input: input space $Z = \{(\overline{PHY}_i, \sigma^2(PHY_i), ASL(PHY_i)), lv_i\}$, number of clusters $k$,
imbalance rate threshold $irt$, nearest neighbor $knn = 20$ Error! Reference source not found..
  Output: The synthesized link quality sample after the over-sampling.
  begin
  clusters $\leftarrow$ $K\text{-}means(\overline{PHY}_i, \sigma^2(PHY_i), ASL(PHY_i))$ ;
  filtered cluster $\leftarrow \varnothing$ ;
     for C $\in$ clusters do
        Calculate imbalance ratio of link quality samples by eq(3);
        If imbalance ratio < $irt$ then
          filtered cluster $\leftarrow$ filtered cluster $\cup \{C\}$;
        end
     end
     for $f$ $\in$ filtered cluster do
       average minority distance $\leftarrow$ $mean(euclidean\ distance(f))$ ;
       Calculate $density(f)$ by eq(4);
       Calculate $sparsity(f)$ by eq(5);
     end
     sparsity sum $\leftarrow \sum_{f \in filtered\ clusters} sparsity\ (f)$ ;
     sampling weight $\leftarrow \frac{sparsity(f)}{sparsity\ sum}$ ;
     generated link quality samples $\leftarrow \varnothing$ ;
     for $f$ $\in$ filtered cluster do
        number of samples $\leftarrow \|sampling\ weight(f) \times n\|$ ;
        generated link quality samples $\leftarrow$ generated samples $\cup$
       $SMOTE(f, number\ of\ samples, knn)$ ;
     end
     return generated link quality samples
  end
```

6. Link Quality Estimation

In different experimental scenarios, the link quality level distribution of the collected sample data is imbalanced. The K-means SMOTE method is used to randomly linearly interpolation on the original samples to increase the number of minority class samples, and improves the classification accuracy of the estimation model for minority class samples.

6.1. Construct estimation model

Link quality estimation is a process of estimating link quality level based on selected link quality parameters, which is essentially a multi-classification problem. Random forest algorithm [29] is not prone to over-fitting in the training process, and can identify overlapping samples between classes, which is suitable for multi-classification problems. Two random processes are introduced into the random forest. One is to randomly extract samples by Bootstrap resampling method when constructing training subsets to increase the differences among the subsets. The second is to randomly select features from the total number of features to construct the decision tree in order to ensure the diversity of the decision tree and reduce the similarity between the trees. Since the randomness is guaranteed by the two random processes, the decision tree may not be pruned during construction process, and it can also ensure that the random forest is not prone to over-fitting.

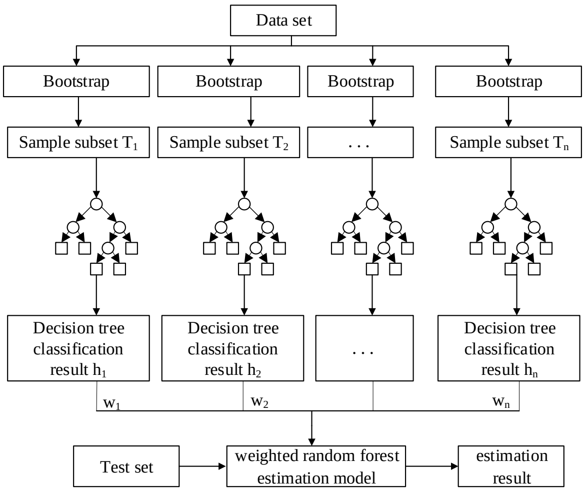
Data set
Bootstrap
Bootstrap
Bootstrap
Bootstrap
Sample subset T_1
Sample subset T_2
. . .
Sample subset T_n
Decision tree classification result h_1
Decision tree classification result h_2
. . .
Decision tree classification result h_n
w_1
w_2
w_n
Test set
weighted random forest estimation model
estimation result

Fig. 1. Link quality estimation model based on weighted random forest.

The traditional random forest algorithm produces the final result through combined voting and gives the same weight to each decision tree. This method gives a too big weight to the decision tree with low classification performance, leading to reducing the overall classification performance. In this paper, a WRF [30] algorithm is applied to construct a link quality estimation model. The model structure is shown in Figure 1. WRF is an improved algorithm for random forests. It assigns smaller weights to decision trees with low classification performance and bigger weights to decision trees with high classification performance.

Let the training sample set processed by K-means SMOTE be $T=\{(x_i,y_i)\}$,where $x_i=\{\overline{PHY}_i,\sigma^2(PHY_i),ASL(PHY_i)\}$, $y_i=\{lv_i\}$, $i=1,2,...,N$, N is the total number of samples, x_i is the vector composed of the selected link quality parameters, and y_i is the link quality level value. In the training process of WRF, the training sample set is split into different training subsets $T_1, T_2, \cdots, T_n$ by Bootstrap resampling method, and the training process for each subset is independent of each other and does not affect each other. The link quality decision trees are constructed for the training subsets, shown in Fig.1. The key to decision tree learning is to select the optimal partitioning attribute. ID3 decision tree learning algorithm selects attributes with large information gain when dividing nodes, but the information gain criterion will bring certain deviation. For attributes with a large number of values, the corresponding information gain is larger. In order to reduce this influence, C4.5 decision tree algorithm selects attributes with large gain rate when dividing nodes. In this paper, C4.5 algorithm is used in the process of decision tree construction, that is, gain rate is used to select features when dividing attributes. The calculation process of gain rate is as follows:

1) Calculating information entropy

$$Ent(T) = \sum_{k=1}^{|y|} p_k \log_2 p_k \tag{6}$$

Where T is the current sample set, p_k is the proportion of the k th sample in the sample set, and $|y|$ is the number of categories of the sample set, which is the number of link quality levels in this paper.

2) Calculating information gain

$$Gain(T,a) = Ent(T) - \sum_{v=1}^{V} \frac{|T^v|}{|T|} Ent(T^v) \tag{7}$$

Where T^v is the subsets according to attribute a, V is the number of divided subsets, $|T^v|$ is the number of samples in the subsets, $|T|$ is the total number of samples, and $Ent(T^v)$ is the information entropy calculated from subsets $|T^v|$.

3) Calculating gain rate

$$Gain\_ratio(T,a) = \frac{Gain(T,a)}{IV(a)} \tag{8}$$

Where,

$$IV(\mathrm{a}) = -\sum_{v=1}^{V} \frac{|T^v|}{|T|} \log_2 \frac{|T^v|}{|T|} \tag{9}$$

A decision tree is constructed for each training subset separately, and when selecting the optimal partition attribute, the attribute with the largest gain rate is selected. Since the training and classification processes of each decision tree are independent of each other, the construction and classification processes of the decision tree can be parallelized so as to save program running time.

In order to determine the weight of each decision tree, WRF divides the samples into training sets and test sets at a ratio of 3:1. The out-of-bag data is employed to estimate the accuracy of the decision tree, and to calculate the voting weight w_j of the decision tree. In the implementation process of WRF, the training set includes 75% of the initial samples. For each decision tree, about 50% of the in-bag data is used to train the decision tree, and 25% of the data is used to evaluate the classification performance of the decision tree and to calculate the voting weight.

$$w_j = \frac{X_j^{correct}}{X_n}, j = 1, 2, \mathrm{L}\ , n \tag{10}$$

Where $X_j^{correct}$ is the number of out-of-bag samples with the correct classification, and X_n is the total number of out-of-bag samples.

After the weight of decision tree is calculated on the training set, the estimation model is applied on the test set. For each test sample, the output sequence $\{h_1(X), h_2(X), \mathrm{L}\ , h_n(X)\}$ of n decision tree classifiers can be obtained, and the weight w_j of decision tree is used to construct a combined classification model.

$$H(x) = \arg\max_{Y} \sum_{i=1}^{n} w_j \cdot I(h_i(x) = Y) \tag{11}$$

Where $H(x)$ is the combined classification model, $h_i(x)$ is the i-th decision tree classification model, Y is the output variable, which is the value of link quality level, and $I(\cdot)$ is the indicative function.

6.2. Evaluation Metrics of models

Accuracy is often used as the evaluation metric of link quality estimation models. But for models with imbalanced samples, accuracy does not well reflect the performance of them. For example, there is a test set of 1000 samples including 100 negative samples. If a model classifies all samples into positive, then the accuracy of the model is 90%. From the view of accuracy, it can be seen that the estimation effect of the model is very good, but the model does not identify even one negative sample. So such model has no meaning. Therefore, this paper uses the precision, recall and F1 values to evaluate the performance of proposed model.

For the two-class classification problem, the combination of the real class of the sample and the predicted class of the classifier has four cases: true positive, false positive, true negative and false negative. The confusion matrix formed by the classification results is shown in Table 2.

Table 2. Confusion Matrix**.**

Real class	Predict result	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Confusion matrix can be used to evaluate the precision and recall of classifiers.

$$precision = \frac{TP}{TP + FP} \tag{12}$$

$$recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{14}$$

7. Experiments and Analysis

In the experiments, TelosB nodes of Crossbow Company and the wireless sensor network link quality test platform were used to collect the required data, and K-means SMOTE algorithm and WRF estimation model were implemented through python platform.

7.1. Experimental Scene Setting

Three experimental scenarios are set up, namely, forest, indoor and campus parking. In each experimental scenario, a small star link quality testing network is deployed to collect corresponding link quality data. The experimental parameter settings are shown in Table 3.

Table 3. Experimental Parameter Setting.

Parameter attribute	Parameter value
Transmit power /dBm	31
Channel	26
Number of probe packets	30
Packet transmission rate	5
Transmission period/ms	200
Test period /s	10

7.2. Analysis of experimental results

In the experiments, the training set and the test set are randomly distributed at a ratio of 3:1. For the data in the training set, K-means SMOTE is used to balance samples. WRF is used to construct a link quality estimation model for the original samples and the samples processed by K-means SMOTE respectively. The evaluation confusion matrix diagrams under different scenarios are shown in Figure 2 to Figure 7.

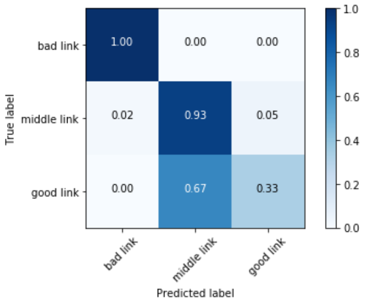

Fig. 2. Confusion matrix of original samples in parking scenes.

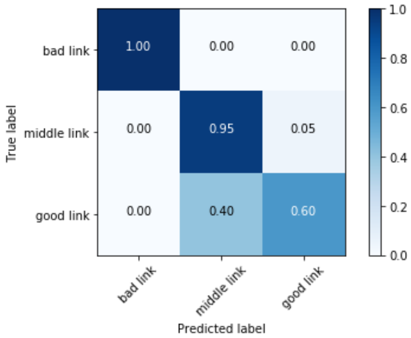

Fig. 3. Confusion matrix of K-means SMOTE processing samples in parking scenes.

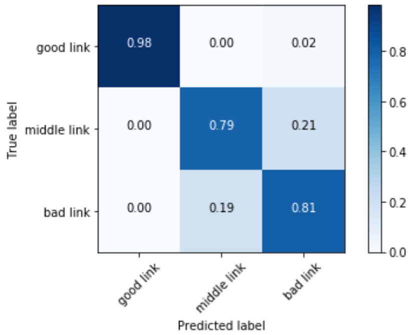

Fig. 4. Confusion matrix of original samples in indoor scenes.

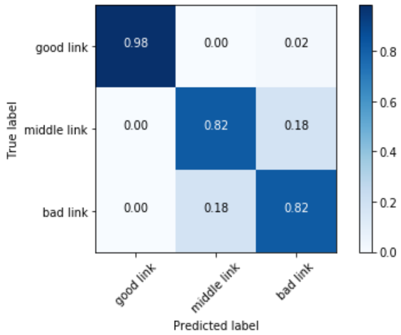

Fig. 5. Confusion matrix of K-means SMOTE processing samples in indoor scenes.

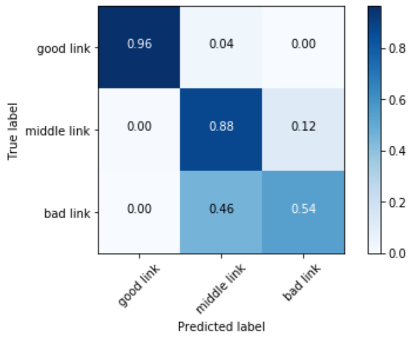

Fig. 6. Confusion matrix of original samples in forest scenes.

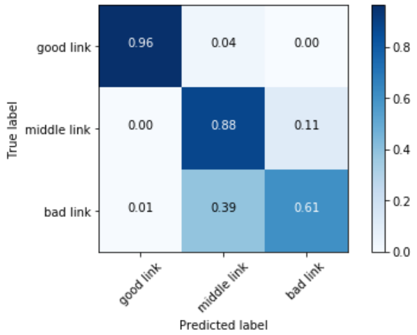

Fig. 7. Confusion matrix of K-means SMOTE processing samples in forest scenes.

It can be seen from Figure 2 and Figure 3 that the estimation accuracy of good link in the parking scene is 33% for the original samples, 60% after K-means SMOTE processing, and the estimation accuracy is increased by 27%. In the indoor scene, the estimation accuracy for middle link has increased by 3%, and in the forest scene, the estimation accuracy for bad link has increased by 7%. The estimation results of three experimental scenes show that the data processed by K-means SMOTE can obviously improve the estimation effect of the model on minority class samples.

In order to further reflect the estimation effect of the model, we calculate the precision, recall and FI value of the original samples and the samples processed by K-means SMOTE under different scenes respectively.

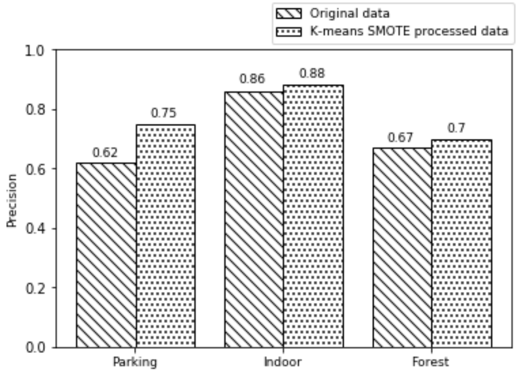

Fig. 8. Comparison of precision.

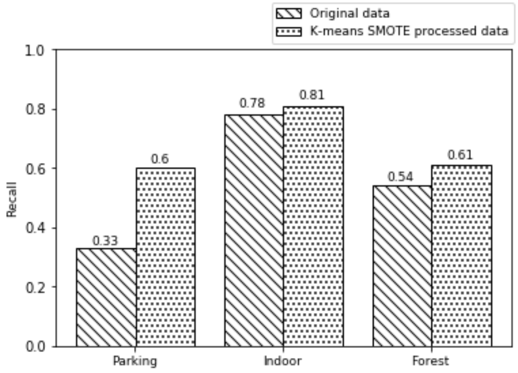

Fig. 9. Comparison of recall.

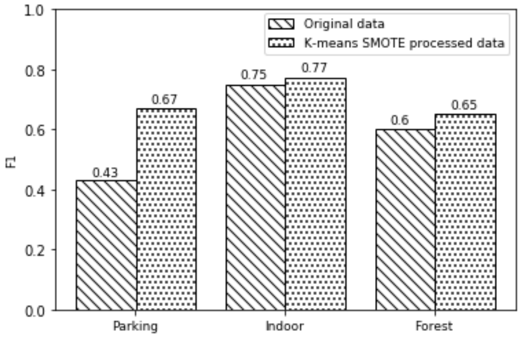

Fig. 10. Comparison of F1.

Figure 8 to Figure 10 show the comparison of precision, recall and F1 of the original data and K-means SMOTE processed data under different scenes respectively. As can be seen from the Figures, in the parking scene, the precision, recall and F1 value of the data processed by K-means SMOTE have increased by 13%, 27% and 24%, respectively, with the most obvious performance improvement. In other scenes, the evaluation metrics of K-means SMOTE with the data processed are improved compared with ones of the model with the original data, indicating that K-means SMOTE has good performance in dealing with imbalanced data.

In order to further verify the estimation performance of WRF, this paper uses the data, processed by K-means SMOTE to train WRF model, and compares the trained WRF model with NB, LR and KNN models in three experimental scenarios.

The comparison results of precision, recall and F1 in each scenario are shown in Figure 11 to Figure 13.

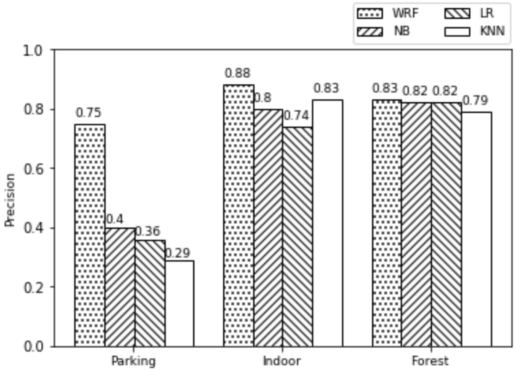

Fig. 11. Precision comparison result 1.

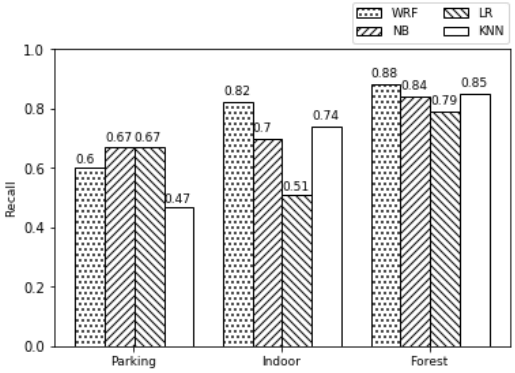

Fig. 12. Recall comparison result 1.

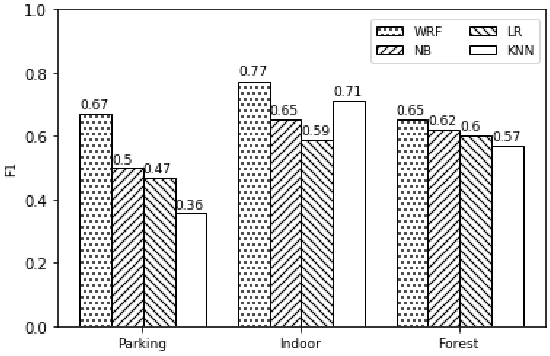

Fig. 13. F1 comparison result 1.

As can be seen from Figure 11 to Figure 13, the precision of the WRF model is 35% bigger than that of the LR model in the parking scene, and the precision and F1 of the WRF model in the three scenes are bigger than those of the other three models. The recall values in the indoor scene and the forest scene are also bigger than that of the NB, LR and KNN models. The experimental results show that the WRF has good estimation performance in different experimental scenes.

8. Conclusion

The main work of this paper is as follows: The average, variance and asymmetry metric of physical layer parameters are selected as link quality parameters, and link quality level are divided according to PRR values to estimate link quality. K-means SMOTE is applied to generate minority samples, and the samples are divided into different clusters by K-means, thus noise data can be avoided. For different clusters, minority class samples are generated by random linear interpolation in each cluster, so that the number of minority class samples is increased, thus solving the imbalance of link data. This paper constructs a link quality estimation model based on WRF, assigns smaller weights to decision trees with low classification performance and bigger weights to decision trees with high classification performance in the combined model, and uses out-of-bag data to evaluate the accuracy of the decision trees. Precision, recall and F1 are used to evaluate the performance of proposed model. The experimental results in three scenarios show that the proposed model with samples processed by K-means SMOTE presents better performance. Compared with NB, LR and KNN estimation models, WRF model has better estimation performance.

Acknowledgements: This work was supported by the National Natural Science Foundation of China (Grant No. 61962037, 61762065, 61363015, 61501218), the Natural Science Foundation of Jiangxi Province (Grant No. 20171ACB20018, 20171BAB202009, 20181BAB202015).

References

1. Buşoniu, L., Babuška, R., Schutter, B. D.: Multi-agent Reinforcement Learning: An Overview. Innovations in Multi-Agent Systems and Applications, Vol. 38, No. 2, 156-172. (2010)
2. Babuška, R., buşoniu, L., and De Schutter, B.: Reinforcement learning for multi-agent systems. IEEE International Conference on Emerging Technologies and Factory Automation. IEEE, 1-7. (2006)
3. Liu, Z., Wang, F.: Scale-free topology for wireless sensor networks with energy efficient characteristics, Journal of Beijing University of Posts and Telecommunications, Vol. 38, No. 1, 87-91. (2015)
4. Cao, N., Liu, P., Li, G., Zhang, C.: Evaluation models for the nearest closer routing protocol in wireless sensor networks, IEEE Access, Vol. 6, No. 1, 77043-77054. (2018)
5. Gao, D., Zhang, S., Zhang, F.: RowBee: A routing protocol based on cross-technology communication for energy-harvesting wireless sensor networks, IEEE Access, Vol. 7, No. 1, 40663-40673. (2019)
6. Lowrance, C. J., Lauf, A. P.: Link quality estimation in ad hoc and mesh networks: A survey and future directions, Wireless Personal Communications, Vol. 96, No. 1, 475-508. (2017)
7. Bote-Lorenzo, M. L., Gómez-Sánchez, E., Mediavilla-Pastor, C.: Online machine learning algorithms to predict link quality in community wireless mesh networks, Computer Networks, Vol. 132, No. 1, 68-80. (2018)
8. Lu, J., Zhu, Y., Xu, Z.: A reliable wireless sensor network routing method for power transmission line monitoring, Power System Technology, Vol. 41, No. 2, 644-650. (2017).
9. Jayasri, T., Hemalatha, M.: Link quality estimation for adaptive data streaming in WSN, Wireless Personal Communications, Vol. 94, No. 3, 1543-1562. (2017)
10. Baccour, N., Koubâa, A., Youssef, H.: F-lqe: A fuzzy link quality estimator for wireless sensor networks, in Proc. 2010 European Conference on Wireless Sensor Networks, Coimbra, Portugal, 240-255. (2010)
11. Lai, X., Ji, X., Zhou, X.: Energy efficient link-delay aware routing in wireless sensor networks, IEEE Sensors Journal, Vol. 18, No. 2, 837-848. (2018)
12. Boano, C. A., Zúniga, M. A., Voigt, T.: The triangle metric: Fast link quality estimation for mobile wireless sensor networks, in Proc. 19th International Conference on Computer Communications and Networks, Zurich, Switzerland, 1-7.(2010)
13. Zhang, Y., Fu, S., Jiang, Y.: An LQI-based packet loss rate model for IEEE 802.15.4 links, in Proc. IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communications(PIMRC), Bologna, Italy, 1-7.(2018)
14. Sun, W., Lu, W., Li, Q.: WNN-LQE: Wavelet-neural-network-based link quality estimation for smart grid WSNs, IEEE Access, Vol. 5, No. 1, 12788-12797. (2017)
15. Mi, X., Zhao, H., Zhu, J.: Research on EWMA based link quality evaluation algorithm for WSN, in Proc. 2011 Cross Strait Quad-Regional Radio Science and Wireless Technology Conference, Harbin, China, 757-759.(2011)
16. Qin, F., Zhang, Q., Zhang, W.: Link quality estimation in industrial temporal fading channel with augmented Kalman filter, IEEE Transactions on Industrial Informatics, Vol. 15, No. 4, 1936-1946. (2019)
17. Rojas, C., Decotignie, J.: Leveraging MAC preambles for an efficient link estimation, in Proc. International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Limassol, Cyprus, 1-10.(2018)
18. Sharma, A., Bansal, A., Rishiwal, V.: Selection of high quality path through MXAODV in mobile ad-hoc network, International Journal of Systems Control and Communications, Vol. 10, No. 1, 1-17. (2019)

19. Liu, W., Xia, Y., Luo, R.: Lightweight, fluctuation insensitive multi-parameter fusion link quality estimation for wireless sensor networks, IEEE ACCESS, Vol. 8, No. 1, 28496-28511. (2020)
20. Liu, T., Cerpa, A. E.: Data-driven link quality prediction using link features, ACM Trans on Sensor Networks, Vol. 10, No. 2, 1-35. (2014)
21. Shu, J., Liu, S., Liu, L.: Research on link quality estimation mechanism for wireless sensor networks based on support vector machine, Chinese Journal of Electronics, Vol. 26, No. 2, 377-384. (2017)
22. Pan, L., Li, J.: K-nearest neighbor based missing data estimation algorithm in wireless sensor networks, Wireless Sensor Network, Vol. 2, No. 2, 115-122. (2010)
23. Luo, X., Liu, L., Shu, J., AL-KALI, M.: Link quality estimation method for wireless sensor networks based on stacked autoencoder, IEEE Access, Vol. 7, No. 1, 21572-21583. (2019)
24. Baccour, N., Koubâa, A., Youssef, H.: Reliable link quality estimation in low-power wireless networks and its impact on tree-routing, Ad Hoc Networks, Vol. 27, No. 1, 1-25. (2015)
25. Bildea, A., Alphand, O., Rousseau, F., Duda, A.: Link quality estimation with Gilbert-Elliot model for wireless sensor networks, in Proc. IEEE 26th Annual Int. Symp. Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong, China, 2049-2054.(2015)
26. Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., Ning, G.: Class weights random forest algorithm for processing class imbalanced medical data, IEEE Access, Vol. 6, No. 1, 4641-4652. (2018)
27. Galar, M., Fernandez, A., Barrenechea, E.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches, IEEE Trans on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 42, No. 4, 463-484. (2012)
28. Batista, G. E., Prati, R. C., Monard, M. C.: A study of the behavior of several methods for balancing machine learning training data , ACM SIGKDD explorations newsletter, Vol. 6, No. 1, 20-29. (2004)
29. Chawla, N. V., Bowyer, K. W., Hall, L. O.: SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, Vol. 16, No. 1, 321-357. (2002)
30. Douzas, G., Bacao, F., Last, F.: Improving imbalanced learning through a heuristic oversampling method based on K-means and SMOTE, Information Sciences, Vol. 465, No. 1, 1-20. (2018)
31. Liaw, A., Wiener, M.: Classification and regression by random forest, R News, Vol. 2, No. 3, 18-22. (2002)
32. Winham, S. J., Freimuth, R. R., Biernacka, J. M.: A weighted random forests approach to improve predictive performance, Statistical Analysis and Data Mining: The ASA Data Science Journal, Vol. 6, No. 6, 496-505. (2013)

Linlan Liu born in 1968, and received the Bachelor degree in computer science from the National University of Defense Technology, Changsha, China, in 1988. Currently she is a full Professor, Internet of Things Technology Institute, Nanchang Hangkong University, Nanchang, China. She was a Visiting Scholar at Wilfrid Laurier University, Waterloo, Ontario, Canada. She has authored/coauthored more than 70 papers. Her research interests include wireless sensor networks and embedded system (765693987@qq.com).

Yi Feng born in 1995. She received the Master degree from Nanchang Hangkong University, Nanchang, China, in 2020. She is now with the Department of School of

Engineering, Zhejiang Normal University Xingzhi College, Jinhua, China. Her research interests include wireless sensor networks. (458018002@qq.com).

Shengrong Gao born in 1994. He received the Master degree from Nanchang Hangkong University, Nanchang, China, in 2019. His research interests include wireless sensor networks. (1322415547@qq.com).

Jian Shu born in 1964. He received the M.Sc.degree in computer networks from Northwestern Polytechnical University. He is currently a Professor with Nanchang Hangkong University. His research interests include wireless sensor networks, embedded systems, and software engineering. (shujian@nchu.edu.cn).

Received: December 18, 2020; Accepted: May 22, 2021.