IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 11, 2020, accepted December 21, 2020, date of publication December 28, 2020, date of current version January 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047648

A Link Quality Estimation Method for Wireless Sensor Networks Based on Deep Forest

MU HE AND JIAN SHU
School of Software, Institute of Internet of Things and Big Data Technology, Nanchang Hangkong University, Nanchang 330063, China
Corresponding author: Jian Shu (shujian@nchu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61962037, Grant 62062050, and Grant 61762065; in part by the Natural Science Foundation of Jiangxi Province under Grant 20202BABL202039 and Grant 20181BAB202015; and in part by the Innovation Foundation for Postgraduate Student of Jiangxi Province under Grant YC2019090.

ABSTRACT In wireless sensor networks, sensor nodes, the miniature embedded devices, have limitation of energy, storage, computing, and etc. One of the tasks of the nodes is to use their limited resources to complete work efficiently. Choosing high quality link communication can effectively save energy. In this paper, we propose a link quality estimation model that is based on deep forest. To avoid a noise sample becoming a center point in the clustering, we use an improved K-medoids algorithm based on step increasing and optimizing medoids (INCK) when dividing the link quality grades. During the sample preprocessing stage, the Pauta criterion is used to delete the noise link samples, and we fill the mean value of each grade into the missing values. The feature extraction performance of deep forest is improved by combining the stratified sampling to change the unbalance distribution of link quality samples. And then the Stratified Sampling Cascade Forest link quality estimation (SCForest-LQE) is constructed by combining stratified sampling with cascade forest. The experiments are conducted in three real application scenarios. Compared with the existing six link quality estimation models, SCForest-LQE has better estimation performance and stability.

INDEX TERMS Wireless sensor networks, link quality estimation, deep forest, stratified sampling.

I. INTRODUCTION

Wireless sensor networks (WSNs) are formed by self-organizing multiple sensor nodes through wireless communication technology. The nodes are organized into a sensor network in a certain way to cover the monitoring scope, and the sensor nodes calculate the collected information and transmit it to the sink node. In WSNs, the quality of the link communication can reflect the real state of the link, and the selection of high quality for wireless link communication can avoid the energy consumption of rerouting and data retransmission caused by the influence of the environment on the sensor nodes. It can improve the reliability of the network protocol and algorithm by applying it to the actual industrial Internet of things, agricultural monitoring, location perception and so on. The nodes are distributed in a complex and diverse environment, for example, a grove where there are stones and trees blocking the nodes, an office covered by Bluetooth and Wi-Fi, a road with a flow of people and vehicles, and so on. These barriers will lead to multipath weakness and path loss of radio signal propagation, which will result in fluctuations in the link communication quality. In addition, climate factors will also affect node communications. Weather changes have a negative correlation on the signal strength, especially the effect of temperature and humidity on the Packet Received Rate (PRR) and Received Signal Strength Indicator (RSSI) [1]. A good link quality estimation (LQE) model can be built to accurately evaluate the current link status, which can improve the packet transmission efficiency, decrease the energy consumption of sensor nodes, and prolong the network service life. The link quality distribution is defined as three regions: the connected region, the transitional region and the disconnected region [2]. The link quality in the transitional region is not stable in the long-term process of link communication, and it is easy for it to be disturbed and generate dynamic fluctuations, which increases the difficulty of accurate evaluation. The range of the transitional region depends on the external environment and the characteristics of the radio hardware. In WSNs, the link quality distribution usually presents an unbalanced state, and the links in the transitional region account for more than half of the total number of links [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Xianfu Lei.

In this paper, a link quality estimation model based on deep forest is proposed, so as to reduce the impact of an unbalanced link quality distribution on the estimator and improve the accuracy of the estimation. The major contributions of this article are summarized as follows:

(1) The improved K-medoids algorithm based on step increasing and optimizing medoids (INCK) is utilized to divide the link quality grades. The INCK optimizes the clustering performance of PRR, so that the division accuracy of link quality grades is improved.

(2) A Link quality estimation model based on improved deep forest algorithm (stratified sampling cascade forest, SCForest) is proposed. The stratified sampling of SCForest can reduce the impact of unbalanced data on the estimator, the accuracy of link quality estimation is further improved.

(3) To verify the performance of the proposed LQE method, we carry out experiments in three real campus scenarios and the results show that our method is more accurate and stable in comparison with other LQE methods.

The remainder of this paper is structured as follows. Section II introduces the related research on link quality estimation and deep forest. Section III addresses the method of dividing the link quality grades. The link quality estimation approach for WSNs is proposed in section IV. Section V verifies the performance of the proposed method. We summarize the work in section VI.

II. RELATED WORK

After the development of sensor networks theory and technology, although it has been widely used in all walks of life, but most applications are still limited to a small scale. Accurate estimation of the link quality and reasonable allocation of the resources are still key issues that limit the development of WSNs in large areas. Complex interference sources in the transmission scenario will also reduce the wireless signal strength. This concern makes obtaining an accurate estimation of the link quality more challenging, attracting extensive attention and in-depth research from domestic and foreign scholars. Currently, research on the link quality estimation can fall into link parameter-based, link characteristics-based and machine learning-based methods. The link quality estimation model proposed in this paper belongs to machine learning-based methods.

In early LQE studies, scholars use physical layer parameters to estimate the link quality, such as RSSI, signal-to-noise ratio (SNR) and link quality indicator (LQI)[4]–[6]. However, a wireless channel has a time-varying nature, and hence, the relationship between the link parameters is not suitable for all scenarios. In a dynamic environment, the reliability of estimation models based on link parameters will be reduced. The existing estimation methods based on link characteristics mainly analyze the characteristics of the physical layer and the data link layer in WSNs. The scholars apply the complex model theory and link characteristics to the study of LQEs [7], [8]. They periodically analyze the efficiency of data packet transmission in the wireless channels. This approach uses an approximate mapping function to map the link parameters to the link quality, and then, it obtains the link quality change trend. According to this strategy, the spatial, temporal and asymmetry characteristics of the link are addressed, and thus the link quality estimation model based on link characteristics can be applied in many dynamically changing network environments.

The above research provides relevant experience for current studies on link quality estimation. In recent years, some scholars have discovered the potential of applying machine learning to link quality estimation [9]. On the one hand, machine learning-based methods can find the best overall correlation between the PRR and physical layer parameters by analyzing the error between the packets sent amount and the actual received amount. Liu and Cerpa [10], [11] trains Bayesian, neural network, and logistic regression algorithms to establish a link quality evaluation model. They propose 4C to calculate the probability of the next packet being successfully transmitted. Sun *et al.* [12] uses a wavelet neural network to estimate the link quality, and it provide a guarantee for the development of the WSN routing protocol by analyzing whether link quality meet the communications standard. Liu *et al.* [13] uses the lightweight weighted Euclidean distance to fuse SNR and LQI. Then, the link quality estimation model is constructed by logistic regression to reflect the link quality. Xue *et al.* [14] decomposes the raw SNR sequence into the time-varying sequence and stochastic sequence. A random-vector-functional-link-based algorithm is used to predict the two sequences separately. On the other hand, the hierarchical structure of the deep learning network can enhance its learning efficiency on link quality data. There are many different deep learning networks, like dynamic convolution neural network (DCNN) [15], spatiotemporal attention-based long-short-term-memory (STA-LSTM) [16] and layer-wise data augmentation-based stacked autoencoder (LWDA-SAE) [17]. With the characteristics of layer-by-layer processing, deep learning is able to extract features of higher levels, which is of benefit to feature learning of link quality. For example, Sun *et al.* [18] uses LSTM to determine the link reliability confidence interval, which is used to express the link quality in worst case. In addition, in some link quality estimation studies, such as those in [19] and [20], they turn the link quality estimation problem into a classification problem. Machine learning can help the link quality estimation model to continuously adapt to changes in the network environment, and to reduce the impact of interference on the link quality estimation through a feature processing technique. People using machine learning to solve the problem of link quality estimation is a development trend of future research.

Deep forest is also known as the multi-grained cascade forest (gcForest)[21]. The learning performance of traditional deep network algorithms depend on careful hyperparameter adjustment. Furthermore, before training the networks, its architecture has to be determined, and thus, the model complexity is determined in advance. Different from traditional neural networks, the performance of gcForest is quite robust to hyperparameter settings. Moreover, gcForest can automatically determine the number of cascading layers, and adjust the model complexity through data-related methods to make it adaptable to different datasets. Compared with the deep neural networks, gcForest has fewer hyperparameters and it is easier to train, which can control the training cost of model according to the available computational resources. The multi-grained scanning of gcForest is powerful in processing the feature relationships in image data, and then gcForest builds a deep model by introducing a cascade structure to enhance its feature learning ability. In addition, due to the characteristics of decision trees, the cascade forest of gcForest also has good performance in the multiclassification of nonimage fields. At present, good results have been achieved in the application of malicious code classification [22], anomaly detection [23], and multi-instance learning [24].

In this paper, on the basis of studying the relationships of the link parameters, we use the INCK algorithm to divide the link quality grade and use the grades as the classification label for the link quality. The Pauta criterion is used to remove the highly abnormal data from the samples and we fill the mean value of each grade into the missing values. At the same time, considering that the unbalanced data set makes the models less accurate when estimating the minority samples, the majority samples will dominate the final accuracy[25]. By combining the link quality distribution regions, we use the stratified sampling method to process the samples of each grade, so as to optimize the data training stage of deep forests. During the construction of the SCForest estimation model (SCForest-LQE), the unbalanced distribution of the original link quality samples is improved to reduce the impact of the unbalanced nature of the dataset on the performance of the SCForest-LQE.

III. DIVISION OF LINK QUALITY GRADE

Selecting a high-quality link for communication can improve the transmission efficiency of the data packets. Link quality estimation provides a basis for the route to select communication links. Using the link quality grade rather than specific scores as the link quality estimation index will help to simplify the process of routing. At the same time, the link quality grade can be used as the classification index of the estimation model. In experimental scenarios, a node periodically sends probe packets to calculate the PRR, and the link quality is directly reflected in the successful reception rate of the probe packets. We take PRR as the measurement of the link quality grade, and we set the division level to the category of the SCForest-LQE classification. In the existing research, the divisions of the link quality grades are determined by a subjective decision. The clustering algorithm divides the data into different clusters based on similarity. Compared with a subjective partitioning method, a clustering algorithm is more accurate, but it is sensitive to initial medoids. Due to interference from the environmental, the collected link quality samples have noise and isolated values, which will affect the performance of clustering algorithm. We use the INCK algorithm[26] to divide the link quality grades. This approach obtains a subset of candidate medoids based on the deviations of the PRR samples, which can avoid an isolated point becoming a medoids. Based on this principle, the overall deviations state of the PRR samples can be measured according to the variance σ of the sample set. The variance σ is shown in Equation (1).

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1}$$

where $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is the mean of all objects in the dataset.

Because noise data is usually far away from the central region, there is a large deviation between the noise sample and other PRR samples. The variance σ_i of one PRR is defined as

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{1} dist(x_i, x_j)^2} \tag{2}$$

where $dist(x_i, x_j)$ is expressed as the distance between object x_i and object x_j.

According to the definition of the variance σ and variance σ_i, the candidate medoids subset S of PRR can be defined as in Equation (3). The stretch factor λ excludes link outliers from clustering effects. Here, λ is selected based on experience.

$$S = \{x_i | \sigma_i \leq \lambda\sigma, i = 1, \ldots, n\} \tag{3}$$

The selection of clustering medoids will have a significant impact on the final clustering effect. If the medoids are randomly selected in a cluster, the algorithm could fall into a local optimum. Hence, we use the INCK method, which optimizes the K-medoids clustering methods by increasing the number of link quality clusters from 2 to the expected number in a stepwise fashion. To obtain the medoids of the PRR cluster, two initial medoids are selected first. It is assumed that the first initial medoid o_1 is located in the largest density region of S, then we can obtain the first medoid o_1 that satisfies Equation (4).

$$o_1 = \arg\min_{x_i \in S_m} \{d_i \,| i = 1, \ldots, n\} \tag{4}$$

where d_i is the total distance from object x_i to all objects. To ensure that the boundary samples in the real dataset are divided into different link quality grades, the second initial medoid o_2 is the point in S that has the largest distance from o_1, as shown in Equation (5). We distribute the remaining objects in the cluster in which they are closest to the medoids.

$$o_2 = \arg\max_{x_i \in S} \{dist(x_i, o_1) \,| i = 1, \ldots, n\} \tag{5}$$

For the subsequent growth of the medoids, a point farthest from the initial center point o_α in each existing cluster group should be selected as the candidate medoid o'_α. Choosing the

Algorithm 1 Link Quality Grades Division Algorithm

Input: PRR data set: $X_{PRR} = [x_1, x_2, \ldots, x_n]$, the number of link quality grades: G, the stretch factor: λ.
Output: PRR medoids: $o = [o_1, o_2, \cdots, o_G]$, the range of PRR clusters: $C = \{C_1, C_2, \ldots, C_G\}$.
Process:

Get σ according to Equation (1).
Get σ_i according to Equation (2).
if $\sigma_i < \lambda\sigma$ then
  Add x_i to S.
end if
Compute the distance of PRR samples.
Get o_1 according to Equation (4).
Get o_2 according to Equation (5).
Assign the remaining PRR samples to cluster C_1 or C_2.
while $k \leq G$ do
  Compute the new medoid $o'_\alpha = \max(dist(o'_\alpha, o_\alpha))$.
  o.append(o'_α).
  Assign the remaining PRR samples to cluster $C_1, C_2 \cdots, C_k$.
  $k = k + 1$.
end while
Compute the cost $E = \sum_{G=1}^{G} \sum_{x \in C_G} dist(o_G, x)^2$.
Update o to make E become minimum.
return C.

pair of o_α and o'_α with the farthest distance, makes the o'_α become new medoid.

In the INCK algorithm, the medoids updating method for the PRR clusters is similar to that of FastK[27], and thus it can maintain the calculation efficiency. The process of dividing link quality grades base on INCK is shown in algorithm 1.

IV. LINK QUALITY ESTIMATION

The link quality data set is obtained from the office, parking lot and road scenarios on campus. Multiple link parameters are selected as the input features of the link quality estimation model. Some noise data exist in the data set due to the interference in the experimental scenarios. Therefore, the Pauta criterion is used to remove the noise data from the sample set. We use the mean value of each grade to fill in the missing values. Stratified sampling is used to improve the unbalanced distribution of the original link samples, and the preprocessed link samples are used as the input of SCForest-LQE. The link quality grades are classified as a category of cascading forest output to assess the link quality of the current period.

A. SELECTION OF LINK QUALITY PARAMETERS

Physical layer parameters can respond quickly to the link quality, they are easy to read and have low overhead in the CC2420 nodes, and there is a correlation with the link layer parameter PRR. Multi-index link quality parameters can reflect the link quality more comprehensively and are not easily affected by the deployment environment noise or by multipath effects during the link quality estimation. The mean value of LQI, RSSI, SNR (LQI_{mean}, $RSSI_{mean}$, SNR_{mean}) and the coefficient of variation of LQI, RSSI, SNR (CV_{LQI}, CV_{RSSI}, CV_{SNR}) are selected as the features of the SCForest-LQE. They can be obtained directly by a simple calculation through the relevant registers of CC2420 nodes. LQI is used to represent the quality of the received packets; it provides the wireless signal strength and quality information for the network layer when receiving data frames. RSSI represents an indication of the strength of the received signal. CV represents the link quality stability of the three hardware parameters. CV_{LQI} is defined as in Equation (6).

$$CV_{LQI} = \frac{\sigma(LQI)}{\mu(LQI)} \tag{6}$$

where $\sigma(LQI)$ is the standard deviation of a period, and $\mu(LQI)$ is the mean of a period. The formulas of CV_{RSSI} and CV_{SNR} are the same as CV_{LQI}.

B. SAMPLE PREPROCESSING

Since the interference in the wireless sensor networks affects the link communication between the nodes, there will be noise in the samples obtained in the experimental scenarios. During the characteristic engineering stage, noise samples that deviate far from the normal link samples should be removed to reduce the influence of the noise samples on the model training efficiency. We use the Pauta criterion to define the scope of the training samples. The Pauta criterion states that the values of the samples are almost entirely within 3σ, then, the highly abnormal samples, which exceed 3σ are deleted.

Due to the interference of the experimental scenarios, data entry errors and hardware damage, it is easy to have incomplete data collection. The link quality samples with missing data will affect the accuracy of the model [28]. Some hardware parameter information that is lost within a probe period of a node can be processed by filling in the missing data. When a row of data in a feature of the link quality sample set is missing, the mean value of each grade is used to fill in the missing samples. Link quality samples preprocessing is shown in algorithm 2.

C. STRATIFIED SAMPLING OF SCFOREST

The link quality belongs to the text structure. We take the features $[LQI_{mean}, RSSI_{mean}, SNR_{mean}, CV_{LQI}, CV_{RSSI}, CV_{SNR}]$ as the raw input vector of SCForest-LQE. A multi-grained scanning method is not conducive to feature extraction of the link quality, because the important features at the beginning and the end of the dataset can easily to be ignored by using this method. The characteristic of link asymmetry has a large impact on the distribution of the link quality data, with most of them are located in the transitional region. The existing estimators based on machine learning can easily to be influenced by the imbalanced data. To solve this problem, the feature scanning method of gcForest is optimized in this paper. Based on the distribution area of the link quality,

Algorithm 2 Link Quality Samples Preprocessing

Input: Link quality samples $D_{LQ} = [LQI_{mean}, RSSI_{mean}, SNR_{mean}, CV_{LQI}, CV_{RSSI}, CV_{SNR}], label.$
Output: Processed samples $D'_{LQ} = [LQI'_{mean}, RSSI'_{mean}, SNR'_{mean}, CV'_{LQI}, CV'_{RSSI}, CV'_{SNR}]$
Process:

Get $\sigma_{LQ}[n]$ according to Equation (1), $n = 1,2,3,4,5,6$.
Compute the $amount = len(D_{LQ})$.
while $j \leq amount$ do
  if $\sigma_{LQ}[i, n] \geq 3\sigma_{LQ}[n]$ then
    delete $D_{LQ}[i]$
  end if
  $i = i+1$
end while
while $j \leq amount$ do
  Compute the mean of a label in D_{LQ}: $mean_{label}[n]$.
  if one of the parameters in $D_{LQ}[j, n]$ is NULL then
    Use $mean_{label}[n]$ fill in $D_{LQ}[j, n]$.
  end if
  $j = j+1$.
end while
return D'_{LQ}.

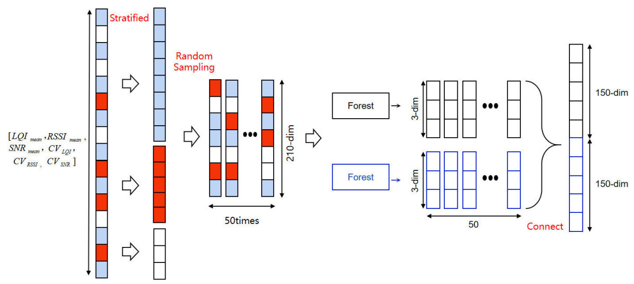

FIGURE 1. Principle of stratified sampling.

we use stratified sampling to stratify the link quality samples, and the features of each of the link quality distribution regions are extracted evenly.

If the sample numbers of the three regions obtained from an experimental scenario are unbalanced, then multi-grained scanning will lead to model training bias. Therefore, we divide the three regions samples according to the asymmetry characteristics of the link. Asymmetrical links are mainly located in the transitional region of medium link quality, while connected and disconnected regions with high or very low PRR tend to be symmetrical [2]. Therefore, according to the PRR labels that are divided by INCK, the medium grade link samples with higher asymmetry can be set to be the transitional region samples φ_t. The good grade samples are set to be the connected region samples φ_c. The bad grade samples are set to be the unconnected region sample φ_u.

We obtain three layers of feature vectors at the end. The principle of the stratified sampling of the SCForest-LQE is shown in Figure.1. According to the proportion of each layer, the feature vectors of each layer are randomly extracted to form multiple sets of sequence data. Three layers of feature vectors extracted from the stratified sampling will be treated as positive/negative instances, which will be used for training a completely random tree forest and a random forest, and then for generating class vectors. We connect all of the class vectors to form the transformed features to be the input of the cascade forest. For example, there are three layers of link samples. Suppose that we extract a 70-dimensional feature vector in each layer. Randomly extracting 50 times on the scale of 3×70, there are then 210-dimensional features randomly extracted from the original feature vectors each time, and we use them to form an instance; in total, 50 instances are generated. Assuming that the link quality grade is divided into 3 classes, and the instances are trained with a completely-random tree forest and a random forest, then 50 three-dimensional class vectors are produced by each forest. Since the link samples of each layer are randomly selected, the feature vectors of each link quality distribution region have the same number; thus, the samples in each link quality distribution region are trained in a balanced way using the forest.

D. LINK QUALITY ESTIMATION MODEL

SCForest-LQE consists of stratified random and cascade forest. Stratified sampling combines the characteristics of the wireless link quality distribution regions to transform the original input samples. It fuses the link characteristics to process the samples, and it enhances the estimation performance of the SCForest-LQE on a few types of connected and unconnected area link samples. Cascade forest is the integration of multiple forests, with training samples in each layer of the cascade structure. Every forest uses cross validation to reduce the risk of overfitting which can improve the accuracy of the model.

Figure.2 summarizes the overall procedure of SCForest-LQE. There are three sizes of samples that are used for stratified sampling. Assuming that the 3×40-dimensional link features are randomly sampled 100 times, the 3×70-dimensional link features are randomly sampled 50 times; and the 3×100-dimensional link features are randomly sampled 25 times, then, it can generate 75, 150 and 300-dimensional transformed feature vectors, connecting all of them as input vectors of the cascade forest.

In ensemble learning, the accuracy and diversity of individual learners can improve the performance of the overall model. Adding randomness to the training process can enhance the diversity of SCForest-LQE. Therefore, we adopt three methods to enhance the diversity of SCForest-LQE. First, we use completely random tree forests and random forests to form cascading forests, so as to encourage the diversity of internal estimators in SCForest-LQE. Second, the estimated class distribution forms a class vector, which is then concatenated with the original feature vector to be input to the next level of the cascade, which allows it to enhance the diversity of the samples. Third, because the completely-random tree forest divides the nodes by randomly selecting features until the tree grows to the leaf nodes, the method of randomly

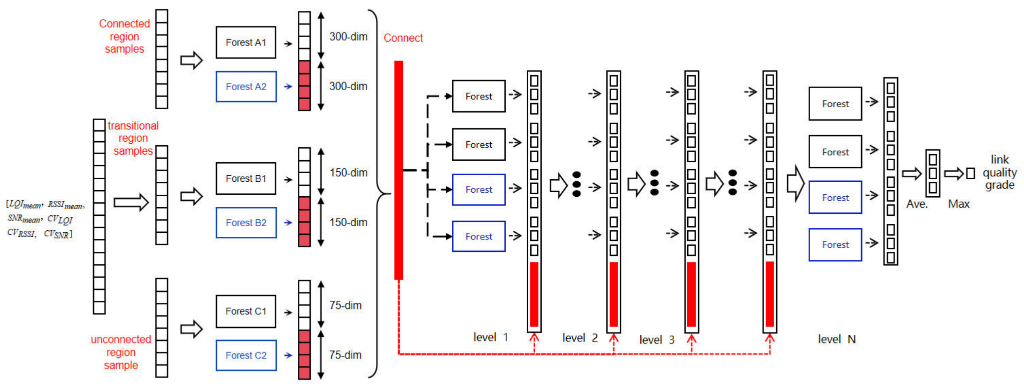

FIGURE 2. Structure of SCForest link quality estimation model.

selecting the features also leads to the different growth rules of each tree. By randomly selecting $\sqrt{d}$ features (d is the total number of input features), the random forest takes the highest Gini coefficient as the attribute value of the node splitting in each tree. The Gini coefficient calculation formula is shown in Equation (7).

$$Gini(D) = 1 - \sum_{i=1}^{G} P_i^2 \tag{7}$$

where D represents the total link quality samples, P_i refers to the probability of link quality grade i, and G represents the number of link quality grades. In cascade forest, each forest generates a link quality estimate class distribution by calculating the percentage of training samples at different link quality grades that fall on the leaf nodes of a decision tree, taking the average of all statistical results. Finally, a G-dimensional class vector is obtained.

To reduce the risk of overfitting, the link quality class vector is generated by K-fold cross-validation. In detail, we split instances into K subsamples, and keep one subsample as the test data to verify the model. Then, the other K-1 samples are used for training, with the average value of the K-1 results as the enhancement feature of the next cascade. There are two ways to terminate the training of SCForest-LQE: (1) The maximum of the cascade structure is set initially, when the cascade forest reaches its value, the cascade forest will stop training, and it will output the maximum average class vector of that layer as the link quality estimation result; (2) SCForest-LQE automatically adjusts the layer number of the cascade forest, and we set the round of stopping growth rounds. At the end of each level of training, subsamples are used to verify the estimation performance of the model. If there is no gain in the link quality estimation performance of each layer within the period (round), the training process will terminate. Therefore, SCForest-LQE can adjust its complexity and maintain the generalization of the model training automatically by adaptive expansion or reduction of its own structure according to the size of the link quality data set.

V. DESIGN AND ANALYSIS OF EXPERIMENTS

To verify the validity of the model, the link quality data are acquired from multiple application scenarios. This paper uses

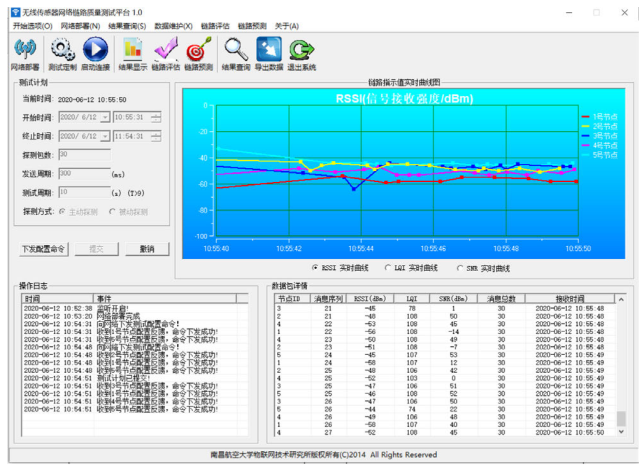

FIGURE 3. The testing platform of the wireless sensor network link quality.

the TelosB node created by Crossbow to send and receive packets, and it collects the link quality information by a wireless sensor network link quality testbed (WSNs-LQT) that was developed by the Institute of Internet of Things and Big Data Technology. WSNs-LQT is shown in Figure.3. We analyze the link quality information and implement a link quality estimation model on the PyCharm platform and Jupyter notebook.

The precision is the proportion of true positive samples among all samples which estimate as positive, as shown in Equation (8). In addition, the recall is the proportion of true positive samples among all positive samples, as shown in Equation (9). F1-score is harmonic average of the precision and recall, which can be calculated by Equation (10). In the experiment, we use the precision, recall and F1-score to evaluate the classification performance of the estimation models.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

In the case of multiclassification, we regard a link quality grade as positive, and the other grades as negative. TP is

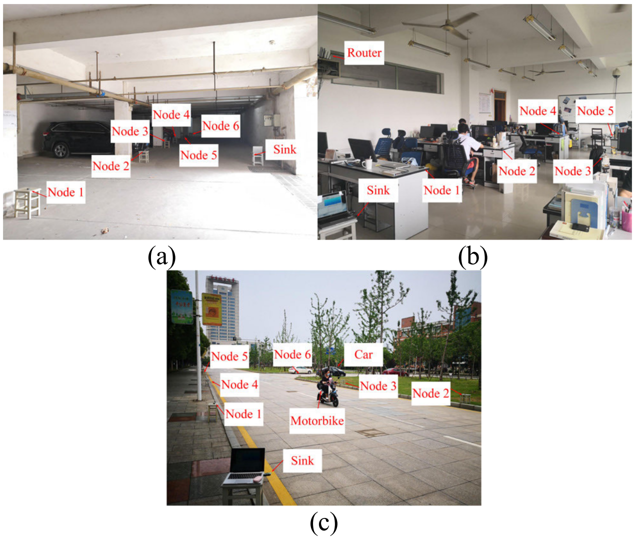

FIGURE 4. **Experimental scenarios: (a) Campus parking lot, (b) Office, (c) Campus road.**

the number of true positives, which represents the samples that are correctly classified as positive. FP is the number of false positives, which represents the negative samples that are falsely classified as positive. FN is the number of false negatives, which represents samples that belong to positive grade, but the samples are classified as negative grade on error.

A. EXPERIMENTAL SCENARIO SETTING AND DATA ANALYSIS

The selection of wireless link communication scenarios is mainly based on common application scenarios, and it considers possible interference in real WSNs environments. In this paper, we set up three experimental scenarios on campus. A parking lot scenario is designed for static obstacles (parked vehicles, load-bearing columns), an office scenario is designed for electromagnetic interference (Wi-Fi, Bluetooth), and a road scenario is designed for dynamic occlusion (moving vehicles, pedestrians). A small link quality testing network is deployed in each experimental scenario. In the campus parking lot and campus road scenarios, we deploy seven nodes, including one Sink node and six perception nodes. We arrange six nodes in the office scenarios, including one Sink node and five perception nodes. Sensor nodes are evenly spaced. The specific experimental scenario is shown in Figure.4.

Table 1 summarizes the experimental parameter settings. To ensure the diversity and reliability of the data, a series of link quality data are obtained by measuring nodes in the same time period for several consecutive days. The PRR is collected by the WSNs-LQT in three experimental scenarios, which is drawn into a time series diagram, as shown in Figures.5–7.

Due to different environmental interference sources in each scenario, the distribution of PRR presents different regulars. Figure.5 represents link quality for campus parking lot scenario. The PRR is concentrated approximately

TABLE 1. **WSNs-LQT parameter settings.**

Parameter attribute	Values
Transmission power	(-24~0)dBm
Channel	26
Number of probe packets	30
Detection method	Active detection
Packet rate	10pcs/s
Test period	10s

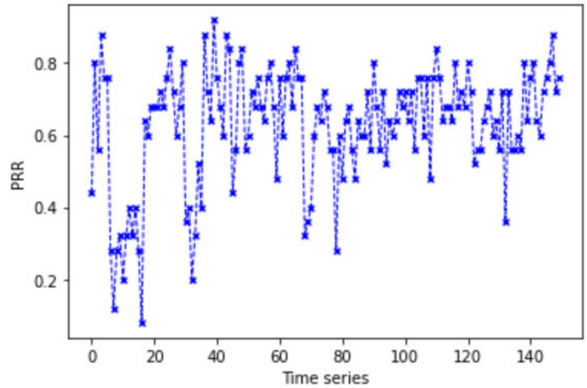

FIGURE 5. **Time series diagram of PRR in the campus parking lot scenario.**

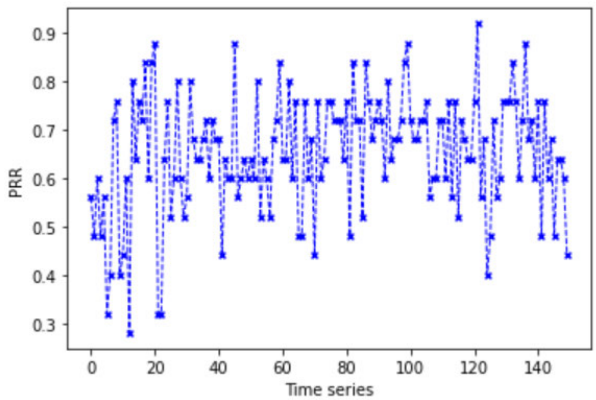

FIGURE 6. **Time series diagram of PRR in the office scenario.**

0.6~0.8. The interference in the parking lot is mostly static interference, so that the link quality often keeps at a good grade. If parked vehicles block some nodes, it will cause slight fluctuations in the link quality between nodes. In some periods, PRR drop sharply. Such as the time series of 5, due to vehicles entering and exiting parking lot, wireless link communication signal strength reduced, the link between nodes is affected.

As shown in Figure.6, the PRR in office scenario is concentrated approximately 0.6~0.9. The Wi-Fi signal of the router which is near node 1 causes continuous influence on the communication link of the node and causes PRR timing fluctuation. The interference mentioned above makes the link often in an unstable state, frequently switching between the transition area and the connected area. The interference in office scenario are computers and mobile phones, which

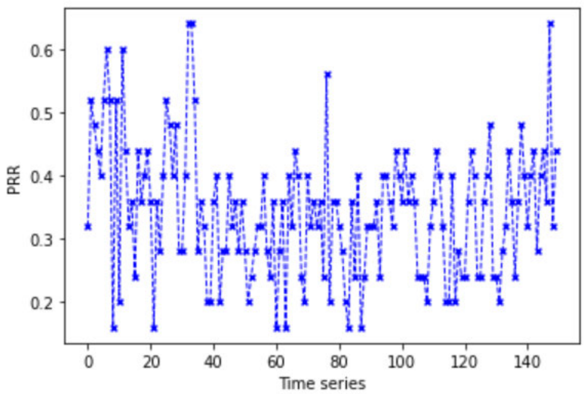

FIGURE 7. Time series diagram of PRR in the campus road scenarios.

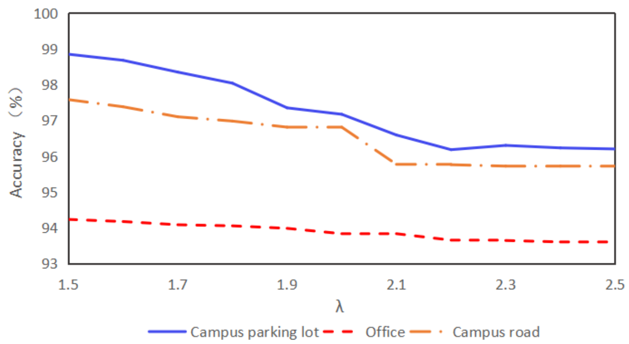

FIGURE 8. The accuracy of SCForest-LQE under different λ values.

result in general link mainly for electromagnetic interference caused by Wi-Fi, Bluetooth of quality large fluctuations.

Figure.7 represents the link quality for the road scenario. In the road scenario, the transition area is widely distributed. When the interference source moves continuously, the small changes in the signal between the nodes will lead to the conversion of links between good and bad, resulting in a burst link.

B. THE CLASSIFICATION OF LINK QUALITY BASED ON PRR

We use the INCK algorithm to divide the link quality grade. By constructing the PRR candidate, the influence of the noise samples on the division result is eliminated, and the medoids are selected at the dense part of the PRR candidate subset. In order to determine the candidate medoids subset S, we must introduce a stretch factor λ. This λ values have been experimentally verified in ranging from 1.5 to 2.5 according to [26]. We compare the accuracy of SCForest-LQE under different λ values, the results of three datasets are shown in Figure.8.

Figure.8 shows that stretch factor λ falls into the range of [1.5, 2.5]. When the λ is 1.5, the accuracy of SCForest-LQE is the highest both in three datasets. Therefore, we set the stretch factor λ as 1.5.

In this paper, the link quality grade is divided into 3 categories. It can be seen from Figure.9 that the samples that belong to Grade 2 have the largest number, followed by

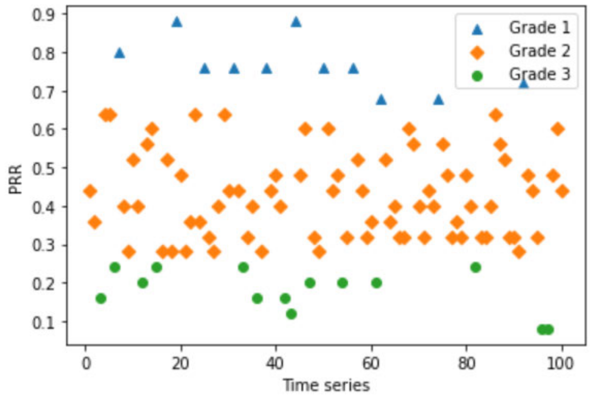

FIGURE 9. The distribution of link quality grades.

TABLE 2. The class proportion of datasets.

	Dataset	Grade1 (%)	Grade 2 (%)	Grade 3 (%)
1	Parking lot	17.20	66.46	16.34
2	Office	20.65	45.55	33.80
3	Road	19.42	60.48	20.10

Grade 1 samples, which have the best link quality. Due to the interference effect between the nodes in experimental environments, wireless links are often in the transitional region. Therefore, the distribution of the link quality data is unbalanced. Table 2 depicts the class proportion of selected datasets in this paper.

C. ANALYSIS OF THE PROPOSED MODEL

The learning performance of SCForest-LQE and gcForest do not depend seriously on careful parameter tuning, they are also able to get excellent performance by using the default setting. The d raw feature of multi-grained scanning takes the default setting as suggested in [21]. SCForest-LQE has one stratified sampling stage, three sampling sizes are used. For M minority samples, we also use the sizes of $[M/16]$, $[M/8]$, $[M/4]$ for sampling. For each size, we take 25, 50, 100 times sampling respectively. In cascade forest, each layer consists of 2 completely random forest and 2 random forest, which will bring greater diversity to our model, as suggested Each layer of the cascade forest is tested with 6-fold cross validation. Since the amount of link quality samples used in this paper is not much, the depth of cascading forest is reduced. We set the maximum layer of cascade as 40 and set the stop growing rounds as 6. The hyperparameter default settings of SCForest-LQE and gcForest are shown in Table 3.

To verify the performance of the improved model, we performed the comparison experiments between SCForest-LQE and gcForest with the same unbalanced datasets. In all of the experiments, SCForest-LQE and gcForest share the same cascade structure.

Figure.10 shows the estimation accuracy of the link quality estimation model SCForest-LQE and gcForest in each layer. The accuracy of SCForest-LQE is approximately 10% higher than gcForest in three experimental scenarios.

TABLE 3. The hyperparameter default setting.

SCForest-LQE	gcForest
Stratified sampling: Sampling size: [M/16], [M/8], [M/4] Cascade forest: Type of forest: Completely random tree forest, random forest No. Forests: 4 Maximum layer of cascade: 40 Stop growing rounds: 6	multi-grained scanning: Sliding window size: [d/16], [d/8], [d/4] Cascade forest: Type of forest: Completely random tree forest, random forest No. Forests: 4 Maximum layer of cascade: 40 Stop growing rounds: 6

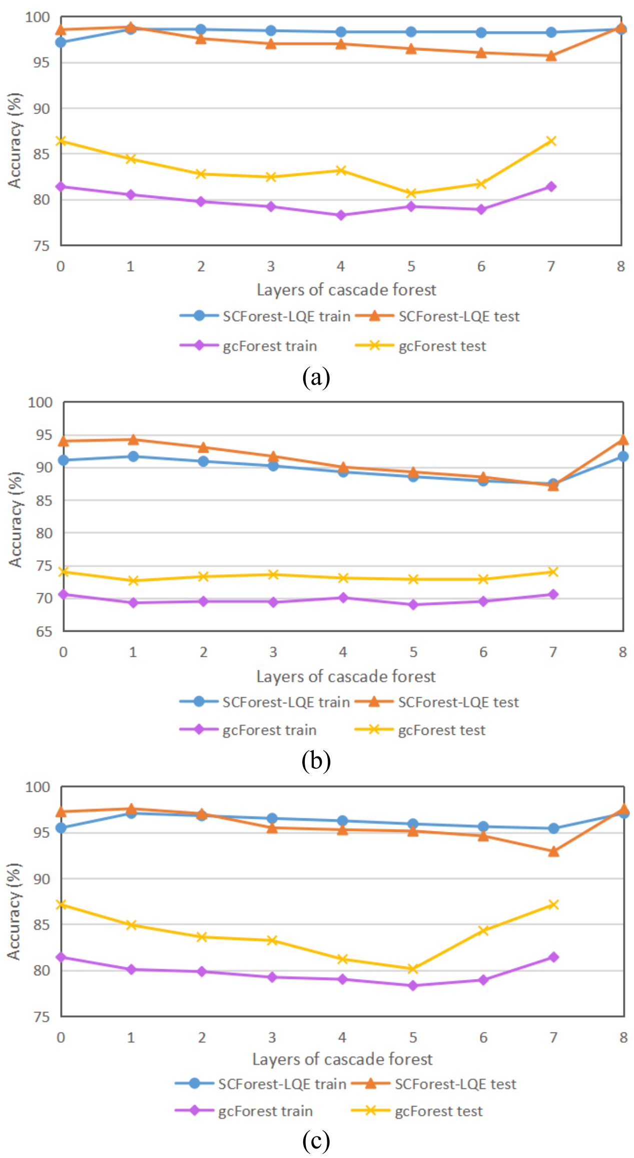

FIGURE 10. The comparison results of SCForest-LQE and gcForest in three scenarios: (a) Campus parking lot, (b) Office, (c) Campus road.

This finding proves that the SCForest-LQE uses the stratified sampling method to significantly improve the estimation ability by changing the unbalanced distribution of the link quality samples. From the 0th layer (initial layer) to the 1st layer of the cascade forest, the training result of SCForest-LQE is grown. At this time, the accuracy of SCForest-LQE is improved, and as a result, the amount of true cascade layers in SCForest-LQE is more than gcForest. When the accuracy of the model gradually decreases after 6 layers of training, the model terminates its training and outputs the best link quality estimation result as the final result. In Figure.9(a), from the 3rd to 4th and from the 5th to 6th layers, the accuracy of gcForest is also improved by 1%, but it still fails to exceed the accuracy of 86.38% at the 0th layer. Hence, the training is terminated after 6 rounds, and the link quality estimation result outputs at layer 7. Similarly, from the 5th to 6th layers, the accuracy does not exceed 87.14% at the 0th layer in Figure.9(c), and therefore gcForest is also terminated at the 7th layer. In the office scenario, the link quality is relatively poor. The accuracy of gcForest has dropped to 74%, while the accuracy of SCForest-LQE remains above 85%.

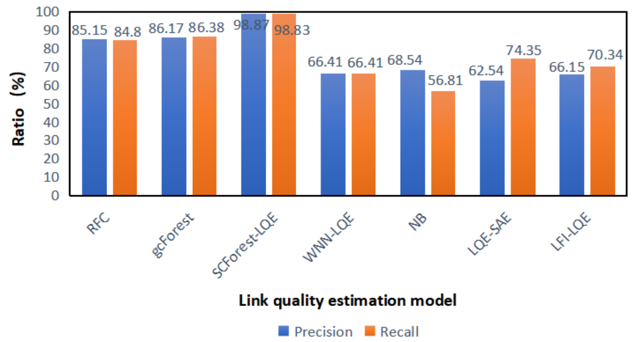

FIGURE 11. Comparison of the precision and recall in the campus parking lot scenario.

D. VERIFICATION AND COMPARISON OF SCFOREST-LQE

To further verify the estimation ability of the link quality estimation model SCForest-LQE, we conduct more comparison experiments, and the results are shown in Figures.11–13. The gcForest-based model (gcForest), the random forest-based model [19] (RFC), the wavelet neural network-based model [13] (WNN-LQE), the naive Bayes-based model [11] (NB), the stacked autoencoder-based model [20] (LQE-SAE) and the lightweight, fluctuation insensitive multi-parameter fusion-based model [14] (LFI-LQE) are chosen to compare with the proposed estimator.

In all of the experiments, the forest shares the same parameters. Each experiment is conducted ten times and we take the average of ten consecutive runs as the final result.

The performance of these estimators under different environments is shown Figures.11–13. It can be seen that SCForest-LQE has better performance compared with other link quality estimators; especially the precision and recall of SCForest-LQE reach the highest values 98.87% and 98.83% in the campus parking lot scenario. Because the nodes in the office scenario are continuously interfered by other wireless signals, the link is often in a fluctuating state. Thus, the performance of some estimators is poor in the office scenario.

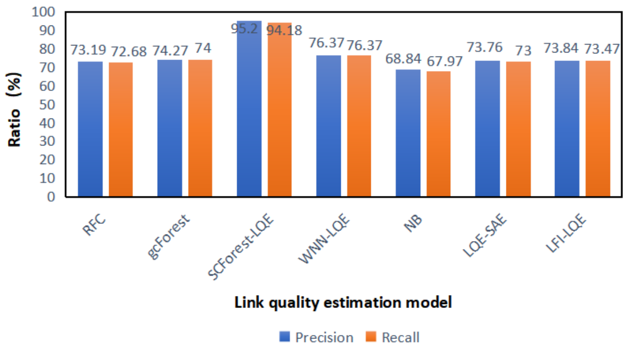

FIGURE 12. Comparison of the precision and recall in the office scenario.

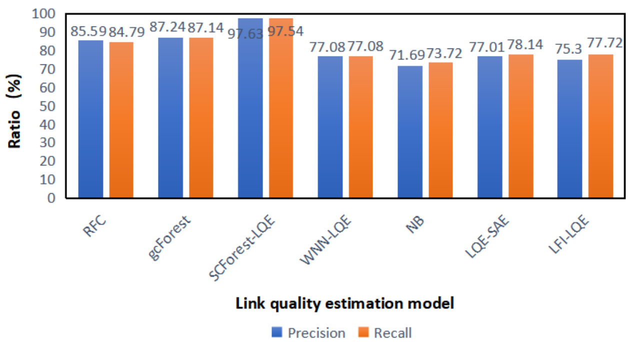

FIGURE 13. Comparison of the precision and recall in the campus street scenario.

TABLE 4. F1-score comparison result of campus parking.

	Grade 1	Grade 2	Grade 3
RFC	0.80	0.89	0.75
gcForest	0.83	0.90	0.74
SCForest-LQE	0.99	0.98	0.99
NB	0.55	0.58	0.51
LQE-SAE	0.71	0.83	0.29
LFI-LQE	0.44	0.81	0.25

Compared with the parking lot scenario and road scenario, the precision and recall of gcForest and RFC in the office decrease significantly. This finding proves that these models can distinguish static or short-term dynamic interference links well, but these models have a weak ability to identify long-term unstable links. Although SCForest-LQE is also affected, the recall and precision remain above 92%. NB, WNN-LQE, LQE-SAE and LFI-LQE have better performance in office scenarios and campus road scenarios, but the precision and recall of them in campus parking lots are poor. This finding indicates that these models have poor estimation performance for scenarios with many static obstructions.

More comprehensively, to compare the performance of these link quality estimators, the F1-score under different link quality grade is calculated, which is shown in Tables 4–6. As the link quality data is unbalanced, it causes the evaluator is easy to ignore the minority class during classification. In Table 4 and Table 6, the F1-score of gcForest is close to SCForest-LQE at Grade 2, but the F1-score of gcForest is

TABLE 5. F1-score comparison result of Office.

	Grade 1	Grade 2	Grade 3
RFC	0.78	0.70	0.73
gcForest	0.78	0.73	0.74
SCForest-LQE	0.98	0.91	0.94
NB	0.75	0.67	0.63
LQE-SAE	0.77	0.73	0.71
LFI-LQE	0.78	0.73	0.72

TABLE 6. F1-score comparison result of campus road.

	Grade 1	Grade 2	Grade 3
RFC	0.91	0.87	0.73
gcForest	0.92	0.90	0.74
SCForest-LQE	0.99	0.96	0.97
NB	0.83	0.79	0.36
LQE-SAE	0.91	0.83	0.46
LFI-LQE	0.91	0.84	0.24

poor at Grade 3. The reason is that the samples of Grade 1 and Grade 3 is fewer than samples of Grade 2, which affects the estimation performance of the estimators for minority link quality samples. In the campus parking lot and road scenario, LQE-SAE and LFI-LQE have higher F1-score at the Grade 1 and Grade 2 than Grade 3. It can be concluded that LQE-SAE and LFI-LQE obtain inaccurate estimations under bad links. Owing to the high proportion of samples in transitional region, when the estimation models have higher accuracy at Grade 2, the final estimation result is better. After using stratified sampling method to combat class imbalance, which makes SCForest-LQE performs better on the three grades. On the other hand, Tables 4–6 show that the F1-score of models with forest structure is higher than LQE-SAE, NB and LFI-LQE in three experimental scenarios, indicating that the forest model is more adaptable to the study of link quality estimation.

E. STABILITY AND REACTIVITY ANALYSIS

The stability of model refers to the ability to tolerate transient variations in link quality [2]. Taking the link quality grade as a benchmark, the deviation between the estimation result and the real value is observed in scenarios. When a link quality shows transient degradation or increase, if an estimation model comparing with other models can ignore this change to keep estimator stable, its stability is better.

The reactivity of model refers to the ability to quickly react to persistent changes in link quality [2]. If the estimation results of the estimators are closer to the real label when link quality grade changes frequently, its reactivity is better.

Experiments between SCForest-LQE and other estimators with similar structures at different status of link quality are conducted to analyze stability and reactivity, which shown in Figures.14–16.

To compare the stability of the LQEs, we can observe their sensitivity to transient fluctuations through the link quality estimating grade. According to the Figure.14, at the 1426th

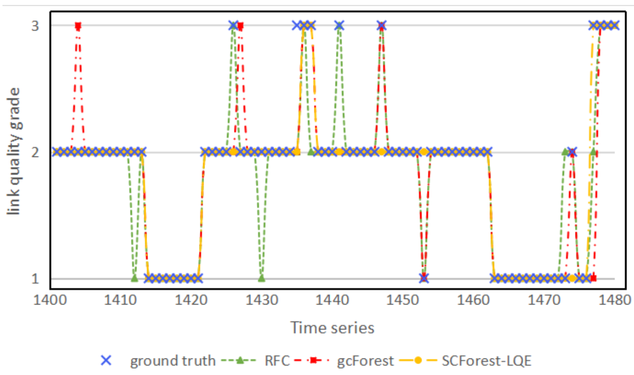

FIGURE 14. Comparison of the estimation results in the campus parking lot scenario.

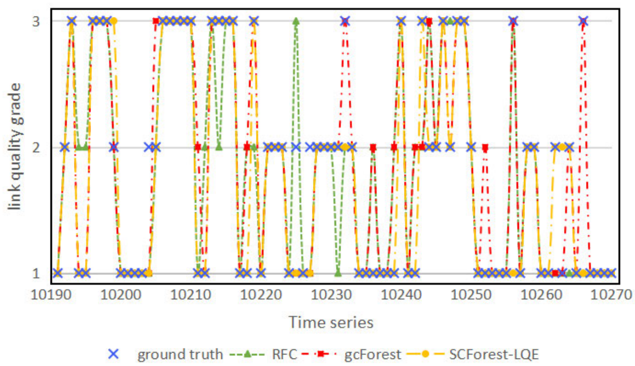

FIGURE 15. Comparison of the estimation results in the office scenario.

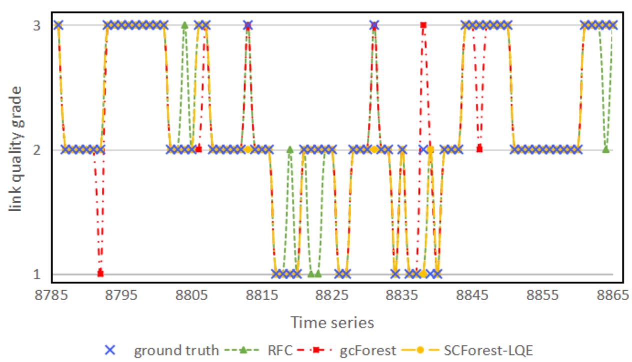

FIGURE 16. Comparison of the estimation results in the campus road scenario.

estimation result of the real value, the link quality grade is changed from the second grade to third grade. The SCForest-LQE model maintains second-grade of the estimation result, while gcForest and RFC quickly change the estimation result, and the same situation occurs at the 1453rd and 1474th estimation result. Hence, we reach conclusion that SCForest-LQE is more stable than gcForest and RFC, which are more sensitive to the transient fluctuations.

As shown in Figure.15, from the 10190th to the 10270th estimation results, the link quality grade fluctuates frequently between three grades in the office scenario, which indicates that the link quality is very unstable. By analyzing Figure.15, it can be seen that the RFC and gcForest model can quickly respond to frequent changes, while their accuracy is relatively poor. The SCForest-LQE model makes fewer mistakes when the link quality grade changes suddenly. This finding shows that SCForest-LQE is more consistent with the ground truth.

Figure.16 contains a short section of stable link quality and a section of frequently fluctuating link quality, in such a way that it can comprehensively evaluate the stability and reactivity of LQEs. According to this figure, we retain the following observations. From the 8802nd to the 8816th estimation results, gcForest and RFC are more reactivity than SCForest-LQE in transient fluctuations. From, the 8833rd to the 8841st estimation result, there is an unstable link quality, and the proposed link quality estimation model can still stably and accurately estimate the link quality.

VI. CONCLUSION

In this paper, we propose a link quality estimation based on SCForest for WSNs. INCK is utilized to divide link quality grade according to the PRR. The noise values of the link samples are deleted by the Pauta criterion. Stratified sampling of the imbalanced link quality samples is conducted according to the link quality distribution regions, which improves the accurate estimation capability of SCForest-LQE in a dynamic transitional region and for minority samples. With the help of cascade forest layer-by-layer learning, the feature training is improved. SCForest-LQE adaptively adjusts its cascade level according to the size of the link quality samples to achieve better estimation capabilities. We conduct experiments in campus parking lots, road, and office scenarios. The results show that SCForest-LQE has better estimation performance than the results of gcForest, RFC, WNN-LQE, NB, LQE-SAE and LFI-LQE. When the link quality grade suddenly changes, SCForest-LQE can still estimate the link quality stably.

One important future issue is to accelerate. As a deep learning model, SCForest-LQE takes a long time to train the model, and the structure of deep forest is not suitable to GPUs. In the future, we will try using distributed computing to implement SCForest-LQE. Another important future work is to test SCForest-LQE in other application scenarios of WSNs and attempt to verify the proposed method in other domain networks.

REFERENCES

[1] A. Guidara, G. Fersi, F. Derbel, and M. B. Jemaa, "Impacts of temperature and humidity variations on RSSI in indoor wireless sensor networks," *Procedia Comput. Sci.*, vol. 126, pp. 1072–1081, Jan. 2018.

[2] N. Baccour, A. Koubâa, L. Mottola, M. A. Zuñiga, H. Youssef, C. A. Boano, and M. Alves, "Radio link quality estimation in wireless sensor networks: A survey," *ACM Trans. Sensor Netw.*, vol. 8, no. 4, pp. 1–33, Sep. 2012.

[3] R. L. Gomes, W. Moreira, E. Cerqueira, and A. J. Abelém, "Using fuzzy link cost and dynamic choice of link quality metrics to achieve QoS and QoE in wireless mesh networks," *J. Netw. Comput. Appl.*, vol. 34, no. 2, pp. 506–516, Mar. 2011.

[4] R. D. Gomes, D. V. Queiroz, A. C. L. Filho, I. E. Fonseca, and M. S. Alencar, "Real-time link quality estimation for industrial wireless sensor networks using dedicated nodes," *Ad Hoc Netw.*, vol. 59, pp. 116–133, May 2017.

[5] N. Baccour, A. Koubaa, H. Youssef, M. B. Jamaa, D. D. Rosario, M. Alves, and L. B. Becker, ''F-LQE: A fuzzy link quality estimator for wireless sensor networks,'' in *Proc. 7th Eur. Conf. Wireless Sensor Netw.*, Coimbra, Portugal, Feb. 2010, pp. 240–255.

[6] Z. Huang, L. Y. Por, T. F. Ang, M. H. Anisi, and M. S. Adam, ''Improving the accuracy rate of link quality estimation using fuzzy logic in mobile wireless sensor network,'' *Adv. Fuzzy Syst.*, vol. 2019, pp. 1–13, Apr. 2019.

[7] Y. Li, Y. Guo, W. Zhao, J. Yu, and M. Daneshmand, ''Statistical characteristics of wireless link in opportunistic networks,'' in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Anaheim, CA, USA, Dec. 2012, pp. 1682–1686.

[8] C. U. Bas and S. C. Ergen, ''Spatio-temporal characteristics of link quality in wireless sensor networks,'' in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Paris, France, Apr. 2012, pp. 1152–1157.

[9] C. J. Lowrance and A. P. Lauf, ''Link quality estimation in ad hoc and mesh networks: A survey and future directions,'' *Wireless Pers. Commun.*, vol. 96, no. 1, pp. 475–508, Apr. 2017.

[10] T. Liu and A. E. Cerpa, ''Data-driven link quality prediction using link features,'' *ACM Trans. Sensor Netw.*, vol. 10, no. 2, pp. 1–35, Jan. 2014.

[11] T. Liu and A. E. Cerpa, ''Temporal adaptive link quality prediction with online learning,'' *ACM Trans. Sensor Netw.*, vol. 10, no. 3, pp. 1–41, Apr. 2014.

[12] W. Sun, W. Lu, Q. Li, L. Chen, D. Mu, and X. Yuan, ''WNN-LQE: Wavelet-neural-network-based link quality estimation for smart grid WSNs,'' *IEEE Access*, vol. 5, pp. 12788–12797, Jul. 2017.

[13] W. Liu, Y. Xia, R. Luo, and S. Hu, ''Lightweight, fluctuation insensitive multi-parameter fusion link quality estimation for wireless sensor networks,'' *IEEE Access*, vol. 8, pp. 28496–28511, Feb. 2020.

[14] X. Xue, W. Sun, J. Wang, Q. Li, G. Luo, and K. Yu, ''RVFL-LQP: RVFL-based link quality prediction of wireless sensor networks in smart grid,'' *IEEE Access*, vol. 8, pp. 7829–7841, Jan. 2020.

[15] X. Yuan, S. Qi, Y. Wang, and H. Xia, ''A dynamic CNN for nonlinear dynamic feature learning in soft sensor modeling of industrial process data,'' *Control Eng. Pract.*, vol. 104, Nov. 2020, Art. no. 104614.

[16] X. Yuan, L. Li, Y. Shardt, Y. Wang, and C. Yang, ''Deep learning with spatiotemporal attention-based LSTM for industrial soft sensor model development,'' *IEEE Trans. Ind. Electron.*, early access, Apr. 9, 2020, doi: 10.1109/TIE.2020.2984443.

[17] X. Yuan, C. Ou, Y. Wang, C. Yang, and W. Gui, ''A layer-wise data augmentation strategy for deep learning networks and its soft sensor application in an industrial hydrocracking process,'' *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 13, 2019, doi: 10.1109/TNNLS.2019.2951708.

[18] W. Sun, P. Li, Z. Liu, X. Xue, Q. Li, H. Zhang, and J. Wang, ''LSTM based link quality confidence interval boundary prediction for wireless communication in smart grid,'' *Computing*, vol. 8, pp. 1–19, May 2020.

[19] L. Liu, S. Gao, and J. Shu, ''Link quality prediction based on random forest,'' *J. Commun.*, vol. 40, no. 4, pp. 202–211, Apr. 2019.

[20] X. Luo, L. Liu, J. Shu, and M. Al-Kali, ''Link quality estimation method for wireless sensor networks based on stacked autoencoder,'' *IEEE Access*, vol. 7, pp. 21572–21583, Feb. 2019.

[21] Z. Zhou and J. Feng, ''Deep forest: Towards an alternative to deep neural networks,'' in *Proc. IJCAI*, Melbourne, VIC, Australia, Feb. 2017, pp. 3553–3559.

[22] X. Lu, Z. Duan, Y. Qian, and W. Zhou, ''A malicious code classification method based on deep forest,'' *J. Softw.*, vol. 31, no. 5, pp. 1454–1464, May 2020.

[23] X. Yang and S. Zhang, ''Anomaly detection model based on multi-granularity cascading isolated forest algorithm,'' *J. Commun.*, vol. 40, no. 8, pp. 133–142, Aug. 2019.

[24] J. Ren, B. Hou, and Y. Jiang, ''Deep forest architecture under multi-example learning,'' *J. Comput. Res. Develop.*, vol. 56, no. 8, pp. 1670–1676, Jun. 2019.

[25] G. Cerar, H. Yetgin, M. Mohorcic, and C. Fortuna, ''Machine learning for wireless link quality estimation: A survey,'' *CoRR*, vol. 9, pp. 1–33, Dec. 2020. [Online]. Available: https://arxiv.org/abs/1812.08856

[26] D. Yu, G. Liu, M. Guo, and X. Liu, ''An improved K-medoids algorithm based on step increasing and optimizing medoids,'' *Expert Syst. Appl.*, vol. 92, pp. 464–473, Feb. 2018.

[27] H.-S. Park and C.-H. Jun, ''A simple and fast algorithm for K-medoids clustering,'' *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009.

[28] J. Deng, L. Shan, D. He, and R. Tang, ''The processing method of missing data and its development trend,'' *Statist. Decis.*, vol. 35, no. 23, pp. 28–34, 2019.

MU HE was born in Yongzhou, Hunan, in 1997. She is currently pursuing the master's degree with Nanchang Hangkong University. Her main research interest includes wireless sensor networks.

JIAN SHU was 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 main research interests include wireless sensor networks, embedded systems, and software engineering. He is also a Senior Member of CCF.