GNSS Multipath Detection Based on Decision Tree Algorithm in Urban Canyons

Yue Wang(✉), Jiawei Xu(✉), Rong Yang(✉), and Xingqun Zhan(✉)

Shanghai Jiao Tong University, Shanghai 200240, China
{johnld,xjw000830,rongyang,xqzhan}@sjtu.edu.cn

Abstract. Multipath detection has long been a fundamental problem in GNSS research and application especially under heavily urbanized condition. In this work, we aim to apply a machine learning algorithm to detect and classify multipath error in urban, kinematic situation based on Rinex datasets provided by the University of Texas. Correspondingly, the data samples are classified into 3 groups according to the chip length from GPS L1 and L5: Short Multipath (0–30 m), Medium Multipath (30–90 m) and Long Multipath ($>$90 m). As a result, the algorithm achieves an average accuracy of 70% in the 5-folded cross validation. Furthermore, the detection result of satellites with various conditions of blockage are compared to give some angle of optimization.

Keyword: GNSS · Multipath detection · Urban canyon · Machine learning · Decision tree

1 Introduction

With the continuous development and improvement of GNSS (Global navigation satellite system), it gradually provides more and more accurate and reliable positioning services, making applications such as unmanned vehicles and drones possible. These applications are often applied in urban dynamic environments and have higher requirements for GNSS positioning services. In this so-called urban canyon environment, GNSS signals may be blocked by buildings around the target, creating multipath errors on the received signals, which is considered to be the main source of GNSS signal errors in this environment. Therefore, the detection and removal of multipath errors is a necessary step before the actual application of GNSS signal.

To remove multipath errors, many previous studies have been done. At the antenna level, multipath errors can be reduced by upgrading the quantity or quality of the antenna [4]. At the receiver level, advanced receiver algorithms such as VDLL can improve the reliability of the signal when it is affected by multipath [3]. While at the software level, inertial navigation is used in [5] to assist GPS systems in positioning under multipath interference, and 3D building models are used in [10] and [6] to label LOS/NLOS signals. These labelled data, in turn, can be used to assist machine learning algorithms [10]. While features calculated based on pseudorange and phase observations are used in [8] and CNN is used as a model in [9].

C. Yang and J. Xie (Eds.): *China Satellite Navigation Conference (CSNC 2021) Proceedings*, LNEE 773, pp. 375–383, 2021.
https://doi.org/10.1007/978-981-16-3142-9_35

Among these methods, the machine learning algorithm shows some advantages with its flexibility and ability to respond to multiple situations or input information. In this study, we used a decision tree machine learning algorithm based on MATLAB's Classification Learner toolbox. Considering ease of use and understanding, the input features are based on the information from Rinex files. The features include carrier-to-noise ratio, pseudorange-doppler residual, pseudorange residual, elevation angle and azimuth angle. In terms of data labelling, two labelling methods relying on Rinex files and true values were used in this study: the pseudorange positioning method and the pseudorange correction method. Considering that the pseudorange observation is not able to distinguish well between the multipath and NLOS signal, and that these 2 interferences have similar effect on pseudorange, they are collectively referred to as multipath interference in this study, For the labelled results, the chip length of GPS L1/L5 was referred (which is 300 m and 30 m separately) to make a classification of 3 categories: short multipath (0–30 m), medium multipath (30–90 m) and long multipath (>90 m).

Overall, in this study, a decision tree machine learning model was trained using input from Rinex files and reference to the ground truth to detect the multipath among the dataset, and finally is able to achieve an average accuracy of 75% against its own labelled results. In addition, this paper also analyses and compares the classification results of multipath errors from the perspective of signal frequency and code rate with respect to the physical significance of multipath generation, and gives some optimization directions for current machine learning algorithms for detecting multipath effects.

2 Data Sources and Labelling

2.1 Data Sources

The source data used in this study are from [7], an open dataset collected by researchers at the University of Texas on 2019/05/09 in the downtown area of Austin, during which the researchers drove from the sparsely built campus area to the heavily built-up downtown area and back when the blockage around changed in terms of height and angle. Thus, the exposure to multipath disturbances encompasses the three scenarios classified in this study. The process last about 2 h.

2.2 Data Labelling

In this study, the labelling result of the data will be determined by a combination of the two methods. Since the variable that determines its final labelling result has the same physical significance (multipath error value), this variable will be averaged from the calculations of the two methods.

2.2.1 Pseudorange Positioning Method

The pseudorange positioning method is calculated using the traditional least squares method [11]. For the target satellite at a certain moment, the pseudorange observations of all satellites in the same system and frequency as the target satellite at the current moment are used for the least-squares positioning. And the difference of this positioning result and the true value from the dataset is taken absolute value as the final considered multipath error value. Thus, this main consideration of this method would be the positioning result.

The reason for using this method is that it is simple and intuitive to calculate. It contains the signal quality of all the targeted satellites of the same frequency at that moment, thus has a better performance with receiver at different positions. On the other hand, the method also has the obvious drawback that it cannot distinguish the errors resulted from different satellites, which is why we need to combine it with the second method for more accurate and targeted labelling.

2.2.2 Pseudorange Correction Method

The pseudorange correction method takes a single pseudorange observation of the target satellite as the main discriminator at a given moment. First, the current position of the target satellite is solved from the ephemeris file. Then the difference between the receiver's position from true value and the position of the target satellite is taken as the distance true value. This pseudorange observation is corrected according to the following pseudorange observation model [11].

$$\rho = \boldsymbol{r} + \delta_{tu} - \delta_t + I + T + \in +\mathrm{M} \tag{1}$$

Where ρ is the pseudorange observation, r is the satellite-receiver distance, M is the perceived multipath error value, ϵ is the thermal noise, δ_{tu} is the receiver clock difference, δ_t is the satellite clock difference, I is the ionospheric error, and T is the tropospheric error. The first two clock differences are calculated from ephemeris files, and the last two atmospheric delays are calculated from the model.

Finally, the difference between the true value of the satellite-receiver distance and the corrected pseudorange is taken absolute value as multipath error value for calibration.

Although this method can distinguish between satellites, its reliability is not high enough in the complex urban dynamic environment. For example, clock difference calculation error, atmospheric delay error, etc. may make the result far from the real multipath error value, so it needs to be combined with the previous method.

2.3 Labelling Result

The labelling results for short multipath (0−30 m), medium multipath (30−90 m) and long multipath (>90 m) are obtained according to the above methods from the selected 23555 samples of GPS L1, 21690 of GAL E1 and 19230 of E5b. The number of epochs is converted into percentages for comparison purpose (Table 1).

3 Selection and Calculation of Features

All features are calculated and applied in model training or validation following two principles. One is that when the absolute value of the feature significantly exceeds the error caused by the normal multipath effect, the data at that point is considered unusable. And the second is that when one of the observations used in the signal is stumped, the data at that point is considered unusable.

Table 1. Result of labelling.

Percentage	GPS L1	GAL E1	GAL E5b
Short multipath	33.43%	34.80%	21.70%
Medium multipath	33.24%	33.42%	41.00%
Long multipath	33.33%	32.20%	37.30%

1. Carrier-to-Noise Ratio: $\boldsymbol{C/N_0}$

   The carrier-to-noise ratio C/N_0 is a common characteristic that indicates the strength of the signal received by the receiver. According to the characteristics of signal propagation, the signal will be significantly reduced in strength when it is reflected and blocked by walls. Therefore, the carrier-to-noise ratio is used as a feature here. The value can be read out directly from a Rinex file.
2. Pseudorange-Doppler Residuals: $\boldsymbol{\rho d}$

   The Doppler shift is another observation related to the multipath, which characterizes the rate of change of the pseudorange. However, it is calculated differently from the pseudorange in the receiver's algorithm, the former given by the code tracking loop, while the latter is given by the carrier frequency. Therefore, the two observation can be considered as independent when multipath happens and the difference between the pseudorange rate of change and the Doppler shift can indicate the consistency of the internal calculations of the receiver, which in turn reflects the receiver receiving multipath interference. The calculation requires the use of both pseudorange and Doppler shift observations from the Rinex file.

   First, the pseudorange is differenced between epochs. Then the doppler value is converted into length. Finally, difference the two value and take absolute value to get the feature.

   Here, the effect of time-differenced pseudorange is ignored as the interference of multipath is considered to last for a while, namely around 10 s.
3. Pseudorange Residuals: $\boldsymbol{\rho_r}$

   The pseudorange residual exploits the inconsistency between the pseudorange observations of certain frequency of the target satellite and the results of the overall pseudorange positioning. According to the conclusions of [8], this feature can characterize the degree of multipath error when the number of observed satellites is sufficient.

   The method and principle of its calculation are almost identical to the method of calculating the multipath error by the pseudorange correction method in the labelling part (Eq. (1)), except that the true value position is replaced by the result of pseudorange positioning of the corresponding frequency at the current moment. Thus, the value of this feature is much smaller than pseudorange positioning labelling, namely 0–5 m.
4. Elevation angle

   The satellite elevation angle refers to the angle between the satellite and the receiver line and the horizontal plane. In the same urban environment, it is clear that satellites with low lift angles are more likely to be blocked by surrounding obstacles than satellites with high lift angles, thus generating multipath errors. In the case where the model considers multiple satellites with different lift angles, this value has considerable relevance to the absence of multipath errors.This value is calculated from the ephemeris data at the time during the pseudorange positioning.

5. Azimuth Angle

The satellite azimuth refers to the angle turned by the receiver clockwise from the due north direction line to the horizontal direction line of the satellite. In a similar urban canyon environment, the degree of obstruction in different directions may also vary, which can lead to different levels of multipath interference for satellites at different azimuths. Although this correlation is diminished in the dynamic case, the azimuth of satellites can still play a role in model judgments in short time and small range conditions. Similarly, this value is calculated from the prevailing ephemeris data during pseudorange positioning.

4 Model Training and Classification Results

In the experimental phase, we used all satellite data of GPS L1, GALILEO E1 and E5b in the dataset.

4.1 Model Training

In this section, we use the Rinex files of GPS L1, GAL E1 and E5b frequencies in the dataset to perform the labelling, feature calculation and training. A total of 20,604 samples from 5 satellites were used for GPS L1and 19475, 17018 samples from 5 satellites were used for GAL E1 and E5B.The overall self-test accuracy with all the above 5 features used is shown in the table below (Table 2).

Table 2. Accuracy of the model

	GPS L1	GAL E1	GAL E5b
Model accuracy (%)	68.4%	80.3%	80.7%

To further verify the meaning of these three sets of self-test accuracies, recall rates of different labelling groups were calculated using the model predictions (Table 3).

Table 3. Recall of the model of different labelling groups

Recall (%)	GPS L1	GAL E1	GAL E5b
Short multipath	85%	92%	84%
Medium multipath	61%	75%	90%
Long multipath	59%	73%	68%

From the data in the table, it can be seen that the model training results for different frequencies vary widely. For GPS L1 and GAL E1 with higher carrier frequencies, they have better distinguishing capabilities for short multipath, but poorer for medium and long multipath. For the medium and long multipaths, the model classifies these samples as two other wrong multipath cases with 20% probability respectively, leading to a lower overall accuracy. For E5b, a frequency point commonly used for multi-frequency combination analysis, its carrier frequency is lower while its level of multipath error is higher (which will be discussed in detail later). This is reflected in the prediction results as the model tends to classify samples with corresponding features as medium multipath, resulting in a higher recall rate for medium and short multipath, but long multipath cases are often misclassified as medium multipath. Thus, the E5b model has a higher overall accuracy.

In order to verify the contribution of each feature, we also exclude one of the features and re-trained model to see its accuracy. The results are shown in the following Table 4.

Table 4. Accuracy of the model with certain feature excluded

Feature excluded	Accuracy
None (with all features)	80.7%
Carrier-to-noise ratio	81.4%
Pseudorange residual	78.5%
Pseudorange-Doppler residual	78.8%
Elevation angle	76.0%
Azimuth angle	77.7%

It can be seen that all feature, except the carrier-to-noise ratio, reduce the overall accuracy after being removed. Among all features, the effect of lift angle and azimuth angle are more significant, which indicates that both of them can better distinguish the situation of multipath interference based on the previous experience in the dynamic situation of similar environment. On the contrary, the correlation between the carrier-to-noise ratio and multipath interference in the dynamic environment may not be strong enough to support our machine learning model, and may produce some interference instead.

4.2 Verification of Classification Results

This section focuses on the comparison of the between different frequencies, i.e., different carrier frequencies and code rates. Thus, the results of two frequencies from the same satellite system, E1 and E5b of GAL, are selected (Table 5).

According to the previously labelled results, the E5b frequency point with lower carrier frequency, longer wavelength, and higher code rate suffers more long multipath and medium multipath cases than the E1 frequency point does, i.e., E5b suffers more serious multipath errors at the level of pseudorange observation than E1. This is different

Table 5. Labelling result of GAL E1 and E5b

Percentage	GAL E1	GAL E5b
Short multipath	34.80%	21.70%
Medium multipath	33.42%	41.00%
Long multipath	32.20%	37.30%

from our expected results: considering that the lower carrier frequency gives E5b a longer carrier wavelength, which makes the same multipath propagation cause a relatively larger code phase delay for E5b; the final multipath error is the code phase delay multiplied by the chip length, while the code rate of E5b is 10 times that of E1 (10.23 MHz: 1.023 MHz), and the chip length is 0.1 times (30 m: 300 m).

There are several possible reasons for this error. One is the processing mode compatibility of the receiver platform with the special AltBOC (alternating binary offset carrier) modulation method for the GAL E5b frequency point. According to [12], the AltBOC modulation method, although with high tracking accuracy, requires a high RF bandwidth and sampling rate of the receiver, which may have produced a larger error at the stage of generating the code phase delay. Secondly, the steps of pseudorange positioning and correction may produce an offset. By observing the above two labelling benchmarks on E5b, it is found that the latter part of the route yields higher positioning and pseudorange errors under the same open conditions. This may be due to the fact that the positioning and correction algorithms do not apply higher-order smoothing algorithms such as carrier or Doppler smoothing, resulting in the accumulation of random errors in the moving process not covered by the observation model (Fig. 1).

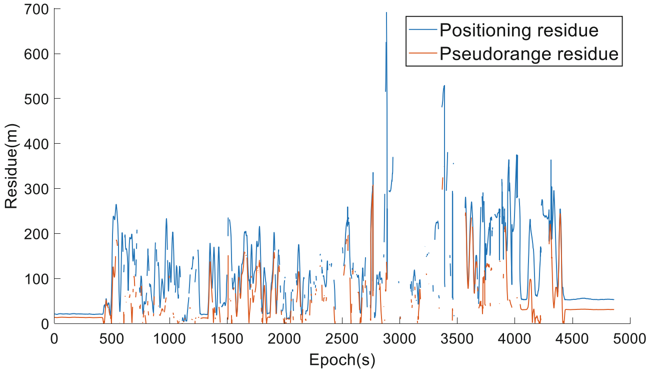

Fig. 1. Increase of labelling benchmarks over time

5 Conclusion

In this study, a decision tree algorithm is applied to give a solution to the multipath classification detection problem in urban-canyon environments. A total of five feature from Rinex files output are used to train a machine learning model. As a result, the model is able to achieve an average classification accuracy of about 75% in cross-validation. While carrier-to-noise ratio may produce negative effect on accuracy, elevation and azimuth angle contribute more than other features that are applied. Compared to the existing work that shares the same method such as [6], our solution mainly deals with dynamic situations and verifies the availability of machine learning algorithm under such condition. However, the physical significance and actual effect of each feature may differ from static condition. Finally, based on the initial labelling algorithm, we analyze the reason why the E5b frequency point of the Galileo system suffers more from multipath errors than the E1 frequency point based on the physical significance of the multipath error generation. The conclusion is that the error may be generated by the hardware of the receiver or the labelling algorithm. As a pioneering study, the feasibility and reliability of machine learning algorithms for detection and rejection of multipath errors at the Rinex file level is verified in this study.

6 Future Outlook

Based on this study, we have plans for more in-depth research and development. First, we plan to use a more comprehensive data labelling algorithm, combining the 3D city model, the multipath error reference values given by the receiver itself to make a weighted comprehensive labelling. Also, the least-square positioning can be modified to allocate more weight to satellites with higher elevation angle so that the result can be more validated in terms of feature calculation. Following that, we can add multi-dimensional feature variables, and conduct joint machine learning algorithm research for multipath problems. The research currently underway has machine learning algorithms based on receiver correlators as features. And in the future, we hope to apply vehicle vision signals to further improve the reliability and strain of the algorithms. Eventually, different machine learning algorithms, such as convolutional neural networks, can be applied to better integrate the impact of each feature quantity on the classification conclusion. Overall, we plan to conduct our own data-collection test to verify our method in terms of universality and apply above-mentioned improvement.

References

1. Munin, E., Blais, A., Couellan, N.: Convolutional neural network for multipath detection in GNSS receivers. In: 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), Singapore, pp. 1–10 (2020). https://doi.org/10.1109/AIDA-AT48540.2020.9049188.
2. Groves, P.D.: Shadow matching: a new GNSS positioning technique for urban Canyons. J. Navig. **64**(3), 417–430 (2011)

3. Hsu, L.-T., Jan, S.-S., Groves, P.D., Kubo, N.: Multipath mitigation and NLOS detection using vector tracking in urban environments. GPS Solutions **19**(2), 249–262 (2014). https://doi.org/10.1007/s10291-014-0384-6
4. Jiang, Z., Groves, P.D.: NLOS GPS signal detection using a dual-polarisation antenna. GPS Solutions **18**(1), 15–26 (2012). https://doi.org/10.1007/s10291-012-0305-5
5. Chiang, K.-W., Duong, T., Liao, J.-K.: The performance analysis of a real-time integrated INS/GPS vehicle navigation system with abnormal GPS measurement elimination. Sensors **13**(8), 10599–10622 (2013)
6. Hsu, L.: GNSS multipath detection using a machine learning approach. In: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, pp. 1–6 (2017). https://doi.org/10.1109/ITSC.2017.8317700.
7. Narula, L., et al.: TEX-CUP: the university of texas challenge for urban positioning. In: 2020 IEEE/ION Position, Location and Navigation Symposium (PLANS), Portland, OR, USA, pp. 277–284 (2020). https://doi.org/10.1109/PLANS46316.2020.9109873.
8. Hsu, L.T., Tokura, H., Kubo, N., Gu, Y., Kamijo, S.: Multiple faulty GNSS measurement exclusion based on consistency check in Urban Canyons. IEEE Sens. J. **17**(6), 1909–1917 (2017)
9. Quan, Y., Lau, L., Roberts, G.W., Meng, X., Zhang, C.: Convolutional neural network based multipath detection method for static and kinematic GPS high precision positioning. Remote Sens. **10**, 2052 (2018)
10. Xu, B., Jia, Q., Luo, Y., Hsu, L.-T.: Intelligent GPS L1 LOS/Multipath/NLOS classifiers based on correlator-, RINEX- and NMEA-level measurements. Remote Sens. **11**, 1851 (2019)
11. 谢钢. GPS 原理与接收机设计. 电子工业出版社 (2017)
12. 闫温合,何在民,胡永辉.伽利略E5频段信号及其性能研究[J].时间频率学报 **39**(01)，25–32 (2016)