Identification of Wireless User Perception Based on Unsupervised Machine Learning

Kaixuan Zhang, Guanghui Fan, Jun Zeng, and Guan Gui(✉)

College of Telecommunications and Information Engineering,
Nanjing University of Posts and Telecommunications, Nanjing 210003, China
guiguan@njupt.edu.cn

Abstract. Wireless user perception (WiUP) plays an important role in designing next-generation wireless communications systems. Users are very sensitive with the quality of WiUP. However, the bad quality of WiUP cannot be identified with traditional methods. In this paper, we propose an intelligent identification method using unsupervised machine learning. More precisely, we create an algorithm model based on historical data to realize feature extraction and clustering. The most similar cluster to those cells with bad WiUP is identified according to Euclidean distance. The experiment is conducted on the basis of a large amount of historical data. With several contrast experiments, Simulation results show that the method proposed achieves the accuracy of identification of bad WiUP over 93%. The study manifests that unsupervised machine learning is effective in identifying bad WiUP in wireless networks.

Keywords: Wireless user perception · Feature extraction · Clustering · Intelligent identification · Machine learning

1 Introduction

With the development of communication technologies [1–6] in last decades, users pay more attention to their service quality. It has always been an important means for operators to improve their competitiveness to improve the call quality by reducing the drop rate and increasing the coverage rate. Recently, bad wireless user perception (WiUP) appears in some cells, which cannot be identified effectively by key performance index (KPI) or key quality index (KQI). The research motivations come principally from user complaints.

The direct measurement of WiUP involves human participation and it requires a lot of time and effort. To improve quality of experience (QoE), a lot of methods were proposed theoretically [7, 8], which ignores the practical problems in real system [9, 10]. However, WiUP is highly reflected in the practical data. dimensionality reduction [11] is considered one of effective methods to realize the data visualization and lower computational complexity. For example, paper [12] uses principal component analysis (PCA) to achieve dimensional reduction of text features, but PCA based methods cannot explain the complex polynomial relationship between features. Supervised machine learning methods [13] have been used in the classification and identification, but unsupervised learning methods are very limited. Sometimes, it is hard to make the label

G. Gui and L. Yun (Eds.): ADHIP 2019, LNICST 301, pp. 507–515, 2019.
https://doi.org/10.1007/978-3-030-36402-1_54

in practical data. Under this context, unsupervised machine learning methods are better. Recently,

In this paper, we propose an effective WiUP identification method using unsupervised machine learning. Specifically, we create an algorithm model based on historical data to realize feature extraction and unsupervised clustering [19], and find out the most similar one to those with bad WiUP by Euclidean distance. During the process, dimensionality reduction should be paid more attention, we train an Auto-encoder neural network [20] for dimensionality reduction. Auto-encoder neural network for dimensionality reduction discards the disadvantage of linear algorithm and improves the identification accuracy by adjusting parameters. In this way, problems can be found and solutions can be formulated in advance. Computer simulations are provided to confirm the proposed method.

The rest of this paper is organized as follows. The system model is given in Sect. 2. Section 3 is the original data description. Section 4 gives these algorithms in this paper including dimensional reduction and clustering. The experimental results are presented in Sect. 5. Section 6 summarizes the full paper.

2 System Model

To realize intelligent identification of WiUP, the algorithm model is divided into three parts. Part I is data reading and preprocessing. Part II is dimensionality reduction and clustering. The last part is cluster identification. The whole system model can be depicted in Fig. 1.

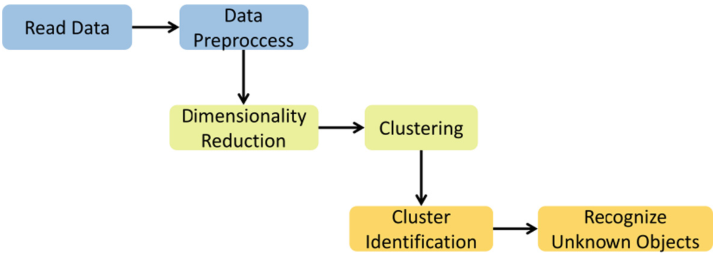

Fig. 1. The system model

3 Data Preprocessing

3.1 Data Description

The data sets are collected from wireless quality monitoring terminals. Table 1 gives the first 4 rows of a data set. Each row in this sheet contains 14 features which reflect the quality of WiUP in each hour. ECI is the ID of a cell. Each cell has 168 rows of data, the data in the whole week. In total, there are more than 10,000 cells. More than 80% of them is unknown historical data of cells and the rest is those identified as ones with bad WiUP. Even though a large number of historical data is available, not all of them is suitable for modeling. The data set need be evaluated in terms of missing value ratio and outlier ratio etc. Before the algorithm model is created, the data needs to be preprocessed, including normalization, correlation analysis [21] and vectorization.

Table 1. The head of the data sheet.

Cell name	ECI	Time	Page response success rate (%)	...	Initial cache delay (ms)
Cell1	126096558	2018-10-16-00	91.10	...	4688
Cell1	126096558	2018-10-16-01	76.32	...	11253
Cell1	126096558	2018-10-16-02	96.68	...	2563

For the sake of simple description, $F_i i = 1, 2, \cdots, 14$ is used to refer to the i-th feature. The map between F and features is listed in Table 2.

Table 2. The map between F and features.

F_i	Features name
F_1	Page response success rate (%)
F_2	Page response delay (ms)
F_3	Page display success rate (%)
F_4	Page display delay (ms)
F_5	Page download rate (kbps)
F_6	Video play success rate (%)
F_7	Video pause time per minute
F_8	Pause time ratio
F_9	Cache time delay (ms)
F_{10}	Stream media rate (kbps)
F_{11}	Instant communication response success rate (%)
F_{12}	Instant communication response delay (%)
F_{13}	Mobile game response success rate (%)
F_{14}	Mobile game response delay (ms)

3.2 Data Preprocessing

Normalization is to eliminate the dimensional influence among different features. And correlation analysis is to eliminate the redundancy among features with high correlation. On the other hand, we need to comprehensively consider the data of the whole day to evaluate the WiUP. Hence, we introduce vectorization to convert the data matrix to a vector.

According to the result of correlation analysis, F_1 is highly correlated to F_3. The high correlation also happens to F_7 and F_9. By this way, we delete F_3 and F_9 to eliminate the redundancy.

The vectorization plays an important role in the algorithm model. The process can be displayed as follows (Fig. 2).

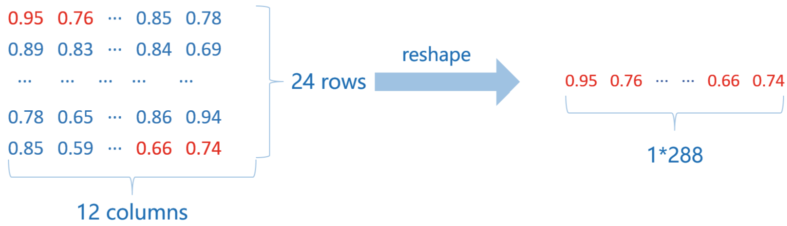

Fig. 2. The process of vectorization.

4 Proposed Methods

4.1 Dimensionality Reduction

Considering the invisibility and high computational complexity in the high dimensional space. Hence, it is necessary to encode the data from high dimensional into low dimensional. PCA is used widely due to its simplicity [22]. We propose a nonlinear generalization of PCA that is an adjustable, Auto-encoder neural network to realize dimensionality reduction. The encoder network transforms the high dimensional data into low dimensional code and the decoder network is responsible to recover the data. Figure 3 gives the structure of this neural network. In Fig. 3, the number in every layer means the number of neurons in this layer. The numbers of neurons and layers are adjustable for higher identification accuracy. w_i, b_i $(i = 1, \ldots, 8)$ represent the weights and deviations of the i-th layer respectively. Both the weights and deviations are updated by back propagation (BP) algorithm.

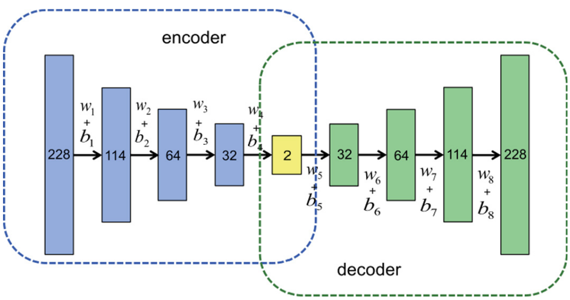

Fig. 3. The neural network for dimensionality reduction.

The neural network is trained with about 1000 pieces of history cells until that the value of the cost function is convergent or steady. The curve of the cost function value in the training process is shown in Fig. 4, and the cost function C is given as below

$$\mathrm{C} = \frac{1}{2n}\sum\nolimits_{i=1}^{n} \left\| y_i^{'} - y_i^{2} \right\| \tag{1}$$

where n equals the number of objects. $y_i^{'}$ and y_i represent the value of output and input of i-th object respectively. $\left\| y_i^{'} - y_i \right\|$ is denoted as the ℓ_2-norm of vector $y_i^{'} - y_i$. ℓ_2-norm is defined as $\|\boldsymbol{v}\|^2 = \sum_{i=1}^{n} |v_i|^2$. The change in a weight is given by

$$w_i^{'} = w_i - \alpha \frac{\partial C}{\partial w_i} \tag{2}$$

where $w_i^{'}$ is the weight changed, α is a learning rate and the $\frac{\partial C}{\partial w_i}$ is the gradient of w_i computed by BP algorithm. In the same way, the change in a deviation is given by

$$b_i^{'} = b_i - \alpha \frac{\partial C}{\partial b_i} \tag{3}$$

where $b_i^{'}$ is the deviation changed, the $\frac{\partial C}{\partial b_i}$ is the gradient of b_i computed by BP algorithm [23]. The model trained is saved to reduce the dimension of other objects, which saves the basic characteristic of the original data.

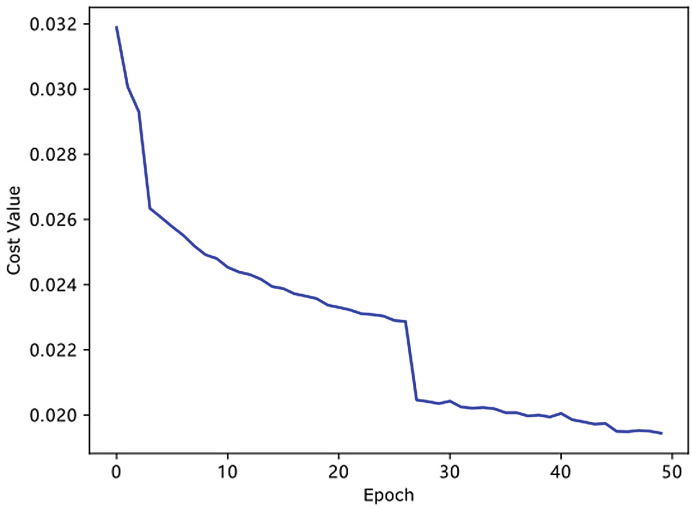

Fig. 4. The change curve of the cost function value in the training process.

In general, the performance of backpropagation deteriorates as the number of hidden layers gets larger. In this paper, we design a neural network with 7 hidden layers. Figure 4 shows the convergence of the neural network after 50 epochs training.

4.2 K-means Clustering

K-means clustering algorithm [24] is one of the most popular algorithms in unsupervised clustering. It owns simple principle and usually works with good results. Suppose that we are given a data set $X = \{x_1, \ldots, x_N\}, x_n \in R^d$. The M-clustering problem aims at partitioning this data set into M disjoint subsets (clusters) $C_1, \ldots, C_M$. The K-means clustering algorithm can be described as following steps:

Step 1: Label the number of clusters.
Step 2: Establish the centroid coordinate.
Step 3: Determine the distance of each object to the centroid.
Step 4: Group the objects based on minimum distance.

K-means clustering uses various distance function to measure the similarity among the objects. The distance function is measured by Euclidean distance [25]. The Euclidean distance between vector X $= (x_1, x_2, \cdots, x_n)$ and Y $= (y_1, y_2, \cdots, y_n)$ can be defined by

$$dist(X, Y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \quad (4)$$

The historical objects after dimensionality reduction are directly input into the k-means algorithm for clustering. The clustering result is shown in Fig. 5. We can observe that 7000 objects from more than 1000 cells are shown in a 2D space after dimensionality reduction. A and B represent the centroids of the two clusters respectively.

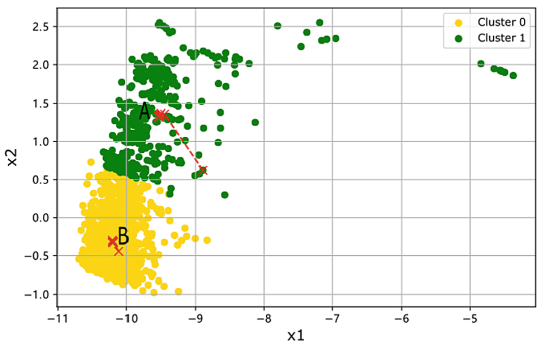

Fig. 5. The result of *K*-means clustering.

4.3 Identification of Target Cluster

Two centroids of clusters are found out by K-means but we cannot identify which cluster is the one with bad WiUP. In this paper, we divide those objects with bad WiUP into test-set and validation-set with ratio 7:3. To identify the target cluster, the validation-set is fed into the neural network for dimensionality reduction and then we compare the Euclidean distance to point A and point B. These objects are grouped into the closer cluster. After clustering, we view the cluster that contains more than 80% objects with bad WiUP as the target cluster. The flow chart is shown in Fig. 6.

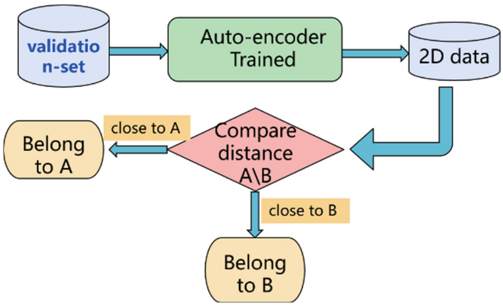

Fig. 6. Flow chart of algorithm for finding target cluster.

5 Experiment Results

5.1 Results Evaluation Method

As for the evaluation method of identification accuracy, we need to feed the test-set into the algorithm model shown in Fig. 6, and finally the proportion of objects correctly classified is output as the identification accuracy.

5.2 Results and Analysis

All 12 features are preserved. In other words, the input data is 228 dimensions. The only variate changed is the number of neurons in the code layer, which means the data is converted into different dimensions to compare the identification accuracy. The results are shown in Table 3. As we can see from the table, the best identification accuracy is achieved when the input is converted to two dimensions. On the other hand, correlation analysis is carried out on the original data to remove the features with high correlation to reduce the input redundancy. We also calculate the variance of all features and eliminate the features with small variance. Table 4 gives the comparison of results with different input dimensions.

Simulation results show that the highest identification accuracy can be achieved by eliminating the input dimensions and converting the data into two-dimension space.

Table 3. Identification accuracy of different dimensions in code layer.

Dimension of input	Dimension of code layer	Number of hidden layers	Recognition accuracy
336	2	7	86.50%
336	10	7	68.53%
336	50	7	61.60%
336	100	7	56.30%

Table 4. Identification accuracy of different input dimensions.

Dimension of input	Dimension of code layer	Number of hidden layers	Recognition accuracy
336	2	7	85.50%
264	2	7	68.85%
228	2	7	62.05%
96	2	7	93.70%

6 Conclusion

This paper has proposed an effective method to identify cells with bad WiUP. The whole algorithm model was created based on a large number of historical data with unsupervised machine learning. The model can be depicted by three parts including data preprocessing, dimensionality reduction and clustering. For higher identification accuracy, many parameters have been adjusted in this model and finally the identification accuracy can increase to 93.7%. The operators can formulate solutions to solve the bad WiUP according the precise identification. However, the identification accuracy need to be further improved. In the further research, the relationship between KQI and KPI should be analyzed in order to find out the basis reasons for bad WiUP.

References

1. Na, Z., Wang, Y., Li, X., Xia, J., Liu, X., Xiong, M.: Subcarrier allocation based Simultaneous Wireless Information and Power Transfer algorithm in 5G cooperative OFDM communication systems. Phys. Commun. **29**, 164–170 (2018)
2. Liu, X., Zhang, X., Jia, M., Fan, L., Lu, W., Zhai, X.: 5G-based green broadband communication system design with simultaneous wireless information and power transfer. Phys. Commun. **28**, 130–137 (2018)
3. Liu, M., Song, T., Gui, G.: Deep cognitive perspective: resource allocation for NOMA based heterogeneous IoT with imperfect SIC. IEEE Internet Things J. 1–11 (2018). https://doi.org/10.1109/jiot.2018.2876152
4. Liu, M., Yang, J., Song, T., Hu, J., Gui, G.: Deep learning-inspired message passing algorithm for efficient resource allocation in cognitive radio networks. IEEE Trans. Veh. Technol. **68**(1), 641–653 (2018)

5. Wu, J., Dong, M., Ota, K., Li, J., Guan, Z.: FCSS: fog computing based content-aware filtering for security services in information centric social networks. IEEE Trans. Emerg. Top. Comput. (2017). https://doi.org/10.1109/TETC.2017.2747158
6. Zhang, N., et al.: Software defined networking enabled wireless network virtualization: challenges and solutions. IEEE Netw. **31**(5), 42–49 (2017)
7. Cicalo, S., Sayadi, B., Faucheux, F., Tralli, V., Kerboeuf, S., Changuel, N.: Improving QoE and fairness in HTTP adaptive streaming over LTE network. IEEE Trans. Circuits Syst. Video Technol. **26**(12), 2284–2298 (2015)
8. Brito, I.V.S., Figueiredo, G.B.: Improving QoS and QoE through seamless handoff in software-defined IEEE 802.11 mesh networks. IEEE Commun. Lett. **21**(11), 2484–2487 (2017)
9. Kato, N., et al.: The deep learning vision for heterogeneous network traffic control: proposal, challenges, and future perspective. IEEE Wirel. Commun. **24**(3), 146–153 (2017)
10. Kato, N.: Challenges of content-centric mobile networks. IEEE Network **31**(1), 2 (2017)
11. Tan, S., Mayrovouniotis, M.L.: Reducing data dimensionality through optimizing neural network inputs. AIChE J. **41**(6), 1471–1480 (1995)
12. Abdulhussain, M.I., Gan, J.Q.: An experimental investigation on PCA based on cosine similarity and correlation for text feature dimensionality reduction. In: Computer Science and Electronic Engineering Conference (CCEC), pp. 1–4 (2015)
13. Li, X., et al.: Joint multilabel classification with community-aware label graph learning. IEEE Trans. Image Process. **25**(1), 484–493 (2016)
14. Fadlullah, Z.M., et al.: State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems. IEEE Commun. Surv. Tutorials **19**(4), 2432–2455 (2017)
15. Wang, Y., Gui, G., Zhao, N., Yin, Y., Huang, H., Li, Y.: Deep learning for optical character recognition and its application to VAT invoice recognition (2018)
16. Huang, H., Gui, G., Sari, H., Adachi, F.: Deep learning for super-resolution DOA estimation in massive MIMO systems. In: 2018 IEEE 88th Vehicular Technology Conference (VTC Fall), Kansas, August 2018
17. Gao, X., Jin, S., Wen, C.K., Li, G.Y.: ComNet: combination of deep learning and expert knowledge in OFDM receivers. IEEE Commun. Lett. **22**(12), 2627–2630 (2018)
18. Wang, T., Wen, C., Wang, H., Gao, F., Jiang, T., Jin, S.: Deep learning for wireless physical layer: opportunities and challenges. China Commun. **14**(11), 92–111 (2017)
19. Lam, D., Wei, M., Wunsch, D.: Clustering data of mixed categorical and numerical type with unsupervised feature learning. IEEE Access **3**, 1605–1613 (2015)
20. Chen, K., Yang, S.: Effect of multi-hidden-layer structure on performance of BP neural network: probe. In: 2012 8th International Conference on Natural Computation, no. 2010, pp. 1–5 (2012)
21. Zhang, Z., Zhao, M., Chow, T.W.S.: Binary-and multi-class group sparse canonical correlation analysis for feature extraction and classification. IEEE Trans. Knowl. Data Eng. **25**(10), 2192–2205 (2013)
22. Agarwal, S., Ranjan, P., Ujlayan, A.: Comparative analysis of dimensionality reduction algorithms, case study: PCA. In: Proceedings of 2017 11th International Conference on Intelligent Systems and Control, ISCO 2017, pp. 255–259 (2017)
23. Liu, Y., Liu, S., Wang, Y., Lombardi, F., Han, J.: A stochastic computational multi-layer perceptron with backward propagation. IEEE Trans. Comput. **67**(9), 1273–1286 (2018)
24. Yang, F., Yang, L., Wang, D., Qi, P., Wang, H.: Method of modulation recognition based on combination algorithm of K-means clustering and grading training SVM. China Commun. **15**, 55–63 (2018)
25. Dokmanic, I., Parhizkar, R., Ranieri, J., Vetterli, M.: Euclidean distance matrices: essential theory, algorithms, and applications. IEEE Signal Process. Mag. **32**(6), 12–30 (2015)