Big Data
Volume 00, Number 00, 2022

DOI: 10.1089/big.2021.0365

Prediction and Big Data Impact Analysis of Telecom Churn by Backpropagation Neural Network Algorithm from the Perspective of Business Model

Jiabing Xu,[1] Jiarui Liu,[2] Tianen Yao,[3] and Yang Li[1,*]

Abstract

This study aims to transform the existing telecom operators from traditional Internet operators to digital-driven services, and improve the overall competitiveness of telecom enterprises. Data mining is applied to telecom user classification to process the existing telecom user data through data integration, cleaning, standardization, and transformation. Although the existing algorithms ensure the accuracy of the algorithm on the telecom user analysis platform under big data, they do not solve the limitations of single machine computing and cannot effectively improve the training efficiency of the model. To solve this problem, this article establishes a telecom customer churn prediction model with the help of backpropagation neural network (BPNN) algorithm, and deploys the MapReduce programming framework on Hadoop platform. Using the data of a telecom company, this article analyzes the loss of telecom customers in the big data environment. The research shows that the accuracy of telecom customer churn prediction model in BPNN is 82.12%. After deploying large data sets, the learning and training time of the model is greatly shortened. When the number of nodes is 8, the acceleration ratio of the model remains at 60 seconds. Under big data, the telecom user analysis platform not only ensures the accuracy of the algorithm, but also solves the limitations of single machine computing and effectively improves the training efficiency of the model. Compared with that of the existing research, the accuracy of the model is improved by 25.36%, and the running time is shortened by about twice. This business model based on BPNN algorithm has obvious advantages in processing more data sets, and has great reference value for the digital-driven business model transformation of the telecommunications industry.

Keywords: business model; BP neural network; telecom users; churn prediction; data mining

Introduction

The rapid development of the Internet has promoted the boom of the telecom industry in China. By 2020, China's mobile phone users have reached 1.596 billion, of which the fourth-generation mobile communication technology (4G) users account for 80.8%, reaching 1.289 billion.[1] With the increase of number of users, the amount of data of telecom operators multiplies. The traditional methods cannot process the emerging big data. Coupled with the rapid development of communication and video voice technology in various Internet social software, the traditional telecom industry has been greatly impacted.

The data mining (DM) method of big data is to search the valuable information in the data with the algorithm, and the technologies involved include statistical data, machine learning, artificial intelligence, and database system.[2] Common DM analysis methods include cluster analysis, regression analysis, correlation analysis, classification, and prediction analysis.[3] The application of DM technology in the telecom industry includes telecom resource fraud monitoring, telecom churn prediction, telecom user classification, and telecom industry precision marketing. Currently, the DM technology is in the theoretical stage, and there are few application cases.[4]

[1]*Yantai Institute of China Agricultural University, Yantai, China.*
[2]*School of Business, The University of Sydney, Sydney, Australia.*
[3]*School of Mathematics, University of Birmingham, Birmingham, England.*

*Address correspondence to: *Yang Li, Yantai Institute of China Agricultural University, Yantai 264670, China,* E-mail: ylcherry17@cau.edu.cn

The rapid development of telecom technology has intensified industrial competition. To develop new users while retaining the old, Chinese telecom operators have vigorously adjusted business models to improve their competitiveness.[5] Under the background of big data, the prediction methods of telecom churn can be divided into two categories. The first method predicts the telecom churn by combining gray prediction and logistic regression algorithms with management knowledge.

The second method combines a classification algorithm with DM technology and builds a churn risk prediction model. At present, there are few studies on the churn risk prediction by the two prediction methods, and the accuracy of the existing models is not high.[6] Here, the DM technology is combined with the second prediction method to establish a prediction model for telecom users and accurately predict the telecom churn.

Here, a prediction model is constructed to promote the business model transformation of telecom operators and cope with both opportunities and challenges faced by telecom operators, combined with DM analysis under the background of big data. First, relevant research and background are introduced, and the research methods are described. Second, the model data are prepared and the telecom churn prediction model is established. Finally, the model is simulated and the experimental results are analyzed. The innovation is to treat data processing as the core of the prediction model. The research results provide a reference for the business model transformation of telecom operators in the era of big data.

Using backpropagation neural network (BPNN) algorithm, the prediction model of telecom customer churn is established, and the telecom customer churn in big data environment is analyzed. The innovations of this study include the following. (1) With the deployment of large data sets, the learning and training time of the model is greatly shortened. (2) The acceleration ratio of the model remains stable. In the big data environment, the telecom user analysis platform not only ensures the accuracy of the algorithm, but also solves the limitations of single machine computing and effectively improves the training efficiency of the model.

(3) The accuracy of the model is improved and the running time is shortened by about twice. This business model based on BPNN algorithm has obvious advantages in processing more data sets, and has great reference value for the digital-driven business model transformation of the telecommunications industry.

Related Works

Internationally, Internet technology has developed earlier, and data extraction technology is more mature than in China. In the research of behavior prediction, Xu et al. proposed a prediction model of user purchase behavior based on information fusion and integrated learning.[7] In the research of the telecom field, Syed et al. constructed and verified the telecom churn model through the combination of relevant algorithms with artificial neural network (ANN) and decision tree.[8] Zheng et al. made a detailed classification of users based on users' reputation, consumption level, and functional services.[9] In addition, Alrwashdeh et al. established a specific churn prediction model with strong applicability.[3]

Under the huge impact of the Internet, the traditional telecom industry began to explore new big data extraction methods. Liu et al. co-operated with well-known coffee companies to collect users' location information and calls and established a user prediction model based on the sample data, thus predicting customers' shopping behavior.[10] Su et al. analyzed the user data collected after the combination of well-known mobile phone brands and Internet social software and grasped the user experience.[11] Youssouf found that the sense of user experience was related to many aspects, including user call quality, telecom network layout, and call drop rate.[12]

Domestically, the development of user prediction is first implemented by Internet enterprises. As the head enterprise of China's search engine, Baidu has a huge database and powerful data processing ability. After data collection and analysis, Baidu forecasts much public domain information. Other big data processing methods are structured, which makes it easier to extract valuable information. Noticeably, the DM technology of China Mobile, as one of the three major operators in China, has developed well, and the company has done abundant research in improving customer experience. Zhang et al. proposed an intelligent management system to provide countermeasures for user complaints, including the content analysis of complaints and problem solving.[13]

Guan et al. used clustering algorithm to divide telecom enterprise customers and studied the strategic model of Guangdong Telecom Market from the perspective of virtual enterprise.[14] Kunle et al. constructed a user prediction model combined with algorithm knowledge to identify users.[15] Olga et al. established a

prediction model by studying rough set theory, which played an important role in the effective extraction of telecom user tests.[16]

Although the mentioned research methods have analyzed churn prediction, most of them are theoretical and lack actual applications, and the model data may not be accurate enough. Therefore, based on the results of previous studies, a model of telecom churn is established. In DM and data analysis, a multilayer backpropagation (BP) network with structural advantages is introduced. Combined with the Hadoop platform, the parallel processing of the user prediction model and platform is completed. The results have reference significance for big data extraction of telecom enterprises.

Theory Introduction and Prediction Model Establishment of Telecom Churn

Institutional review board

The main investigator of the trial is ultimately responsible for the selection of subjects, the implementation of clinical trials, data analysis, and the final transfer of data to the sponsor or publication in medical journals. The rights of subjects must be respected. This is the core of all human trials. We recognize that this test should be avoided if the hazard is unpredictable. If the harm is greater than the benefit, the test shall be suspended.

DM method

The cross-industry standard process for DM (CRISP-DM) model proposed by Non-Conformance Report and Statistical Product and Service Solutions company is the most commonly used DM method. This model provides a complete process description for cross-industry DM, including the business understanding stage for business objectives determination, the data understanding stage for data collection, the data preparation stage for data analysis and conversion, the modeling stage for model parameter calibration, the evaluation stage of feasibility analysis for high-quality models, and the final deployment stage for converting the results into readable text form.[17] Here, the CRISP-DM process is used for DM, and the specific DM process is shown in Figure 1.

Figure 1 illustrates that the six stages of CRISP-DM are linked by data. Where data appear, the work begins, so the order between model stages is not fixed. The arrow in Figure 1 represents the connection between the stages.

Introduction of neural network algorithm

Neural network (NN) algorithm can be combined with DM technology to establish a churn prediction model. Because the ANN is inspired by the central nervous system of animals, it works like the human brain and connects information through neurons.[18] NN can express the complex nonlinear relationship. The hidden relationship of data can be found through weight adjustment between nodes. Through model training, NN can be applied to different information processing requirements.[19]

The structure of the core part of the NN is shown in Figure 2.

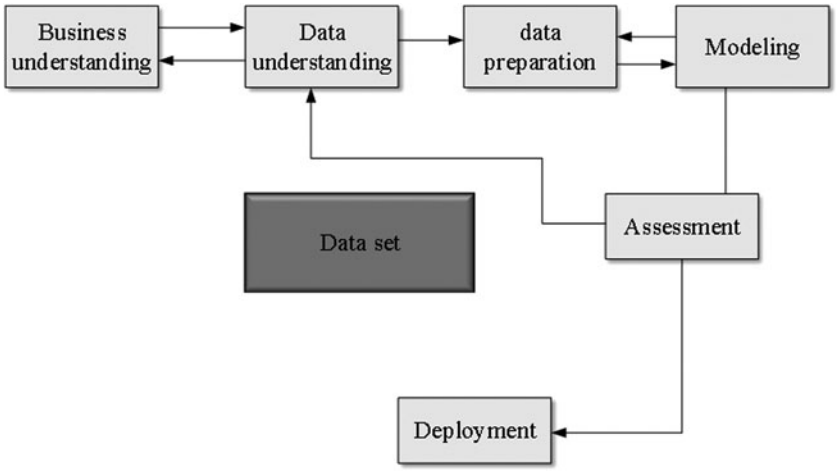

FIG. 1. The flow of the data mining model.

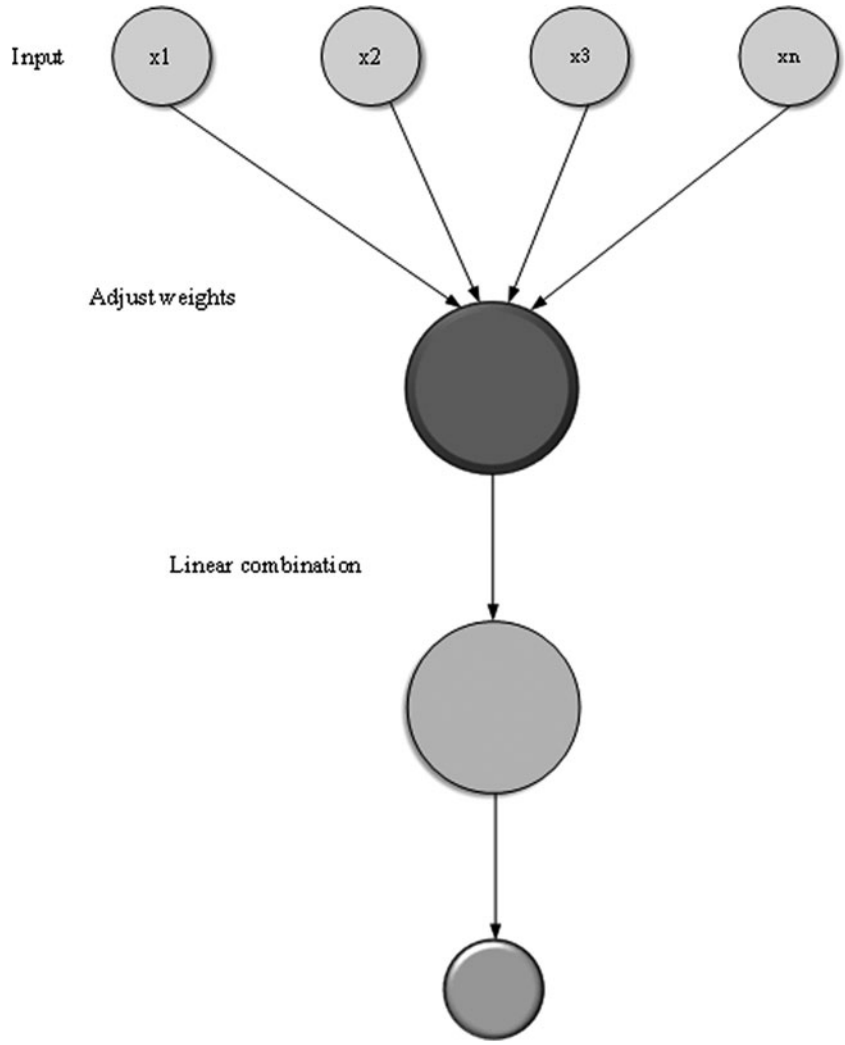

FIG. 2. Structure of neural network perceptron.

Figure 2 displays that the core working part of the NN is based on neurons. After the input is received, the sensor processes the data, the threshold is set, and, finally, the relationship between the processing data and the threshold is compared. In this process, the output of the sensor is calculated as in Eq. (1).[20]

$$f(x_1, x_2, \ldots, x_n) = \begin{cases} 1 & \text{if} B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_n x_n > 0 \\ -1 & \text{othewise} \end{cases} \quad (1)$$

Equation (1) indicates that if the sensor processing data are greater than the threshold, the output is 1, and if the processing data are less than the threshold, then output is 0. In Eq. (1), x represents the input value, B denotes the weight value, and x_0 stands for the bias value. Since the bias value has little effect on the learning process of the NN, it is generally ignored.[21]

The structure of NN generally includes the input layer, output layer, and hidden layer. After the data input, the NN is calculated layer by layer and is propagated forward. After data output, the result is compared with the expected value of the sample. According to the error, the network parameters are modified to fit the calculation results accurately to most of the samples. Sigmoid function and hyperbolic

tangent function are the activation functions of network neurons throughout the process,[22] which are expressed as in Eqs. (2) and (3).

$$f(y) = \frac{1}{1 + e^{-y}}, \tag{2}$$

$$f(y) = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}. \tag{3}$$

In Eq. (2), the range of y is 0–1, and in Eq. (3), the range of y is-1-1. Both functions can simulate the interaction between neurons and describe the output relationship between levels in the NN structure.[23]

Introduction of Hadoop platform

Hadoop is a distributed system infrastructure used for distributed computing and storage of large-scale data. Hadoop implements a distributed file system. Hadoop Distributed File System (HDFS) is one of the components, which is featured by high fault tolerance and provides storage for massive data. Apart from HDFS, MapReduce is also the core design of the Hadoop framework, and MapReduce provides the calculation for data.[24] Here, the telecom churn prediction model is built with Hadoop.

HDFS is mainly used for data storage, and the nodes in the system are divided into working nodes and management nodes. The management node maintains the namespace of the entire HDFS and manages the entire file system, so it is the core of the HDFS, and the fault-tolerant backup mechanism is exactly what the management node needs. The working node mainly calculates, stores data, and regularly reports the data block file name to the management node.

HDFS data read and write process is shown in Figure 3.

Figure 3 demonstrates that the client, HDFS, working node, and management node interact with each other to complete the reading and writing process. The main process is as follows. First, the client receives the user command for reading and sends it to depth first search (DFS). Second, DFS accesses the management node to

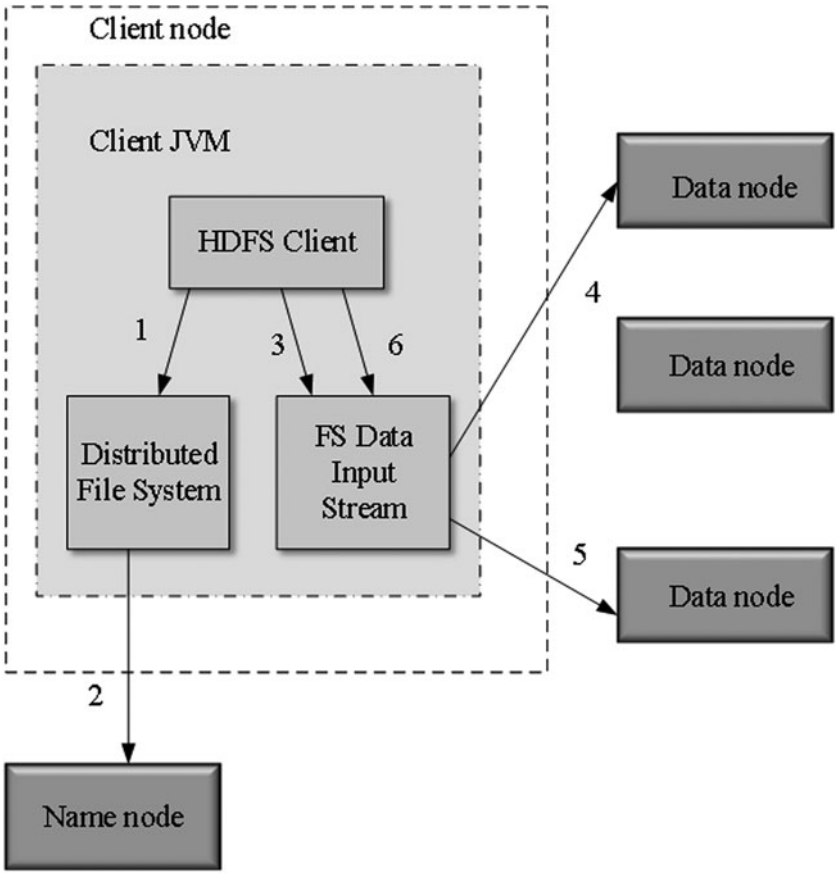

FIG. 3. HDFS process of reading and writing file. FS; HDFS, Hadoop Distributed File System; JVM.

obtain the file location. The client calls the input flow. Finally, the working node reads the input flow and transmits it to the client.

MapReduce distributes data and computing work to the data center to parallelly run data and programs. MapReduce is easy to operate but needs to be optimized to give full play to its function. Input data, relevant configuration, and MapReduce program constitute a whole set of MapReduce workflow. Nodes are classified as job trackers and task trackers. The job tracker node manages and schedules the daemons node, and the daemons node completes and reports the tasks. When a task fails, the job tracker node will inform other daemons to execute the task until the task is completed.[25] MapReduce has the same data storage mechanism as HDFS.

Establishment of telecom churn prediction model

Model establishment preprocessing. Here, combined with the first three steps of CRISP-DM, the data are selected and preprocessed to obtain the sample data set.

First, the stage of business understanding. Most of the reasons for churn are untraceable. Therefore, this part of churn without early warning is specifically studied here. Based on the mentioned content, users without consumer behavior for three consecutive months are defined as churn. The users of telecom enterprises generally include broadband users, mobile phone users, and equipment users. Here, the telecom churn prediction model is constructed only for individual users in mobile phone users.

Next, the data understanding phase. Here, user data behavior, complaint information, voice behavior, account status, and value-added business behavior are collected for churn prediction. Specifically, the data of a telecom company from July to October 2020 are selected. The users who consistently have value-added services, data, and calls from July to October are regarded as retained users, and those without value-added services, data, and calls in October are regarded as churn users. The data from July to September are mainly used to predict the data in October, and the average numerical value of the 3 months is chosen.

Based on the consultation of the staff of telecom enterprises, various information of telecom users is collected, including user payment, complaint information, package accumulation, value-added behavior information, and user's personal information.[26] The information obtained after data sort-out includes the basic attributes of the product instance, customer information, mobile product package accumulation, mobile user value-added behavior-related attributes, mobile user voice behavior-related attributes, mobile user data behavior information, user account information, and customer complaint-related attributes. Data cover a wide range and strive to comprehensively describe user data. The test rule for data follows the principle that 1 indicates yes, and 0 indicates no.[27]

Finally, the data preparation stage involves data sets from preprocessing to input the whole process of the model. Data should be integrated, transformed, cleaned, and selected, and these operations may be executed multiple times without sequence before the final requirements are met because of the high data quality requirements. Data integration combines data in a consistent data storage. Here, the data sources are the electronic design automation system, billing system, and customer relationship management system of a telecom company. The integrated data include all the data in the data understanding stage. In data integration, data value conflict occurs, which is caused by different attributes of data source, errors in pattern integration, duplication in tuple level data, and data redundancy.

Here, the script program is chosen to extract data in the data set processing, and 2,913,440 users are extracted, of which 2,789,783 retained users and 123,647 churn users. Then, data cleaning is conducted mainly for inconsistent data and noise. Data cleaning mainly deals with data missing values, outliers, and noise. Here, the method of deleting attributes and filling mean is used to correct the data missing. In this study, only statistical outliers with gross errors are eliminated according to certain rules. When the data are seriously missing, it will have a great impact on the analysis results.

Therefore, reasonable methods should be used to fill in the excluded abnormal values and missing values. The common methods include average filling, etc. The statistical description technology and the method of deleting outliers are selected for the noise processing in abnormal values. Data protocol processes the data set specification, but it does not destroy the data integrity. Here, the data specification method adopts irrelevant data deletion, churn information, and attributes of missing values. For the attribute value specification, the data are simplified, and the operation of attribute value specification is illustrated through roaming duration.

The obtained data set is transformed, and the construction attribute, the discretization value, and the aggregated data are selected for the transformation

method. In the process of modeling, data should be normalized to increase data independence. Equation (4) is used for data normalization.[28]

$$x'_i = \frac{x_i - \min_x}{\max_x - \min_x}(\underset{x}{nmax} - \underset{x}{nmin}) + \underset{x}{nmin}. \quad (4)$$

Equation (4) normalizes maximum and minimum values for attributes. In Eq. (4), x denotes the numerical property, $\min_x$ represents the minimum, $\max_x$ represents the maximum, x'_I stands for the mapping value of x on the interval [$nmax_x$, $nmin_x$].

Zero-mean normalization uses the standard deviation and the mean of attribute x for data conversion, as shown in Eq. (5).

$$x'_i = \frac{x_i - \bar{x}}{\varphi x}. \quad (5)$$

In Eq. (5), φ represents the standard deviation, zero-mean normalization has the best application effect when the maximum and minimum values of attributes are unknown, and the maximum and minimum values are affected by outliers. Here, two methods (maximum and minimum normalization and zero-mean normalization) are chosen to transform data attributes, and the attribute values are scaled to 0–1.

The classification method directly affects data balance. Based on previous experience and the above user data, 108,395 churn users are obtained using the oversampling method, and 327,210 retained users are obtained by the down-sampling method, totaling ~430,000 users. After the two sets of data are balanced, the ratio of retained users to churn users is 3:1.

At present, the existing correlation evaluation methods usually focus on the new analysis of redundancy between attributes and the evaluation standard of maximum correlation and minimum redundancy. Finally, this article designs two attribute selection algorithms based on attribute correlation analysis. One is to eliminate the redundancy between attributes. It uses decision-independent correlation and decision-dependent correlation to measure the correlation between attributes and class attributes and the redundancy between attributes, respectively. The other is the combination of sorting method and packing method, which is a two-stage method. First, the ranking method uses the maximum correlation and minimum redundancy criteria to select some better attribute subsets, and then the packaging method uses cross-validation to select the best attribute subset.

Here, the model sample attributes are selected from two parts: correlation analysis and attribute importance analysis. The final important attributes selected include 18 attributes, such as mobile Internet time, total telephone time, traffic growth rate, and balance.

The last step of preprocessing splits the data set into the test set and training set. Here, the split ratio is test set: training set = 3:7.

Finally, a total of 435,515 sets of data are input into the user prediction model, with 18 attributes, 130,655 test sets, and 304,860 training sets.

Establishment of telecom churn prediction model. Based on the mentioned content, the telecom churn prediction model is established. Here, a convolutional neural network (CNN) model is chosen, which is divided into three layers: an input layer, an output layer, and a hidden layer.

First, the network weights are initialized to establish the output unit and hidden unit containing the expected number. Afterward, the target output is set as t, and each piece of data in the training set is processed with the following steps.

The first step is to input data x into the network and calculate the output of each unit in the network. The operations can be expressed as in Eqs. (6) and (7).

$$Y_j = f\left(\sum_{i=0}^{N} B_{ij} X_i\right), \quad (6)$$

$$Z_k = f\left(\sum_{j=0}^{M} B_{jk} Y_i\right). \quad (7)$$

Equations (6) and (7) propagate the input along the NN, in which B_{ij} represents the weight of the *i*th neuron in the input layer to the hidden layer unit *j*, B_{jk} denotes the weight of the *j*th neuron in the hidden layer to the output unit *k*, Z_k denotes the actual value of the unit *k*, X_i represents the input of the input unit *I*, and the Y_j stands for the output of the hidden layer unit j.

The second step is the error BP along the network. The output unit and its error term are calculated, and then the weight is adjusted reversely, as expressed in Eqs. (8)–(11).[29]

$$B_{jk} \leftarrow B_{jk} + \Delta B_{jk}, B_{jk} = \eta \varphi_k Y_j, \quad (8)$$

$$B_{ij} \leftarrow B_{ij} + \Delta B_{ij}, B_{ij} = \eta \varphi_j X_i, \quad (9)$$

$$\varphi_k = Z_k(1 - Z_k)(t_k - Z_k), \quad (10)$$

$$\varphi_j = Y_j(1 - Y_j)\sum_{k\in outputs} B_{jk}\varphi_k. \quad (11)$$

In Eqs. (8)–(11), φj represents the error term of hidden layer unit j, φk denotes the error term of output layer unit k, ΔBij indicates the weight change from input unit i to hidden layer unit j, ΔBjk stands for the weight change from hidden layer unit j to output unit k, η represents the learning rate, and t_k is the target value of output unit k.

After the mentioned operations, the calculated error value is compared with the set value. If the error value is greater than the set value, the calculation is carried out again in the first two steps until the error value is less than the set value.

In the actual case, the iterative weights can reduce the training time of the model. The updating methods of weights are shown in Eqs. (12) and (13).[30]

$$\Delta B_{ij}(n) = \eta\varphi_j X_i + \alpha\Delta B_{ij}(n-1), \quad (12)$$

$$\Delta B_{jk}(n) = \eta\varphi_k Y_j + \alpha\Delta B_{jk}(n-1). \quad (13)$$

In Eqs. (12) and (13), α represents the impulse coefficient, and the two operations are repeated until termination conditions are met.

Parameter setting of the model. Here, the number of nodes in the input layer is 18, the hidden layer is tansig function, and the output layer is logsig function. The number boundary of hidden layer nodes is determined by Eq. (14).

$$\frac{n_i + n_0}{2} \le n_h \le (n_i + n_0) + 10. \quad (14)$$

In Eq. (14), n_i represents the number of input layer nodes, n_0 denotes the number of output layer nodes, and n_h indicates the number of hidden layer nodes.

Here, Eq. (14) and the trial-and-error method are combined to train the network. The value of the hidden layer starts from a small value, and the weights are updated 5000 times. Finally, the optimal number of units in the hidden layer is 15.

Here, there is only one output node: either 1 or 0. The learning rate is between 0.01 and 0.8, and the impact coefficient is 0.9.

Network training and simulation experiment. The network model is trained according to the mentioned contents. The maximum number of iterations is set to 1000 times in the training. The error of the iterative results is lower than the set value after 774 iterations. Fifty thousand records are extracted from the training samples to observe the churn distribution.

Finally, the constructed model is evaluated, and the evaluation items include the hit rate, accuracy, and coverage of the model, as expressed in Eqs. (15), (16), and (17), respectively.

$$H = \frac{m}{v}, \quad (15)$$

$$I = \frac{f}{u}, \quad (16)$$

$$J = \frac{g}{w}. \quad (17)$$

In Eqs. (15), (16), and (17), f denotes the number of users predicted and actual churn user, u represents the number of predicted churn users, J indicates the accuracy of the model, g stands for the number of accurately predicted users, and w is the total number of users. H represents the predicted coverage, m denotes the number of accurately predicted churn users, v represents the actual number of churn users, and I represents the predicted hit rate.

In addition, the evaluation items in the simulation experiment include the gain index and the ascension index of the model, as calculated in Eqs. (18) and (19).

$$K = (\frac{h}{p}) \times 100\%, \quad (18)$$

$$L = (\frac{r}{q}). \quad (19)$$

In Eqs. (18) and (19), K represents the gain, h denotes the success number at the quantile, p indicates the total success number, L stands for the ascension index, r is the churn rate at the quantile, and q represents the average churn rate. An ascension index >1 suggests that the model is effective. The larger the ascension index is, the stronger the predictability of the model is.

Implementation of prediction model Hadoop platform. A Hadoop platform is established to predict the telecom churn with high efficiency and high scale. The experimental platform is composed of nine servers. The physical host model is IBM System×3635 M3, the memory is 6 * 4G, the storage space is 2 * 1T, and the software is Hadoop 0.20.2.

The CNN algorithm is parallelized using MapReduce programming, and the efficiencies are compared. The comparison includes the operation efficiency of 2, 4, 6, and 8 nodes, and 50,000, 500,000, 1 million, and 12 million data sets. Finally, the acceleration ratio is introduced to characterize the efficiency of the model. Chen et al. verified the accuracy and training speed of the model algorithm through comparison with the CNN model, classification data set pretraining model, and YOLO algorithm.[31] Hence, the result part compares the proposed model with the existing models.

Results Analysis of Telecom Churn Prediction Model

Results analysis of model data preprocessing stage

Here, the total number of field calls is specifically studied to describe the method of deleting outliers, and the results after deleting outliers are shown in Figure 4.

Figure 4 shows that 99% of the users' call times per month remain <800, and the numeric value 15527 in the collected data is an outlier, and after the outlier is removed the call times chart changes normally. Ninety percent of the users have 453 calls per month, 75% have 201 calls per month, 50% have 134 calls per month, and the call times per month are at least one. This frequency is the call times made per month by 1% of users.

The attribute value specification operation is illustrated with roaming duration, as shown in Figure 5.

Figure 5 illustrates that the roaming duration is mainly concentrated in 40–100 seconds, and the churn is between 0.2 and 0.76. When the roaming duration is >100 seconds, the churn begins to decrease, and, finally, approaches 0. Significantly, when the roaming duration passes 150 seconds, the churn ratio drops sharply, and few users are lost, which can be ignored. Overall, when the roaming duration is longer than 100 seconds, the churn ratio approximates 0. According to the analysis, users who roam for >100 seconds have high stickiness and are not easy to lose.

Model network training results

The number of hidden layer nodes is determined as shown in Figure 6.

Figure 6 displays the number of hidden layer nodes determined here is between 10 and 29. When the number of hidden layer nodes reaches 15, the sum of error squares gradually stabilizes and decreases from 0.05 to

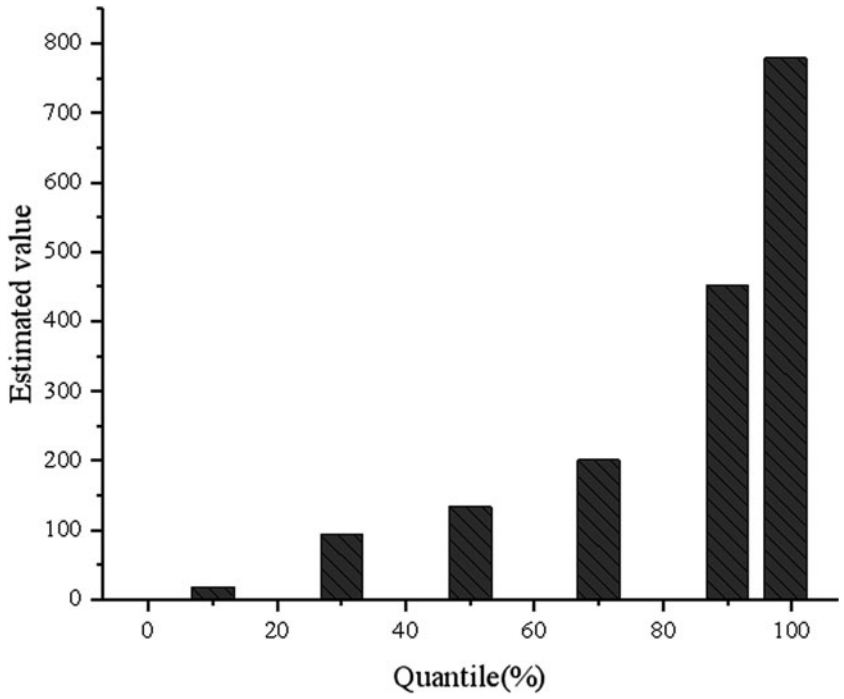

FIG. 4. The change chart of call times.

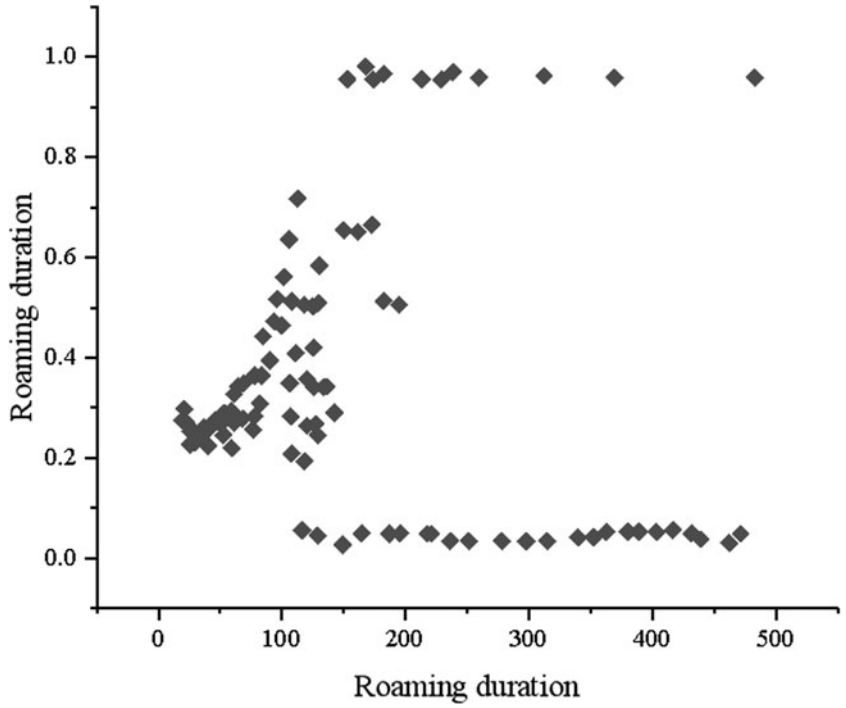

FIG. 5. Churn ratio distribution of roaming call duration. The diamonds indicate the roaming time test point.

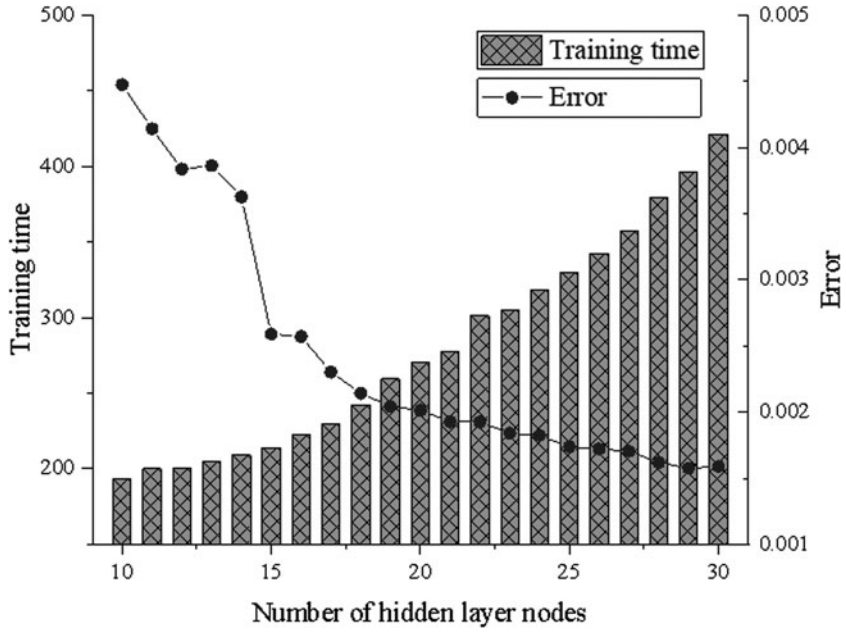

FIG. 6. Comparison of error square sum of hidden layer nodes after 5000 iterations.

0.015. The training time of the model is also short when the number of hidden layer nodes is 15, which is ~220 seconds. Hence, the number of hidden layer nodes in the proposed model should be set to 15.

The distribution of telecom churn is shown in Figure 7.

Figure 7 implies that the churn is distributed in two obvious partitions, mainly distributed between 0 and 35,000 records, and the churn degree is between 0.05 and 0.61. Although the second partition is between 35,000 and 50,000 records, and the churn degree is between 0.61 and 1. Hence, the churn boundary set here is reasonable, ~0.61. In the network model, a churn degree >0.61 is judged to be churn, and a churn degree <0.61 is judged to be retained.

The comparison of hit rate, coverage rate, and model accuracy between the test set and training set is shown in Figure 8.

Figure 8 demonstrates that the three indexes in the training set are generally higher than those in the test set, indicating that the training results for the network in this experiment are good. The model accuracy of the training set is 89.6%, the network hit rate is 85.0%, and the network coverage rate is 76.0%. The model accuracy of the test set of the telecom churn prediction model reaches 89.3%, the hit rate is 82.12%, and the coverage rate is 78.2%. Thus, the evaluation results of common indexes in the proposed model are good, and there is no overfitting. Compared with that of the existing models, the accuracy of the proposed model has increased by 25.36%.

The evaluation results of the model gain index and ascension index are shown in Figure 9.

Figure 9 suggests that the gain index of the model reaches ~82.6% after the gain rate reaches 20%, and the gain index reaches the maximum when the gain rate is 60%, which is 94.3%, and then stabilizes. The highest improvement index is 2.96, proving that the churn predictability is improved ~3 times with the proposed prediction model. Meanwhile, the lowest improvement index is 1, indicating that the worst result with the proposed prediction model equals the original predictability.

Prediction model Hadoop platform building results

The comparison of experimental results of different size data sets in different nodes is shown in Figure 10.

Figure 10 shows that the running time of different nodes in different size data sets is different. The results

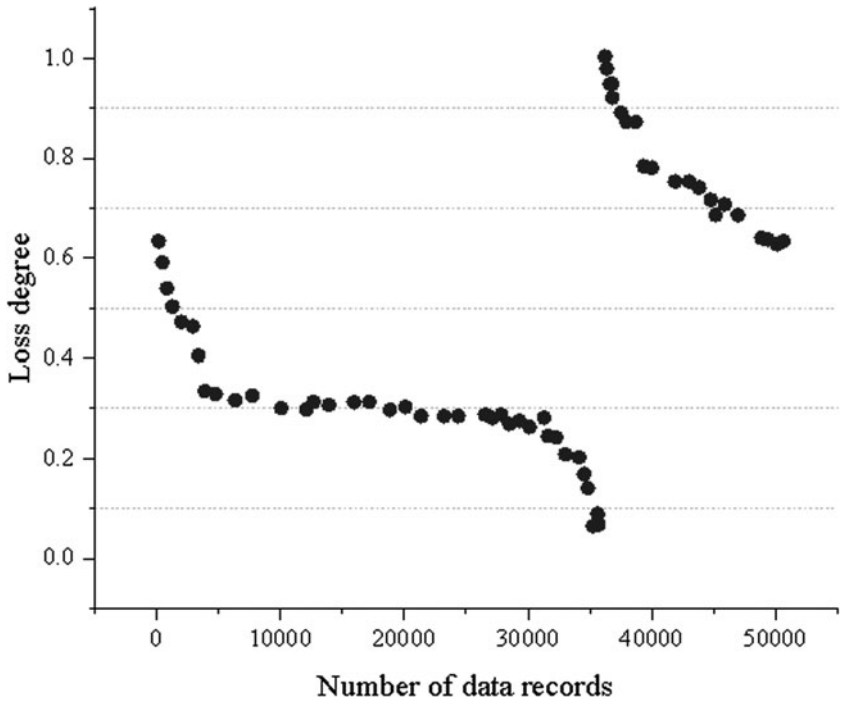

FIG. 7. The distribution of churn degree. The circles indicate the test point of telecommunication loss.

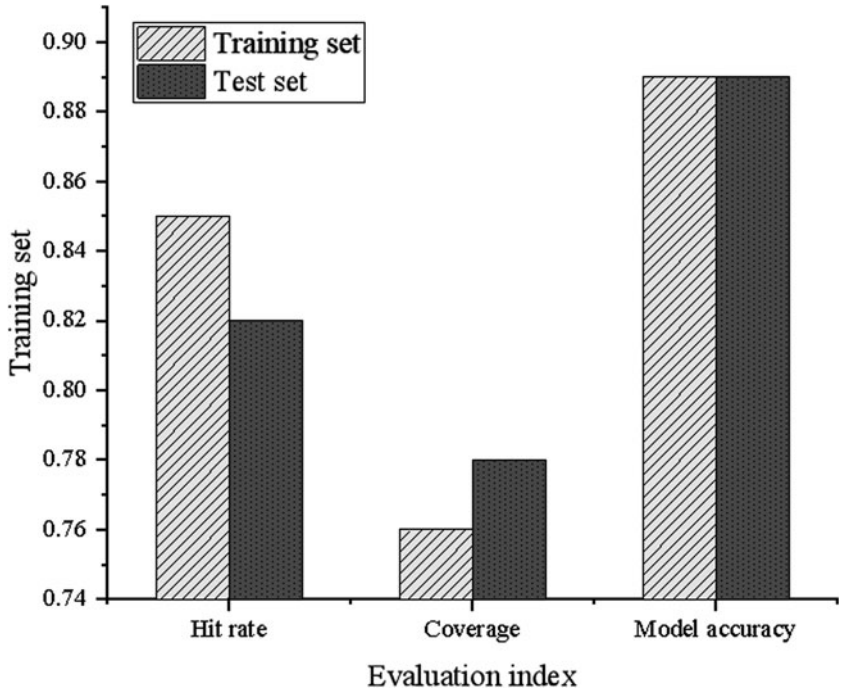

FIG. 8. Comparison results of common evaluation indexes.

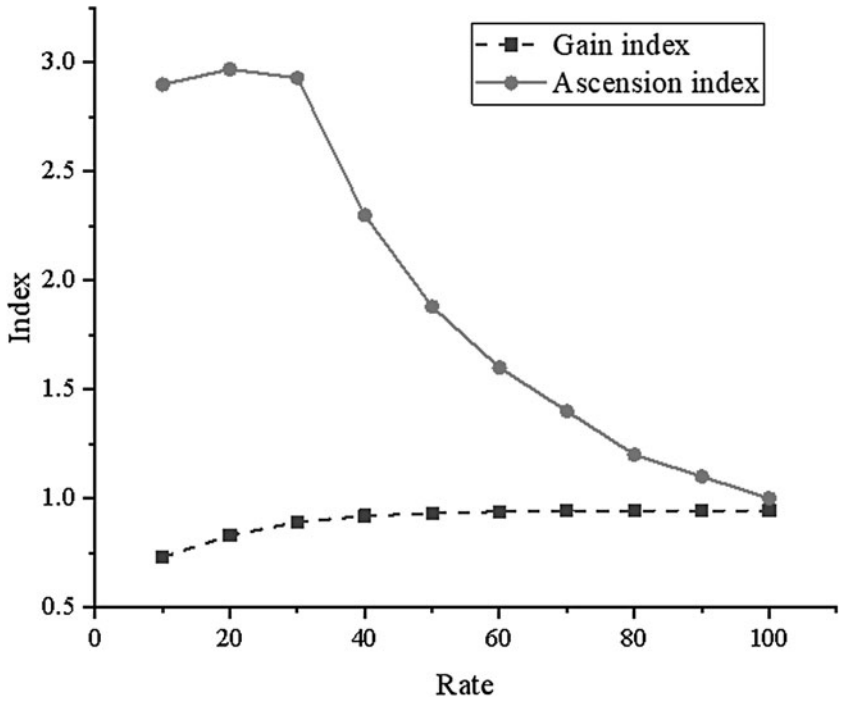

FIG. 9. Evaluation results of gain index and ascension index.

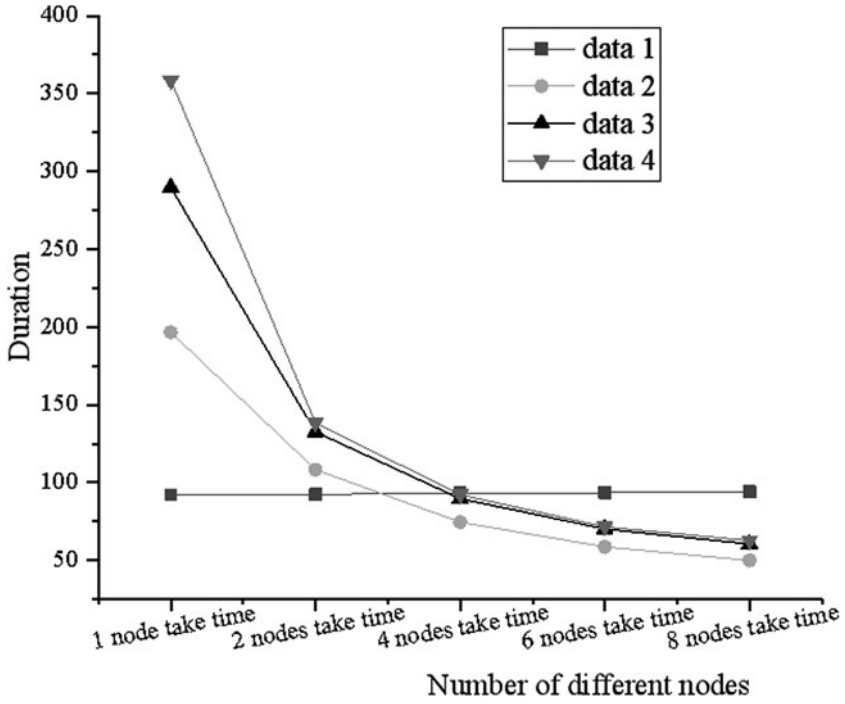

FIG. 10. Running time comparison of prediction network models.

of data set 1 in different number of nodes are almost the same, and the running time is ~100 seconds. The running time of data sets 2, 3, and 4 shortens with the increase of the number of nodes. When the number of nodes of data set 2 is 8, the running time is the shortest, 49.89 seconds. Compared with that of the existing models, the running time of the proposed prediction model is reduced to about two times.

Finally, the acceleration ratio of 8 nodes is studied, and the results show that the acceleration ratio is ~60 seconds, which shows that the proposed telecom churn prediction model has obvious advantages.

Conclusion

To predict telecom customer churn, a telecom customer churn prediction model based on NN is constructed. The proposed prediction model is better than the existing models. With the deployment of large data sets, the learning and training time of the model is greatly shortened. The acceleration ratio of the model remains stable. In the big data environment, the telecom user analysis platform not only ensures the accuracy of the algorithm, but also solves the limitations of single machine computing and effectively improves the training efficiency of the model.

The accuracy of the model is improved and the running time is shortened by about twice. This business model based on BPNN algorithm has obvious advantages in processing more data sets, and has great reference value for the digital-driven business model transformation of the telecommunications industry. But there are also some shortcomings. The method of user feature extraction is relatively simple. In the future, we can conduct in-depth research on user feature extraction to make user information more comprehensive and the research results more convincing.

Data Availability Statement

The data used to support the findings of this study are included within the article.

Author Disclosure Statement

The authors declare that they have no competing interest.

Funding Information

No funding was received for this article.

References

1. Chandra SP. Customer switching behavior towards mobile number portability: A study of mobile users in India. Int J Cyber Behav Psychol Learn. 2020;10:31–46.

2. Kumar NS. Network slicing and SDN: New opportunities for telecom operators. CSI Trans ICT. 2020;8:15–20.
3. Alrwashdeh M, Jahmani A, Ibrahim B, et al. Data to model the effects of perceived telecommunication service quality and value on the degree of user satisfaction and e-WOM among telecommunications users in North Cyprus. Data Brief. 2020;28:104981.
4. Khedra MM, Abd EA, Hamdi H, et al. A novel framework for mobile telecom network analysis using big data platform. Int J Adv Comput Sci Appl. 2020;11:0110863.
5. Deng WB, Deng LS, Liu J, et al. Sampling method based on improved C4.5 decision tree and its application in prediction of telecom customer churn. Int J Inform Technol Manag. 2019;18:93–109.
6. Rahouma KH, Ali A. Applying machine learning technology to optimize the operational cost of the egyptian optical network. Proced Comput Sci. 2019;163:502–517.
7. Xu J, Wang J, Tian Y, et al. SE-stacking: Improving user purchase behavior prediction by information fusion and ensemble learning. PLoS One. 2020;15:e0242629.
8. Syed IM, Hanif A, Jamal FQ, et al. Towards successful business process improvement—An extension of change acceleration process model. PLoS One. 2019;14:e0225669.
9. Zheng YJ, Zhou XH, Sheng WG, et al. Generative adversarial network-based telecom fraud detection at the receiving bank. Neural Netw. 2018;102:78–86.
10. Liu Y, Zhao Q, Yao W, et al. Short-term rainfall forecast model based on the improved BP-NN algorithm. Sci Reports. 2019;9:19751.
11. Su Y, Guo Y, Huang D. Parameter optimization based BPNN of atmosphere continuous-variable quantum key distribution. Entropy. 2019;21:e21090908.
12. Youssouf EA. Advanced prediction of learner's profile based on Felder-Silverman learning styles using web usage mining approach and fuzzy c-means algorithm. Int J Comput Aided Eng Technol. 2019;11:495–512.
13. Zhang M, Xie F, Zhao J, et al. Chinese license plates recognition method based on a robust and efficient feature extraction and BPNN algorithm. J Phys Conf Ser. 2018;1004:012022.
14. Guan S, Wang X. Research on telecom operator's market strategy change from virtual value chain view: Take Guangdong Telecom Company Limited as an example. E3S Web Conf. 2021;235:01062.
15. Kunle L, Ganiyu RA, Nkechi P. An investigation of brand equity dimensions and customer retention: A perspective of postpaid telecom subscribers in Lagos State, Nigeria. Int J Manag Econ. 2020;56:339–350.
16. Olga T, Tcukanova O, Torosyan E, et al. Methods of end-to-end automatization for the telecom company. IOP Conf Ser Mater Sci Eng.2020;940:012096.
17. Czaplewski M. Internet and its impact on the market position of telecom operators. SHS Web Conf. 2018;57:01007.
18. Eugen S, Seppe VB, Katrien A, et al. Profit maximizing logistic model for customer churn prediction using genetic algorithms. Swarm Evol Comput. 2018;40:116–130.
19. Li YX, Hou BZ, Wu Y, et al. Giant fight: Customer churn prediction in the traditional broadcast industry. J Business Res. 2021;131:630–639.
20. Li MX, Yan C, Liu W, et al. An early warning model for customer churn prediction in telecommunication sector based on improved bat algorithm to optimize ELM. Int J Intell Syst. 2021;36:3401–3428.
21. Devriendt F, Berrevoets J, Verbeke W. Why you should stop predicting customer churn and start using uplift models. Inform Sci. 2021;548:497–515.
22. Samira K, Mahmoud ZR. Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior. J Syst Inform Technol. 2017;19:65–93.
23. Adnan A, Sajid A, Awais A, et al. Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing. 2017;237:242–254.
24. Arivazhagan B, Subramanian RS. Customer churn prediction using logistic regression with regularization and optimization technique. Int J Innovat Technol Explor Eng. 2020;9:334–339.
25. Arno DC, Kristof C, Koen DB, et al. Incorporating textual information in customer churn prediction models based on a convolutional neural network. Int J Forecast. 2020;36:1563–1578.
26. Alae C, Hassane IEH. Deep convolutional neural networks for customer churn prediction analysis. Int J Cognit Informat Natural Intell. 2020;14:1–16.
27. Muhammad BAJ, Rameshwar AJ, Khalid MBA. Customer churn prediction in telecom using machine learning in big data platform. J Crit Rev. 2020;7:1991–2001.
28. Sivasankar E, Vijaya J. Hybrid PPFCM-ANN model: An efficient system for customer churn prediction through probabilistic possibilistic fuzzy clustering and artificial neural network. Neural Comput Appl. 2019;31:7181–7200.
29. Zhu B, Baesens B, Broucke SV. An empirical comparison of techniques for the class imbalance problem in churn prediction. Inform Sci. 2017;408:84–99.
30. Eria K, Marikannan BP. Significance-based feature extraction for customer churn prediction data in the telecom sector. J Comput Theor Nanosci. 2019;11:3428–3431.
31. Chen Y, Hu S, Mao H, et al. Application of the best evacuation model of deep learning in the design of public structures. Image Vision Comput. 2020;102:103975.

Cite this article as: Xu J, Liu J, Yao T, Li Y (2022) Prediction and big data impact analysis of telecom churn by backpropagation neural network algorithm from the perspective of business model. *Big Data* 3:X, 1–14, DOI: 10.1089/big.2021.0365

Abbreviations Used

4G = fourth-generation mobile communication technology
ANN = artificial neural network
BP = backpropagation
BPNN = backpropagation neural network
CNN = convolutional neural network
CRISP-DM = cross-industry standard process for DM
DM = data mining
HDFS = Hadoop Distributed File System
NN = neural network