Comparison of Main Algorithms in Big Data Analysis of Telecom Customer Retention

Yuanhu Gu¹, Thelma Domingo Palaoag² and Josephine S. Dela Cruz³

University of the Cordilleras, Baguio City, Philippines

E-mail: ¹tiger18921884431@gmail.com, ²tpalaoag@gmail.com, ³delacruzpen@gmail.com

Abstract. In today's fierce telecommunications market competition, customer churn is very severe. In order to retain customers, telecommunications companies have made various attempts from various data and consumption characteristics analysis to big data analysis. However, since the actual situation of customer churn is very complicated, how to predict customer churn accurately and quickly is a difficult problem. After the researchers successfully conducted big data analysis of customer churn and successfully retained customers, in this article, the researchers mainly compared several commonly used algorithms in order to find a better algorithm for big data analysis of telecommunications customer churn. Compare and analyze the accuracy and efficiency of these several algorithms and suggest that the business support staffs of telecommunications companies adopt major methods for big data analysis. The researcher found that the Decision Tree (CART) algorithm is better for the prediction of customer churn and guided other branch staffs to predict customer churn and retain customers in a timely manner. This kind of big data analysis can be used to retain customers in the telecommunications industry.

1. Introduction

In the 21st century, the telecom industry has developed rapidly. It took 8 years for China to go from 3G to 4G and only 4 years to 5G. But too fast development has also brought strong competition. When all telecom operators provided similar services, the competition became fierce. Especially in recent years, some telecom companies have found that the number of customers has even shown a negative growth. Obviously, many customers are churned. Customer churn means the loss of customers from their company [1]. It is often difficult to recover the churn customers because these they have already used the services of competitors. With the development of big data analysis, researchers have found that using big data analysis can better judge customer churn called Customer Churn Prediction [2].
Big data analysis has a significant impact on the development of modern society [3]. It is increasingly difficult for telecom companies to get new customers, so maintaining existing customers has become particularly important [4]. Using big data analysis to predict customer churn can help telecom companies know in advance which customers may be churn [5] and take certain preventive measures. At present, churn prediction has become one of the main challenges for telecom companies. [6] It is difficulty in modeling customer churn prediction. Dimensionality reduction is an important aspect of considering algorithm efficiency [7]. However, it is also more complicated to determine the reduced dimension.

Although there are many difficulties in predicting customer churn, we can study what algorithm is more suitable for Customer Churn Prediction and train the research results to more data support personnel to perform daily big data analysis. Customer Churn Prediction has been performed in the literature using various techniques including machine learning, data mining, and so on. First of all, let us review the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, KNN, Naive Bayes, and CART. These 10 algorithms are one of the most influential data mining algorithms in recent years [8]. One of the commonly used tools in data mining is C4.5, C5.0 and CART. The decision tree algorithm is cost-sensitive, while improving the robustness [9], [10]. According to the literature, these algorithms can be used for big data analysis and decision-making of our telecom customer churn. It strikes a good balance between accuracy and cost. In machine learning, support vector machines (SVM) must be tried [8]. K-nearest neighbor (KNN) classification is a more complex method [11]. KNN is suitable for low-dimensional data, and it can help decrease computational costs [8]. Naive Bayes method is a very important method. A general discussion of the Naive Bayes method and its advantages is given in [12], [13]. The Naive Bayes model is simple, elegant and robust [14], [15]. After referring to many documents, the researchers know that many algorithms need to be studied and will focus on the impact of these algorithms on the performance and accuracy of telecommunication customer churn analysis. The researchers will mainly compare C4.5, K-means, SVM, KNN, Naive Bayes and Cart algorithms, and find more practical algorithms, so that daily analysis becomes fast and accurate. These algorithms should be simple and easy to use so that more colleagues can analyze various telecommunication services.

The researcher is engaged in data analysis in a branch of China Telecom. Our company is mainly engaged in telecommunications-related services, including fixed-line, mobile, broadband and various value-added services. Among them, the mobile phone business has a high churn rate and fierce competition, which is also our key business. Therefore, the analysis of mobile phone customer churn must be timely and accurate to effectively retain customers. There are many algorithms in big data analysis. What algorithm is better to use? This is a problem that puzzles many data analysts. In order to obtain the best algorithm for a certain problem, so as to train and guide the support staffs at all levels of the company, the researchers analyzed and compared several main algorithms, in order to find a more accurate and faster algorithm to guide various business support staffs conduct big data analysis. The researchers collected 221,770 user data through the support of the superior company, of which 11,282 are churn customers, involving 37 features. This article uses the same data to compare the accuracy and operating efficiency of the following algorithms: DecisionTree (C4.5, CART), KNN,
Linear, SVM (SVC, SVR), Naive Bayes (GaussianNB, BernoulliNB), RandomForest, Neural Network. The research analyzed train accuracy, test accuracy, AUC and runtime of each algorithm. Finally, it is found that it is best to use Decision Tree (CART) to predict customer churn. The researcher shared the research results with colleagues in each branch and trained and guided other staffs to predict customer churn and retain customers in a timely manner.

2. Methodology

For the data of telecom customer churn, what algorithm can improve the prediction accuracy and operating efficiency? The researcher used the same data to evaluate the accuracy and efficiency of different algorithms. Several commonly used algorithms hope to be found to guide various business support personnel in big data analysis. This article mainly studies the accuracy and efficiency of several main algorithms. Figure 1 shows the research method.

First of all, the telecommunications business is very large, with all kinds of data. Data needs to be collected from different departments and platforms. Provincial big data is stored on the Hadoop platform and downloaded to the local ORACLE database every month using ETL tools. The data from the local external department is extracted from the relevant platform through DB link and the program written by JAVA. The data mainly includes user attributes, consumption data, and consumption characteristic data, such as customer age, consumption, online time, online time period, online traffic and other 37 kinds of data.

Secondly, the researcher use python to develop big data analysis program on each algorithm before and after cleaning up abnormal data. The researcher mainly compares the 7 algorithms mentioned in introduction. Record the training accuracy and test accuracy, AUC and running time of each algorithm, and then assess. The efficiency and accuracy of predictions through different algorithms are tested. After repeated predictions with different parameter, evaluations and optimizations, the most accurate and fast algorithm will be found.

The researcher used the data before and after data cleaning to test each algorithm separately. The original data has a lot of null values, and there are data with extremely large values, such as Internet traffic and deposits, which need to be cleaned up. Data cleaning mainly includes cleaning up null values and abnormal values. The purpose of data cleaning is to enhance and maintain the quality of the data [16]. This article also discusses a quick method of clearing abnormal values; simply and easily adjust abnormal data (outliers) to the non-abnormal range.
Finally, the researcher evaluates the prediction results by evaluating the accuracy and efficiency of each algorithm, AUC and running time. In order to improve execution efficiency, the expected test accuracy rate should be at least 80%, the higher the better, the AUC is greater than 0.8, and the shorter the running time, the better. Of course, it also need to consider the complexity of the algorithm and the simplicity of training and use. Through EXECL comparison, after finding the best algorithm, the researcher will train and guide support personnel at all levels to conduct big data analysis in order to react quickly to retain customers.

3. Result and Discussion

3.1. Background
In order to compare the pros and cons of different algorithms for big data analysis of telecom customer churn, the researchers prepared a period of data of churn users with a data volume of 221,770 pieces, which were divided into 161,770 pieces of training data and 60,000 pieces of test data. When the data dimensions are very large, some algorithms will be very difficult to establish an effective model. This is called the "curse of dimensionality" [17]. Feature models are often used to model variability [18]. In this study, the researcher has done a lot of feature analysis and multi-feature analysis based on experience and selects 37 features.

3.2. Data Cleansing
The researcher used the data before and after cleansing to compare the accuracy of each algorithm. There are many out-of-range data in the consumption data of telecom users. These data are called outliers. In order to better analyze big data, outliers should be replaced. In order to better represent the relationship between data sizes, the researchers linearly narrowed the abnormal data to the normal range. The following is batch processing method.

The IQR is from 1 as follows:
\[ IQR = Q_3 - Q_1 \]  
(1)
Q1 is the first quartile, Q3 is the third quartile, Q3 - Q1 is called the Inter-Quartile Ranger (IQR).

The 1*IQR+Q3 is from 2 as follows:
\[ Q_5 = IQR + Q_3 = 2Q_3 - Q_1 \]  
(2)
Q5 is 1*IQR+Q3. The researchers treat data larger than Q5 as outliers and narrow down this type of data to make it within the inner fence. The formula is from 3 as follows:
\[ Q = (Q - Q_5)/(Q_{\text{max}} - Q_5) * IQR/2 + Q_5 \]  
(3)
Qmax is the maximum value of the column. The researchers first replace the null value with an appropriate value, and then convert data beyond Q5 to within 1.5* IQR+Q3. As we all know, 1.5*IQR+Q3 is inner fence.

This is a simple, batch, and fast way to handle outliers. After the researchers’ test, it can better improve the prediction accuracy of big data.

3.3. Customer Churn Prediction of Each Algorithm
The researcher used the data before and after data cleansing to test each algorithm separately, and carefully adjust the parameters of some algorithms in order to obtain a better test accuracy and recall rate. Each algorithm is tested and evaluated as below.

3.3.1. Decision Tree C4.5. The researcher first used the simpler C4.5 algorithm, and the training accuracy was 85%, the test accuracy was 83%, the AUC (AUC stands for "Area under the ROC Curve.") was 0.78, and the ROC (Receiver Operating Characteristic) curve is shown in Figure 2.

3.3.2. Decision Tree CART. The researcher again uses the CART algorithm in the decision tree, and the training accuracy is 86%, the test accuracy is 82%, the AUC is 0.81, and the ROC curve is shown in Figure 3. This shows that the C4.5 and CART algorithms of decision tree are relatively similar, and the difference is not very big.

3.3.3. KNN. The researcher continued to use the same data and used KNN to make predictions. The training accuracy rate was 89%, the test accuracy rate was 80%, and the AUC was only 0.72, as shown in Figure 4. This shows that the training situation is good, but the test situation is not ideal. And the AUC is low, indicating that there is over fitting.
3.3.4. Linear Regression. The researcher also tested the linear regression algorithm. The training accuracy and test accuracy are both low, as shown in Figure 5. Such prediction results cannot be used. This may indicate that the complexity of telecommunications customer churn data is very large, and it is difficult to use linear regression algorithms to have better prediction results.

![Figure 6. SVM SVC (C=1 gamma=1)](image)

![Figure 7. SVM SVC (C=4 gamma=0.1)](image)

3.3.5. SVM. The researcher used SVM to predict the telecom customer churn data. When taking the default value C=1 gamma=1, as shown in Figure 6, the training accuracy rate is 95%, the test accuracy rate is 79%, but the AUC is only 0.59. After trying to adjust the parameters, the best result is shown in Figure 7, where C=4 and gamma=0.1. The training accuracy rate is 86%, the test accuracy rate is 84%, and the AUC is 0.73. When using the SVR algorithm, the best training accuracy rate is only 36%. Although after a lot of adjustments, the test effect is very poor, as shown in Figure 8.

![Figure 8. SVM SVR (C=1 gamma=0.03)](image)

![Figure 9. Naive Bayes (before cleansing)](image)

3.3.6. Naive Bayes Then, the researchers studied the Naive Bayes algorithm. Using the GaussianNB algorithm, the effect is shown in Figure 9. The training accuracy rate is 69%, the test accuracy rate is 62%, and the AUC=0.78, the accuracy rate is relatively low. After cleansing the outlier data, the researchers compared the GaussianNB and BernoulliNB algorithms. The results are almost same only
GaussianNB shown in Figures 10 and Figures 11. The test accuracy has increased by 10%, indicating that the cleaning of abnormal data is very important. The effect of these two algorithms is not much different, but the accuracy is relatively low. The comparison shows that data cleansing is very important to Naive Bayes algorithm. In addition, using this algorithm must first deal with null values.

![Figure 10. Naive Bayes GaussianNB](image1)

![Figure 11. Naive Bayes BernoulliNB](image2)

3.3.7. Random Forest. The researcher used RandomForest to perform big data analysis. The training accuracy rate is very high, reaching 99%, and the test accuracy rate is also 80%, with AUC=0.76, as shown in Figure 12, indicating that the RandomForest algorithm has a high over fitting phenomenon. After trying several times, AUC from 0.76 to 0.80 has a relatively large range of variation, indicating that the results of each operation of the RandomForest are significant different, which proves that it has a random result, as shown in Figure 13. Maybe this algorithm is suitable for scientific research, but not very suitable for enterprise production.

![Figure 12. Random Forest (1)](image3)

![Figure 13. Random Forest (2)](image4)

3.3.8. Neural Network. The researcher finally tried the Neural Network algorithm. As shown in Figure 14, the training accuracy rate is 83%, the test accuracy rate is 80%, and AUC=0.78. This is closer to the decision tree algorithm, but the execution time is much longer.

![Figure 14. Neural Network](image5)
3.4. Algorithms Comparison

The researcher compared these algorithms, as shown in Table 1. Researcher believes that it is better to use the decision number (CART) method for big data analysis such as telecom customer churn. This method is simple and practical, and the efficiency and accuracy are high. Although the accuracy of SVM is high, the AUC is not high, and it takes too long to adjust the parameters for repeated training. RandomForest has also obtained good accuracy and AUC, but the outliers need to be sorted in advance. The results of each operation of RandomForest algorithm are quite different, and it is necessary to try to find better prediction results many times. The neural network also has good accuracy and AUC, but it takes too long.

![Figure 14. Neural Network](image)

Table 1. Algorithms comparison

| Algorithm        | Train_Accuracy | Test_Accuracy | AUC | Runtime |
|------------------|----------------|---------------|-----|---------|
| DecisionTree C4.5| 85%            | 83%           | 0.78| 14:33   |
| DecisionTree CART| 86%            | 82%           | 0.81| 14:00   |
| KNN              | 89%            | 80%           | 0.72| 15:42   |
| Linear/Regression| 38%            | 24%           | 0.82| 26:17   |
| SVM SVC(C=1, gamma=1) | 95% | 79%           | 0.59| 17:34   |
| SVM SVC(C=4, gamma=0.1) | 86% | 84%           | 0.73| 13:52   |
| SVM SVC(C=1, gamma=0.001) | 36% | 22%           | 0.8 | 12:32   |
| Naive Bayes(GaussianNB) (outlier) | 69% | 62%           | 0.78| 14:53   |
| Naive Bayes(GaussianNB)       | 74%            | 72%           | 0.8 | 10:30   |
| Naive Bayes(GaussianNB)       | 73%            | 73%           | 0.76| 14:46   |
| RandomForest (1)            | 99%            | 80%           | 0.76| 15:41   |
| RandomForest (2)            | 98%            | 81%           | 0.8 | 14:45   |
| Neural Network              | 83%            | 80%           | 0.78| 43:28   |

In short, through these experiments and analysis, the researchers found that the CART algorithm is most suitable for big data analysis of telecommunications companies retaining customers, because it has better performance in accuracy and running speed. In addition, the algorithm is simple, does not require complex parameter adjustments, and does not require too high technology. It is suitable for data support personnel at all levels to perform big data analysis.

Finally, telecommunications companies can use the K-means clustering method [19] to classify churn customer and introduce corresponding policies to retain customers.

4. Conclusion

The analysis of customer churn is a particularly difficult problem faced by telecom companies in the fierce market competition. When data acquisition, outlier processing, and analysis using various big data algorithms become daily tasks, it is necessary to further study how to conduct big data analysis accurately and quickly. In this paper, the researcher first discussed a quick and practical method of data cleaning, and then compared and analyzed several major algorithms in order to obtain better prediction results, and further train and guide analysts in our company. After comparing the main algorithms, the researcher found that the Decision Tree (CART) algorithm is most suitable for predicting customer churn. The algorithm is fast and accurate. The accuracy of the model on the test set reached 82%. In addition, the CART algorithm is simple and does not require too high technology, so it is suitable for data support personnel at all levels to quickly analyze big data for customer churn. For telecommunications companies, professional personnel can be used to find suitable algorithms for
specific problems to train and guide support personnel at all levels to perform big data analysis. Through this process it can save time, standardize operations, and respond quickly to the market to retain customers.

References

[1] Maji G, Mandal S, Bhattacharya S and Sen S 2017 Designing combo recharge plans for telecom subscribers using itemset mining technique 2017 IEEE Int. Conf. on Industrial Technol. (ICIT) DOI: https://doi.org/10.1109/ICIT.2017.7915539

[2] Amin A, Anwar S, Adnan A, Nawaz M and Huang K 2017 Customer churn prediction in the telecommunication sector using a rough set approach Neurocomputing vol 237 no 10 pp 242–54 DOI: https://doi.org/10.1016/j.neucom.2016.12.009

[3] Iqbal R, Doctor F, More B, Mahmud S and Yousuf U April 2018 Big data analytics: computational intelligence techniques and application areas Technological Forecasting and Social Change 153 p 119253 DOI: https://doi.org/10.1016/j.techfore.2018.03.024

[4] Abd-Allah M N, El-Beltagy S R and Salah A 2017 DyadChurn: Customer Churn Prediction using Strong Social Ties IDEAS 2017: Proc. of the 21st Int. Database Engineering & Applications Symp. pp 253–63 DOI: https://doi.org/10.1145/3105831.3105832

[5] Tiwari A, Sam R, Shaikh S Analysis and prediction of churn customers for telecommunication industry 2017 Int. Conf. on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) pp 218–22 DOI: https://doi.org/10.1109/I-SMAC.2017.8058343

[6] Abd-Allah M N, Salah A and El-Beltagy S R 2014 Enhanced customer churn prediction using social network analysis DUBMOD ’14: Proc. of the 3rd Workshop on Data-Driven User Behavioral Modeling and Mining from Social Media pp 11–2 DOI: https://doi.org/10.1145/2665994.2665997

[7] Li H, Yang D, Yang L, Lu Y and Lin X Oct 2016 Supervised massive data analysis for telecommunication customer churn prediction 2016 IEEE Int. Conf. on Big Data and Cloud Computing (BDCloud) pp 163–9 DOI: https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.35

[8] Wu X et al 2008 Top 10 algorithms in data mining Knowl. Inf. Syst. 14 pp 1–37 DOI: https://doi.org/10.1007/s10115-007-0114-2

[9] Han L, Li W and Su Z 2019 An assertive reasoning method for emergency response management based on knowledge elements C4.5 decision tree Expert Systems With Applications 122 pp 65–74 DOI: https://doi.org/10.1016/j.eswa.2018.12.042

[10] Y Chen L, C Wu C and Tang K 2016 Time-constrained cost-sensitive decision tree induction Information Sciences vol 354(1) pp 140–52 DOI: https://doi.org/10.1016/j.ins.2016.03.022

[11] Tao D, Li X, Wu X, Hu W and Maybank S J 2007 Supervised tensor learning Knowl. Inf. Syst. 13(1) pp 1–42

[12] Domingos P and Pazzani M 1997 On the optimality of the simple Bayesian classifier under zero-one loss Mach. Learn. 29 pp 103–30

[13] Hand D J and Yu K 2001 Idiot’s Bayes—not so stupid after all? Int. Stat. Rev. 69 pp 385–98

[14] Friedman N, Geiger D and Goldszmidt M 1997 Bayesian network classifiers Mach. Learn. 29 pp 131–63
[15] Ridgeway G, Madigan D and Richardson T 1998 Interpretable boosted naive Bayes classification Proc. of the Fourth Int. Conf. on Knowledge Discovery and Data Mining ed R Agrawal et al (California: AAAI Press) pp 101–4

[16] Alotaibi S B 2017 ETDC: An efficient technique to cleanse data in the data warehouse ICAIP 2017: Proc. of the Int. Conf. on Advances in Image Processing pp 135–8 DOI: https://doi.org/10.1145/3133264.3133296

[17] Elitedatascience 2017 Dimensionality Reduction Algorithms: Strengths and Weaknesses Retrieved from https://elitedatascience.com/dimensionality-reduction-algorithms

[18] L A Zaid, F Kleinermann and O D Troyer 2011 Feature Assembly Framework: towards scalable and reusable feature models VaMoS ’11: Proc. of the 5th Workshop on Variability Modeling of Software-Intensive Systems pp 1–9 DOI: https://doi.org/10.1145/1944892.1944893

[19] W Qadadeh and S Abdallah 2018 Customers segmentation in the insurance company (TIC) dataset Procedia Computer Sci. vol 144 pp 277–90 DOI: https://doi.org/10.1016/j.procs.2018.10.529