Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/49/20241167

An overview study of bank customer classification

Qianying Shao

Department of Marketing and E-Commerce, Nanjing University, Nanjing, 210023, China

201098312@smail.nju.edu.cn

Abstract. Under the trend of combining the banking industry with the Internet, bank customer classification based on data analysis is important for enterprises to conduct accurate marketing and thus increase revenue. Based on the current situation, this overview study takes banking customer classification as the theme, examining the relevant research in this field between 2013 and 2023. The characteristics of customer classification in this industry, the focuses of existing research, and the future direction of attention are summarized. Due to the large volume and relative multidimensionality of customer data involved in the banking industry, most of the classification models are built on the basis of the standard model, which is improved to make the operation more efficient and accelerate the speed of convergence. On this basis, this paper proposes that bank customer classification should be improved in the algorithmic model while pay more attention to its operating effect based on real-world scenarios. Meanwhile, feature engineering, which plays an important role in data mining, is attracting attention, and more research may be carried out in this direction in the future. In addition, the research on customer segmentation dynamics is important but seldom addressed, which is an area for deeper cultivation.

Keywords: Banking Industry, Customer Classification, Customer Segmentation, Data Mining, Data Analysis.

1. Introduction

Since the first decade of this century, the Internet financial business, represented by mobile payment, has been developing rapidly, penetrating into many subfields of finance and attracting wide attention from the society. At the same time, traditional finance seems to have lost the favor of the public. Taking the central bank data as an example, China's RMB deposits in the banking system decreased by 940.2 billion yuan in January 2014, which means a year-on-year increase of 2.05 trillion yuan less. Under the powerful impact of the Internet, traditional commercial banks have been forced to carry out a deep reform of fintechization. According to Gong [1], banks should incorporate this emerging financial model in the following aspects in order to break through: using the internet to expand banking customers and marketing channels; adjusting interest rate levels dynamically through the open market, etc. At the same time, those organizations should also capitalize on their past capital accumulation, data collection and high credibility during the process. Considering these industry characteristics as a starting point, banks can utilize data mining techniques - a technology oriented towards business applications, with the main task of discovering hidden patterns from large amounts of data through a variety of methods including classification, clustering, regression, and so on [2]. Customer segmentation is one of the typical business

DOI: 10.54254/2755-2721/49/20241167

problems it can solve. Based on the classified customer groups, enterprises can have more access to the needs of customers and their potential value, and carry out personalized marketing and service for different groups, ultimately achieving the ultimate goal of maximizing profits and sustainable development. It can be deduced that under the trend of fintechization, banks can utilize their past databases for data mining, so as to optimize customer management and expand sales, promoting profitability [3].

Wei and Ling mention that before 2013, there was relatively little research on the practical application of big data in the banking industry globally, implying that the techniques of data mining were not often improved by applying them to the classification of bank customers at that time [4]. They also mentioned that many practitioners and researchers had already proposed the possibility of big data applications at that time. This may indicate that studies combining the banking industry and customer segmentation have begun to appear since then. In this paper, researches published between 2013 and 2023 are collected. The keywords "bank customer classification" and "bank customer segmentation" are used during the selection in the published papers database of China Knowledge Network Infrastructure (CNKI) and Google Scholar. After summarizing and reflecting on the results, this paper discusses the characteristics and current status of data mining in this field and speculates on possible future research directions.

The article is organized as follows. In the second part, data mining and customer classification are briefly described, and other concepts are well defined. The difficulties in bank customer classification are also discussed. In the third section, a literature review is presented for the mainstream customer classification algorithms, including K-means clustering and artificial neural networks (ANN). In the fourth part, there is a discussion on feature engineering, a current promising research direction. The last part is the conclusion.

2. Concept definition

This review study is carried out in response to published studies on customer classification in the banking sector. Banks in this context mainly refer to commercial banks, whose traditional business is focused on areas such as operating deposit and loan businesses. In the process of conducting business, institutions usually use information such as transaction flow and product sales records to keep customer transaction data. They also keep users' personal information. At the same time, with the added influence of Internet finance on traditional banking, banks are also able to count a wide range of personalized data including customers' transaction channels, risk preferences and product holdings, thus establishing the system of customer relationship management (CRM). With the support of such data, banks have accumulated the initial capital to conduct precision marketing.

Customer classification has become an important element of customer relationship management in commercial banks, but it may still be limited. For example, some institutions are still differentiating the importance of customers based on the number of their financial assets, which is one-dimensional and does not help to provide targeted services. In contrast, a customer classification system involving big data can synthesize customer characteristics for more detailed segmentation. Depending on the type of customer, banks can differentiate their marketing to increase revenues, for example, by placing products through the right distribution channels, creating more popular investment products. In addition, such customer segmentation allows institutions to focus on "long-tail" customers that were previously ignored, gaining the new revenue-enhancing possibilities.

Some scholars may think that the meaning of "customer classification" and "customer segmentation" are not exactly the same. In this view, customer classification is more macroscopic, mainly through data mining and other techniques to classify customers into several categories based on the similarity of their attributes. Today, the concept of it is more akin to the original concept of customer segmentation, which was introduced by Smith Wendell in 1956 [5]. To contrary, the latter is nowadays more often used in practical situations - it comes with systems such as CRM that are commonly used by modern companies - and refers to the classification of customers into several predefined and interpretable categories. The ABC classification is used, for example, to differentiate between customers with different levels of

DOI: 10.54254/2755-2721/49/20241167

importance in order to facilitate future targeted marketing campaigns. In the research field, however, scholars rarely distinguish between the concepts involved. They tend to mix the two terms and use them to refer to the meanings referred to by customer classification above. Based on this fact, this literature review is done on the basis of considering both as synonyms. Firstly, both keywords were used in the thesis collection phase. Secondly, in the following narrative, customer classification and customer segmentation are used to convey the same meaning, depending on the choice of words used in the cited papers.

The figure 1 shows the evolution of the number of papers found on CNKI using the keywords "bank customer classification" or "bank customer segmentation". This is only a selection from 2013 and as the year 2023 has not yet come to an end, the latest studies are not included in the statistics. As can be seen, the average number of publications in the last five years has risen, indicating that this area is gaining attention from researchers. It can be seen from the figure that the average number of relevant studies published has risen over the last five years, indicating that this area is gaining attention from researchers.

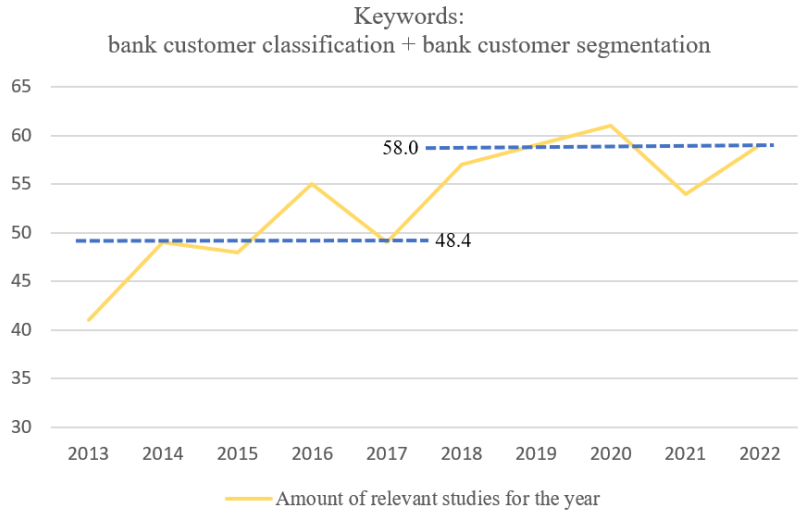

Figure 1. Number of papers with keywords "bank customer classification" or "bank customer segmentation".

The need for a separate discussion on bank customer classification is due to the special characteristics of the industry. First, banks are in the financial sector and the data they obtain is highly private. If the research must be carried out within the organization, attention must be paid to the legitimacy of the data source and the security of the data during the work process. At the same time, it can be found that the publicly available research in the thesis library is mainly based on part of the data disclosed by the bank or a small amount of simulation data, both without the need for data cleaning. This provides favourable conditions for researchers to focus on improvements in steps such as algorithms, but it also has some negative impacts due to limiting data capacity. Specifically, banking data, which is extremely complex, is also characterized by large volume and multi-dimensionality. Large volume commercial banks may have hundreds of millions of individual or group customer numbers, and the data volume will expand rapidly when customers make multiple transactions. When the feature dimension of the data increases, the performance of many machine learning algorithms will be significantly reduced, entering the "Curse of Dimensionality" state. Since the classification of bank customers is based on data mining, how to optimize the algorithm so that it can complete the accurate segmentation at a relatively low cost (including time cost and memory overhead) should become the main challenge for researchers.

DOI: 10.54254/2755-2721/49/20241167

3. Literature review on algorithms

3.1. Review on K-means

Clustering is one of the most important research contents in the field of data mining and it plays an important role in the practice of marketing [6]. Simply put, clustering can be understood as the following. Through data preparation and feature processing, the original dataset is transformed into a series of points in a multi-dimensional space. Then the distance between the points is calculated by a pre-selected distance formula. Finally, points with a small spacing are grouped into the same cluster. A cluster can be described as a dense collection of data points, distinguished from other regions by the relatively low density of the surrounding space. In some algorithms such as K-means, the number of clusters is set manually before the calculation. Obviously, entities within the same cluster are like each other, while entities between different clusters are less similar. K-means is the most classical algorithm under this classification, mainly characterized by the pre-set number of clusters, K, and the use of the average of the distances between the points to accomplish the convergence of the algorithm. It has the advantage of being able to converge to a conclusion relatively quickly, but also has several significant drawbacks [7]. First, the K-means algorithm sometimes stops immediately after obtaining a local optimal solution, which may affect the acquisition of the best conclusion. Second, it can only work for datasets where the class clusters are convex and may misclassify in other shapes. Third, K-means requires some initialization to start running, but the manual intervention on k value in this process may cause bias in the results and cannot meet the accuracy requirements. In addition, this algorithm has unique disadvantages when computed for large volumes of bank customer data. Large time overhead and memory consumption are the fatal drawbacks of K-means clustering with multiple clustering centers for big data, which also leads to high data mining costs and is difficult to be commercially used by banks. As a result, many studies on bank customer segmentation using this algorithm have tried to improve the methodology to avoid these problems as much as possible.

Although the study by Maryani et al. is based on data from credit companies rather than banks, it is cited as a case study belonging to the same financial industry for comparison purposes. In this study, 82,648 transactions from a total of 102 customers were split into two class clusters by the unmodified K-means algorithm. Obviously, since the study only classified a small number of customers and the RFM model it used to serve as a clear indicator of segmentation quality, it did not require special optimization of the algorithm [8]. Mahdiraji et al. also involves RFM model and K-means algorithm, three judgmental parameters and evaluation based on distance from average solution (EDAS) are jointly applied in this study for 20,000 bank accounts. These additional methods enhance the objectivity of K-value selection [9]. Yu et al. proposed an improved K-means algorithm with average maximum similarity (AMS), which simultaneously achieves the goal of obtaining a global optimal solution and significantly increases the convergence speed [10]. However, the above studies are all based on small sample size data or simulation data and have not been certified to be capable of meeting the objectives in real-world scenarios with higher complexity and magnitude.

In future research or practice, analysts can consider combining multiple algorithms to solve the dilemma caused by big data, such as implementing multilevel K-means clustering based on minimum spanning tree, or adopting dynamic class center adjustment and Elkan triangulation in the accelerated K-means clustering algorithm [11]. The latter has been verified to ensure fast convergence speed and low memory usage even when the data size reaches 100,000 items and the number of class clusters is larger than 20.

3.2. Review on Artificial Neural Networks (ANN)

ANN is another hot topic in today's public discourse, named after the way it processes information based on an abstraction of the human neuron network. It gathers knowledge by detecting patterns and relationships in data and learns through experience (or manual training) rather than through programming [12]. Mimicking the way the human brain processes information, neural networks (NN) are non-linear and non-limiting, which to some extent makes it possible to identify information with a

DOI: 10.54254/2755-2721/49/20241167

high degree of accuracy from big data models. However, it is prone to suffer from high memory consumption, very slow convergence or even non-convergence when facing a large amount of multi-dimensional data. This is fatal to those who attempt on its application to banking big data. As a result, most researches focus on balancing the speed and accuracy of the algorithm.

Qian et al. noticed that algorithms like NN would face the problem of decreasing accuracy as the sample size increases. So, they applied multi-class extreme learning machine (a special type of feedforward neural networks) to learning from label proportions (LLP) and then compared the quality of the algorithms under different supervision level frameworks. LLP is a learning problem that can categorize data into bags labeled with specific proportions. Experimental results show that this innovative LLP algorithm has an accuracy advantage over the other algorithms mentioned in the case of large bag sizes. As an unsupervised learning algorithm, its accuracy is slightly lower than that of supervised learning, but the cost investment is much lower as a result. Taken together, this method is a cost-effective alternative [13]. For the other two branches of NN, Tang et al. [14] combined a BP neural network with a genetic algorithm (a parallel stochastic search optimization method) to create an adaptive genetic neural network that is more accurate and faster than general ones. The problem with this study is that the simulation data contains only 2000 samples, and it is difficult to judge whether the algorithm will suffer from Curse of Dimensionality when faced with dozens of times more data.

4. Literature Review on Feature engineering

The above researches mainly focus on the algorithm itself, which only achieve the purpose of customer classification by optimizing the tools used in the data analysis phase. However, the full lifecycle of data mining also includes data acquisition and processing, data desensitization and security, and feature extraction and selection. Aiming at the feature extraction step that has yet to be optimized, an emerging concern called "feature engineering" has come into researchers' view in recent years. It is a process of transforming raw data into new features and selecting the best features from them to improve the performance of classification methods. A well-tuned feature engineering approach (including feature transformation and feature selection) can effectively compute the original data, initially filter out reliable and easy-to-use information, and thus improve the performance, accuracy and scalability of data mining models.

The figure2 shows the changes in the number of published papers searched on CNKI with the keyword "feature engineering + bank".

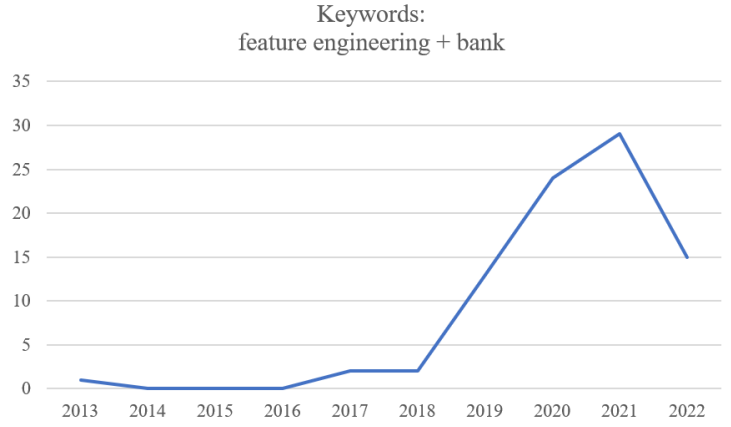

Figure 2. Number of papers with keywords " feature engineering + bank ".

Prior to the timeframe included in Figure 2, i.e. before 2013, no relevant cross-disciplinary studies were published. However, since 2019, research addressing the application of feature selection in data mining in the banking industry has shown an increase. This illustrates the growing importance of this neglected direction. Such a trend implies that the combination of this technique with specific industries

DOI: 10.54254/2755-2721/49/20241167

may become a hot topic in the future in both research and practice areas. At the same time, the fact that only a handful of studies have dealt with the field of bank customer classification suggests that there is considerable scope for its future development.

The practical advantages of using feature engineering in bank customer classification are manifold. Firstly, customer characteristics may be diverse, including demographic data, consumer behaviors and habits, occupations and salary levels, which may even include many non-numeric records. Failure to screen large amounts of data may result in negative impact on the analysis results. In addition, for common algorithms such as K-means clustering, non-numerical data is often binary-numericalized by creating a large number of dummy variables, which can also lead to biased results. With feature engineering, analysts can transform complex customer data into a numerical set of features that can help algorithms do a better job of customer segmentation. Secondly, customer data contains many features with correlations, such as the correlation between age, education level and income. Feature engineering can eliminate such correlations and reduce the information overlap between features, thus reducing the complexity of customer segmentation models. Finally, a good feature engineering can assist in data processing, thus providing a more accurate basis for model training and prediction. Specifically, it avoids outlier sensitivity when filling in missing values by using algorithms such as KNN, and it also eliminates subjective bias caused by manual removal of noise and outliers.

Duan et al. realized that there are very few systematic studies on feature selection of bank customer data, and the current studies on feature selection methods for high-dimensional data are mainly confined to a single perspective, with few comparisons of the effectiveness of multiple methods. Therefore, the team conducted an empirical study based on bank customer credit card data, and compared the data mining effects of four feature selection methods from both qualitative and quantitative perspectives, from which the optimal solution was identified. They concluded that bank customer segmentation requires the use of feature classification methods, and also proposed a feature selection method based on multimodal late fusion in their study, obtaining relatively optimal data analysis results [15]. The problem with this study is that the new feature selection method created in the paper requires the involvement of a perceived adjusted feature contribution coefficient. The model performance may fluctuate considerably due to the reasonableness of the coefficient setting. Abedin et al. mainly investigated how different feature engineering methods affect the results of different classifiers, and observed that the feature transformation methods have a significant impact on the performance of classification methods. By experimenting with five different types of feature engineering, they also found that proper feature transformations must be applied in order to achieve optimized classification performance based on machine learning algorithms, suggesting that proper data must first be obtained before actual customer classification can take place [16]. It should be noted, however, that such results may not be fully applicable to data from other banks. This is due to the fact that each bank's geographical location and main customers may be different from each other, resulting in customer data that may present different dynamics and may respond better to different feature engineering methods. In addition, due to the bank's data confidentiality policy, it is not possible for an external researcher to analyze the features in the dataset, and it is difficult to validate the results based on other bank's data.

5. Conclusion

In the field of bank customer classification, studies mainly focus on algorithmic innovation by combining existing data mining methods with industry characteristics. Published papers under the category of K-means clustering and neural networks are mainly collected in this research. And the result shows that most teams have optimized the convergence speed, accuracy and other efficiencies, but few of them have actually applied the new methods on the primary data from the banking industry. This may lead to the fact that these algorithms may still face performance degradation in real-world applications. In the future, more big data samples should be put to use. Another issue is that data mining is often done from data rather than from interpretation. This means that, unlike the research phase which focuses on one or more algorithms and works on optimizing them for the purpose, data scientists faced with the practical problem of classifying bank customers may have difficulty in discerning directly from the

DOI: 10.54254/2755-2721/49/20241167

analytical results the direction of product marketing that can be applied to reality, resulting in getting lost in the data. The need to resolve this dilemma can only rely on a deep organic integration of practice and research. This means that practitioners need to be aware of the latest research advances and be able to apply them to real-world business, while researchers need to understand the needs and limitations of real-world applications. In addition, the promising research area of feature engineering methods has only begun to emerge today. This branch may become a major research direction in the future.

Except for a very few studies, most projects are based on static customer data at a certain moment in time. In reality, however, the type of customer to which a customer belongs may change due to the implementation of different marketing approaches, changes in the industry environment, and customer initiatives. In the age of new media and social networks, for example, customer needs and expectations are changing rapidly, which leads to the possibility that classifications based on relevant data may quickly become obsolete. In recent years, little has been done to explore the dynamics of customer segmentation within specific industries. Again, this is a possible direction for future research, such as tracking the movement of specific types of customers or observing changes in customer segmentation under specific marketing activities.

References

[1] Gong X 2013 J. *South China Finance* **35** 86-8
[2] Wang G and Jiang P 2004 J. *Journal of Tongji University(Natural Science)* **49** 246-52+115
[3] Liu M and Jiang W 2020 J. *Modern Economic Science* **42** 56-68
[4] Wei Z and Ling H 2013 J. *Shanghai Finance* **34** 28-32+116
[5] Smith W R 1956 J. *Journal of marketing* 21 3-8
[6] Jain A K, Murty M N and Flynn P J 1999 J. *ACM Computing Surveys* **31** 264−323
[7] Sun J, Liu J and Zhao L 2008 J. *Journal of Software* **19** 48-61
[8] Maryani I, Riana D, Astuti R D, Ishaq A, Sutrisno and Pratama E A 2018 *Third International Conference on Informatics and Computing (ICIC)* 1-6
[9] Mahdiraji H A, Tavana M, Mahdiani P and Kamardi A A A 2021 J. *Benchmarking: An International Journal* **29** 456-95
[10] Yu H and Han X 2018 J. *Journal of Xiangtan University(Natural Science Edition)* **40** 125-28
[11] Jin X and Zhang L 2018 J. *Journal of Jilin University(Science Edition)* **56** 1187-92
[12] Zhan S, Ku T and Zhou H 2016 J. *Application Research of Computers* **33** 413-16
[13] Agatonovic-Kustrin S and Beresford R J. *Journal of Pharmaceutical and Biomedical Analysis* **22** 717-27
[14] Qian Y, Tong Q and Wang B 2019 J. *Procedia Computer Science* **162** 421-28
[15] Tang Y, Huang H and Cheng Z 2014 J. *Computer Technology and Development* **24** 192-5
[16] Duan G, Wang Y and Ma X 2022 J. *Computer Engineering and Applications* **58** 302-12
[17] Abedin M Z, Hajek P, Sharif T, Satu M S and Khan M I 2023 J. *Research in International Business and Finance* **65** 101913