DOI: 10.23977/jaip.2024.070120
ISSN 2371-8412 Vol. 7 Num. 1

Research on the application of decision tree algorithm in private universities

Qing Guan

Nanchang Normal University of Applied Technology, Nanchang, 330108, China

Keywords: Decision Tree Algorithm, Online Lending Behavior, Data Mining, Risk Prediction, Anomaly Detection

Abstract: This paper explores the application of decision tree algorithms in the analysis of private university students' online lending behavior. We utilize the decision tree classification algorithm to analyze and predict the risk levels of students' online lending behavior, and employ association rule mining techniques to identify potential risk patterns. Additionally, various data analysis methods are discussed to identify abnormal online lending behavior. The research results indicate that by comprehensively applying these methods, it is possible to effectively identify and prevent online lending risks.

1. Introduction

With the development of internet finance, the online lending behavior of college students has become an increasingly prominent social concern. Due to lack of experience and information asymmetry, private university students are prone to online lending risks. This study aims to apply data mining techniques, particularly the decision tree algorithm, to analyze and predict the online lending behavior of college students, providing a scientific basis for risk prevention.

2. Decision Tree Classification Algorithm

2.1. Algorithm Overview

The decision tree algorithm is a popular and practical machine learning technique widely used for data classification and regression analysis. It constructs a model by learning decision rules from a labeled dataset, which can be used to predict labels for new data. The core idea of a decision tree is to break down a complex decision process into a series of simple decision steps, forming a tree-like structure. Each internal node represents a test on an attribute, each branch represents the result of the test, and each leaf node represents a class label (decision outcome).

When constructing a decision tree, the algorithm selects the optimal attribute to split the dataset, based on the attribute's ability to partition the data. The quality of the split is usually evaluated using criteria such as information gain, gain ratio, or Gini impurity. Information gain measures the change in information entropy before and after splitting the dataset on an attribute, with the attribute that maximizes information gain chosen as the splitting criterion. Gini impurity assesses the disorder of the data and selects the attribute that minimizes the impurity for splitting.

The advantages of the decision tree algorithm lie in its easily understandable and interpretable model, as it can be visualized as a tree structure. Additionally, decision trees can handle both numerical and categorical data with low data preprocessing requirements. However, it also has some drawbacks, such as susceptibility to overfitting and instability for certain types of data. Strategies like pruning techniques and ensemble methods (e.g., random forests) can be employed to address these issues.

2.2. Data Collection and Preprocessing

Before applying the decision tree algorithm, data collection and preprocessing are essential steps. Data collection forms the foundation for constructing any data mining model. In this study, the primary data collected pertain to the online lending behavior of private university students. These data may include personal information (e.g., age, gender, major), academic performance (e.g., grades, attendance rate), financial status (e.g., family income, personal spending patterns), online behavior (e.g., browsing history, online shopping records), and online lending history (e.g., loan amounts, repayment status).

Collected data are often incomplete, noisy, and may exist in different formats. Therefore, data preprocessing becomes a crucial step to ensure the effectiveness of the model. Data preprocessing includes data cleaning, integration, transformation, and reduction. Data cleaning involves handling missing values and outliers, such as filling missing values with the median or mean and identifying/deleting outliers. Data integration merges data from different sources to provide a consistent data view. Data transformation converts raw data into a format suitable for algorithm processing, such as converting text data to numerical data. Finally, data reduction simplifies data through dimensionality reduction or compression techniques, reducing data volume while maintaining integrity.

2.3. Model Training and Testing

Model training is a critical step in applying the decision tree algorithm. In this study, the Classification and Regression Trees (CART) algorithm is utilized. The CART algorithm offers a mechanism to handle both numerical and categorical variables, using Gini impurity as the splitting criterion when constructing the decision tree.

Before initiating model training, the dataset is divided into training and testing sets. Commonly, cross-validation methods are used, such as using 70% of the data for training and the remaining 30% for testing. The training set is used to build the model, while the testing set is employed to evaluate the model's performance.[1]

During the training process, the CART algorithm starts from the root node, selecting the best splitting attribute, and recursively building the decision tree. Each selection is based on maximizing the reduction in Gini impurity. Once constructed, the decision tree model is used to predict the classes of the data in the testing set.

Model performance evaluation involves assessing the predicted results on the testing set. Performance metrics typically include accuracy, recall, F1 score, among others. When evaluating the model, attention should be given to the issue of overfitting. To avoid overfitting, pruning techniques such as pre-pruning and post-pruning can be applied, or ensemble learning methods like random forests can be used.

Finally, the trained model is deployed into real-world applications for predicting the classes of new data. Regular evaluations and updates to the model are conducted to ensure its predictive ability remains effective over time.

3. Association Rules

3.1. Concepts and Algorithm Selection

Association rule mining is a crucial technique in the field of data mining, used to discover interesting relationships among different items in large datasets. This technology finds wide applications in areas such as retail market analysis, network usage analysis, bioinformatics, and more. Its core objective is to identify frequent patterns, associations, correlations, or structures, particularly when these attributes are statistically associated with a specific outcome.

A typical association rule mining problem can be described as follows: "If a person buys product X, what is the probability they will also buy product Y?" The solution to such problems relies mainly on two algorithms: the Apriori algorithm and the FP-growth algorithm.[2]

The Apriori algorithm is one of the earliest and most famous association rule mining algorithms. It is based on the concept of frequent itemsets, i.e., sets of items that frequently appear in the dataset. The Apriori algorithm uses an iterative approach called a level-wise search, where k-itemsets are used to explore k+1-itemsets. The algorithm initially generates frequent itemsets by calculating the support of all individual items, then generates itemsets containing two elements, calculates support, and so on. This process continues until no higher-level frequent itemsets can be found. The key advantage of the Apriori algorithm lies in its simplicity and ease of understanding, but it may encounter efficiency issues when dealing with large datasets.

The FP-growth algorithm is another effective method for mining frequent itemsets, avoiding the candidate set generation and testing process in the Apriori algorithm. The FP-growth algorithm first constructs a data structure called an FP-tree (Frequent Pattern tree) and then mines frequent itemsets by applying recursive decomposition on this tree. Compared to Apriori, the main advantage of FP-growth is its performance, especially when dealing with datasets containing numerous frequent patterns, longer patterns, or dense databases. Additionally, due to the compressibility of the FP-tree, the FP-growth algorithm exhibits excellent space efficiency.[3]

In the data mining research on private university students' online lending behavior, association rule mining can reveal potential connections between students' online lending behavior and other behaviors (such as spending habits, social media usage, etc.). The choice of the appropriate algorithm depends on the specific features of the dataset and the research objectives. If the dataset is relatively small or researchers aim for a more intuitive understanding, the Apriori algorithm is a good choice. For larger or more complex datasets, the FP-growth algorithm may be more suitable, providing higher efficiency and scalability.

3.2. Data Mining and Analysis

Before conducting association rule mining, it is necessary to collect and prepare appropriate data. For the study of private university students' online lending behavior, this may include students' personal information (e.g., age, gender, major), financial information (e.g., family income, personal spending habits), academic information (e.g., grades, attendance), social media activities (e.g., posting frequency, content types), and online lending history (e.g., borrowing frequency, amounts, repayment status). This data can be collected from the school's database, social media platforms, financial institutions, among other sources.[4]

The first step in data mining is data preprocessing, which involves cleaning data, handling missing values, normalizing data formats, and more. Additionally, some data transformation may be required, such as discretizing continuous variables for better application of association rule mining algorithms.

Once the data is prepared, the Apriori or FP-growth algorithm can be used to mine association rules. This process typically involves two main steps: first, generating frequent itemsets, and then generating association rules from these frequent itemsets. When generating association rules, minimum support and minimum confidence need to be set. Support refers to the frequency of an itemset occurring in all transactions, while confidence is the conditional probability, i.e., the probability of the conclusion itemset occurring given the premise itemset.

In the analysis phase, researchers evaluate the generated rules and attempt to identify meaningful patterns. For example, they may discover that specific spending behaviors are closely associated with high online lending risk. These association rules can assist schools and financial institutions in better understanding students' online lending behavior and designing targeted intervention measures.

Furthermore, results obtained from association rule mining can be used to enhance risk assessment models, improving the accuracy of predicting online lending defaults. For instance, if certain spending patterns or social media behaviors are found to be highly correlated with online lending defaults, this information can be integrated into the risk assessment model to assist in making more accurate loan decisions.

In conclusion, association rule mining provides a powerful tool for understanding and predicting university students' online lending behavior. By revealing hidden relationships between students' spending habits, social behaviors, and online lending behavior, it can offer schools and financial institutions more effective risk management and prevention strategies. However, it should be noted that association rules can only reveal correlation, not causation. Therefore, when applying these rules, a comprehensive judgment should be made in conjunction with other information and real-world considerations. [5]

4. Identifying Anomalous Online Lending Behavior

4.1. Data Sources

Before delving into the discussion of identifying anomalous online lending behavior, it is crucial to have a thorough understanding of the data sources. Understanding and analyzing data from multiple channels and dimensions are essential for effectively identifying potential risk behaviors. The current digital era provides a wide range of diverse data sources, including but not limited to the following key areas:

4.1.1. Transaction Data Analysis

Transaction data is a primary basis for identifying anomalous online lending behavior. By deeply analyzing borrowers' transaction histories, frequencies, amounts, and timings, we can reveal their financial situations and spending habits. For example, frequent large transactions or irregular transaction patterns may imply financial stress or unstable income sources for borrowers.

4.1.2. Login Pattern Data Analysis

Login pattern data provides detailed information about user interactions with the online lending platform. This includes login frequency, login times, login durations, and the types of devices used. Anomalous login patterns, such as frequent logins at unconventional times or multiple logins and logouts in a short period, may indicate fraud risk or suspicion of account compromise.

4.1.3. Social Media Behavior Analysis

Social media behavior has become a crucial component of modern data analysis. By analyzing an individual's activities on social media, we can indirectly understand their lifestyle, social circle, and even psychological state. Certain patterns of behavior on social media may be correlated with financial difficulties, providing valuable supplementary information for assessing the risk of online lending behavior.

4.1.4. Other Data Sources

In addition to the aforementioned primary data sources, other sources include geographic location data, device information, credit history, user feedback, and reports. Geographic location data can reveal a borrower's residential and work environments, aiding in assessing their creditworthiness. Device information analysis (such as device type, operating system, IP address) can be used to identify unconventional device access behaviors, thereby preventing fraud risk. Credit history data provides information about a borrower's past credit behavior, helping evaluate their repayment capability and willingness. User feedback and reports serve as direct risk indicators and can be used to validate the results of other data analyses.

Through comprehensive analysis of these multi-channel and multi-dimensional data, we can construct a comprehensive borrower profile, enabling more precise identification of anomalous online lending behavior and the implementation of corresponding preventive measures.[6]

4.2. Methods and Applications

After obtaining sufficient data, various methods need to be applied to analyze this data and identify anomalous online lending behavior. This involves complex data processing techniques, including but not limited to the following aspects:

4.2.1. Behavioral Analysis

Behavioral analysis involves identifying potential risks by analyzing individual activity patterns. In the context of online lending, this includes the analysis of transaction patterns, login behavior, social media activities, and more. Establishing a baseline for a user's normal transaction patterns can help identify abnormal transaction behavior deviating from the baseline. Login behavior pattern analysis aids in identifying abnormal account access attempts, potentially indicating account compromise or fraud risk.

4.2.2. Device Information Analysis

Device information analysis focuses on the characteristics of devices used by users. By analyzing information such as device type, operating system version, IP address, etc., abnormal device access behavior can be identified. For example, logging in suddenly from an unusual location or using an uncommon device may signal a risk.

4.2.3. Machine Learning and Artificial Intelligence Techniques

Machine learning and artificial intelligence technologies play an increasingly important role in identifying anomalous online lending behavior. These technologies can handle large volumes of data and learn patterns to identify anomalous behavior. By training classification models (such as decision trees, random forests, neural networks, etc.), we can automatically identify potential risk behaviors. These models can quickly identify anomalies in large datasets, providing real-time risk monitoring.

4.2.4. Privacy and Security Considerations

When applying these methods, attention must be given to privacy and security issues. All data collection and analysis activities must comply with relevant data protection regulations to ensure that individual privacy is not violated.

By integrating these methods, we can significantly improve the accuracy and efficiency of identifying anomalous online lending behavior. This is crucial for preventing financial fraud, protecting borrowers' interests, and maintaining the stability of financial markets. The ongoing development of technology and the enhancement of data analysis capabilities will bring more innovation and progress in identifying and preventing anomalous online lending behavior in the future.

5. Empirical Analysis

5.1. Case Study Design

To delve into the application of decision tree algorithms and association rule mining techniques in analyzing the online lending behavior of private university students, we selected a private university with a diverse student population and typical behavioral characteristics as the case study. The aim is to gain comprehensive insights within this real-world context.

5.1.1. Case Selection

The chosen university not only has a representative student body but also exhibits diverse and challenging online lending behaviors. We will extensively collect data on students' personal information, academic performance, social media activities, and past online lending history to ensure the case's breadth and depth.

5.1.2. Data Collection and Preprocessing

During the data collection and preprocessing phase, our focus is on protecting student privacy while ensuring the quality and consistency of the obtained data. Through detailed data cleaning and anonymization processes, we will construct a high-quality dataset to provide a reliable foundation for subsequent analysis.

5.1.3. Application of Decision Tree Algorithm

Using the decision tree algorithm, we will build a model aimed at predicting the risk level of students' online lending behavior. Model training and testing will utilize carefully partitioned datasets, adjusting parameters and employing pruning techniques to enhance the model's performance and generalization capability.

5.1.4. Analysis Using Association Rule Mining

Simultaneously, we will apply association rule mining techniques to conduct in-depth analyses of students' consumption and online behaviors, seeking potential association patterns leading to online lending risks. Using algorithms like Apriori and FP-growth, we will mine frequent itemsets, generate a series of association rules, and reveal potential connections between specific consumption habits, social media usage patterns, and high-risk online lending behavior.

5.1.5. Objectives and Significance

This case study aims not only to identify students engaged in high-risk online lending behavior but also to understand the underlying reasons behind these behaviors. By combining decision tree algorithms and association rule mining techniques, this study provides a comprehensive approach, allowing for a more in-depth analysis and understanding of students' online lending behavior. This can assist private universities in developing more effective strategies for online lending risk management, promoting healthy financial behavior among students, and providing decision support to ensure the stability of financial markets.

5.2. Results and Discussion

Through empirical research and analysis, we obtained a series of meaningful discoveries and insights. The decision tree model successfully identified a group of students with high online lending risk, achieving high levels of accuracy and recall. This suggests that the decision tree algorithm is an effective tool for predicting and evaluating the risk of online lending behavior among university students.

Through association rule mining techniques, we discovered several patterns associated with high online lending risk behavior. For instance, specific consumption habits and social media usage patterns were closely correlated with higher online lending risk. These findings offer new perspectives for understanding student online lending behavior and may aid universities and financial institutions in implementing more effective measures for risk prevention and management.

Additionally, the study revealed some challenges in the practical application of data mining techniques, such as data quality control, model interpretability, and ensuring privacy protection throughout the data analysis process. These challenges underscore the need for a comprehensive strategy and methods to ensure the validity and security of results when applying these technologies in practical scenarios.

In summary, this case study not only demonstrates the potential application of data mining techniques in identifying and preventing online lending risks among private university students but also emphasizes the importance of integrating various technologies and methods. These findings provide valuable information and strategies for university administrators and financial institutions to more effectively manage and prevent student online lending risks. Furthermore, the research outcomes lay the groundwork and direction for future studies in this field.

6. Conclusion

By applying decision tree algorithm and association rule mining technology, this study effectively identifies and predicts the risk of online loan behavior of private college students. The research shows that the comprehensive application of various data mining technologies can provide powerful risk prevention and management tools for universities. At the same time, these methods need to be constantly adjusted and optimized in practice to adapt to the changes in student behavior and network environment.

Acknowledgement

Science and Technology Research Project of Jiangxi Provincial Department of Education, no. GJJ2203220 "Research on the Application of Data Mining Technology in the Prevention of Online Loan Behavior Risk of College Students in Private Universities"—Take Nanchang Normal University of Applied Technology as an example ".

References

[1] Shi Yansong. Analysis of blind advanced consumption behavior of college students based on "online lending platform" [J]. Chinese market. 2022(35):54-57.
[2] Yang Shan. Research on network learning behavior based on clustering algorithm and decision tree algorithm [J]. Computer knowledge and technology. 2021,17(10):213-216.
[3] Zhou Li. A CSSA model based on the decision tree algorithm [J]. computer simulation. 2021,38(05):264-268.
[4] Zhang Xiaoyun. Design and evaluation of a network intrusion detection system based on a decision tree algorithm [J]. information technology. 2023,47(02):117-122.
[5] Zhong Yunsheng. Research on the optimization method of situational awareness for network information security based on decision tree algorithm [J]. Information and computer (theoretical version). 2023,35(06):239-241.
[6] Song Lili. Intelligent evaluation algorithm for power supply reliability of real civil buildings based on decision tree [J]. total utilization of PCA. 2023,37(06):128-133.