Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/57/20241318

Data mining in economics: Unraveling merchant transactions for strategic insights

Shirui Li

University College London, London, UK

2482516799@qq.com

Abstract. This paper explores the application of data analysis and mining techniques in the domain of economics, with a specific focus on understanding merchant transaction characteristics. The study delves into fundamental theories, including classification tasks, regression missions, and relevance analysis, showcasing the versatility of these techniques in addressing economic challenges. Neural network models, such as Multilayer Perceptron and Auto Encoder, are introduced for handling complex economic data. The research emphasizes the importance of data mining in extracting valuable insights from real-world merchant transaction data, leading to the creation of the Merchant Transaction Feature Standard Database. A detailed data preprocessing method tailored to merchant transaction data is presented, addressing issues such as missing data, noise reduction, data integration, and transformation. The unique characteristics of real merchant transaction data, including sensitivity, concentration, sparsity, and the lack of label diversity, are outlined. The study concludes by highlighting the potential benefits of employing data mining techniques in optimizing marketing, merchant management, and risk management strategies. Overall, this paper contributes to advancing the understanding and practical applications of data analysis and mining techniques in economics.

Keywords: Data analysis, data mining, economics, merchant transactions, classification tasks, regression analysis.

1. Introduction

In the rapidly evolving landscape of data analysis and mining techniques, the application of these methodologies to real-world scenarios, particularly in the realm of economics, holds immense potential. This section introduced the fundamental theories underlying data analysis and mining techniques, emphasizing their role in extracting valuable insights from extensive datasets. The focus on classification tasks, regression missions, and relevance analysis highlighted the versatility of these techniques in addressing various economic challenges. Additionally, the integration of neural networks, specifically Multilayer Perceptron and Auto Encoder models, demonstrated their significance in handling complex economic data. The subsequent section delved into the standardized study of merchant transaction characteristics, emphasizing the importance of data mining in unraveling patterns and trends in merchant behavior [1]. The exploration of the data mining process, hierarchical clustering methods, and the creation of the Merchant Transaction Feature Standard Database laid the groundwork for understanding and leveraging merchant transaction data.

2. Basic Theories of Data Analysis and Mining Technique

2.1. Research Content of Data Analysis and Mining Technique

In the era of big data, the fundamental theories underlying data analysis and mining techniques play a crucial role. These theories not only aid in understanding the essence of data but also provide robust tools for addressing real-world challenges within various economic domains. Data analysis encompasses aspects such as data collection, cleaning, exploration, and visualization, while data mining primarily focuses on extracting valuable insights and patterns from extensive datasets. In the realm of economics, data analysis and mining techniques are extensively applied to areas such as market trend analysis, risk management, and customer relationship management. By delving into the foundational theories of these technologies, we can develop a deeper understanding and harness economic data effectively. This, in turn, supports informed decision-making and strategic planning, driving progress and innovation within the economic sphere.

2.1.1. Classification tasks

In the realm of data analysis and mining techniques within the field of economics, classification tasks are a common supervised learning method. Their objective is to learn a classification model from a dataset to categorize input data into predefined classes. Classification tasks usually involve training a classifier using a training dataset and applying it to a test dataset to predict the class labels of input data . The performance of classifiers is typically assessed by calculating metrics such as accuracy, recall, precision, and F1 score. Various algorithms are available for classification tasks, including decision trees, naive Bayes, support vector machines, logistic regression, and more. For instance, the Support Vector Machine algorithm's core idea is to map data points from different classes into a high-dimensional space and find an optimal hyperplane that effectively separates them. In practice, SVM maximizes the margin between data points and the hyperplane to determine the optimal classification boundary. Furthermore, feature selection and feature extraction are crucial steps in classification tasks, enhancing classifier accuracy and generalization ability [2]. Classification tasks find extensive applications in the field of economics, including medical diagnosis, financial risk assessment, customer segmentation, and image recognition, among others.

2.1.2. Return mission

In the domain of data analysis and mining techniques, regression tasks involve establishing a mathematical model based on known dataset features and their corresponding target values to predict the target values for new data. The primary objective of regression tasks is to model the relationship between independent variables and dependent variables and predict continuous numeric target variables. Regression tasks are commonly used to explore relationships between variables and predict the impact of one variable on another. The performance of regression models is typically assessed using metrics such as Mean Squared Error (MSE) and Coefficient of Determination (R^2), which measure the degree of error between predicted and actual results. Regression tasks hold extensive potential applications in economics, such as forecasting stock prices, housing prices, currency exchange rates, and more.

In Economic and Data Analysis Techniques, regression tasks refer to the use of mathematical models, based on known dataset features and corresponding target values, to predict target values for new data. The main objective of regression tasks is to model the relationship between independent and dependent variables, predicting continuous numeric target variables [3]. Regression tasks are typically used to explore variable relationships and forecast the impact of one variable on another. In regression tasks, model performance is often assessed using metrics like Mean Squared Error (MSE) and Coefficient of Determination (R^2) to measure the level of error between predicted and actual results. Regression tasks have extensive potential applications in economics, such as predicting stock prices, housing prices, currency exchange rates, and more.

2.1.3. Relevance Analysis

In the field of economics and data mining, association analysis is an important technique that aims to discover relationships between two or more attributes within a dataset. This means that when the value of one attribute changes, the values of other attributes may also change correspondingly. The core of association analysis is the mining of frequent item sets and association rules in the dataset. Frequent item sets refer to sets of items that frequently occur in the dataset, while association rules are conditional statements that describe relationships between item sets. Commonly used methods in association analysis include the Apriori algorithm and FP-Growth algorithm. These methods help individuals gain deeper insights into the relationships among items in the dataset, thereby providing more precise decision support and business recommendations in the field of economics [4].

Table 1 summarizes the key aspects of data analysis and mining techniques, particularly in the context of economics.

Table 1. Overview of Data Analysis and Mining Techniques in Economics.

Category	**Definition**	**Objective**	**Methods/Techniques**	**Applications in Economics**
Data Analysis	Use of statistical methods to study, summarize, and apply data.	To fully harness the potential of data and maximize its utility.	Data collection, cleaning, exploration, visualization.	Market trend analysis, risk management, CRM.
Data Mining	Critical step in KDD, representing an outcome of the process.	To unearth hidden information from data.	Pattern recognition, machine learning.	Unearthing economic patterns and insights.
Classification Tasks	Supervised learning method to categorize data into classes.	To learn a classification model from a dataset.	Decision trees, naive Bayes, SVM, logistic regression.	Medical diagnosis, financial risk assessment, customer segmentation.
Regression Tasks	Establishing a model to predict target values for new data.	To model the relationship between variables and predict numeric targets.	Mathematical modeling, MSE, R^2.	Forecasting stock prices, housing prices, currency exchange rates.
Association Analysis	Discovering relationships between attributes in a dataset.	To find frequent item sets and association rules in the dataset.	Apriori algorithm, FP-Growth algorithm.	Providing decision support and business recommendations.

2.2. Neural Networks in Data Mining

2.2.1. Multilayer Perceptron

The Multilayer Perceptron (MLP) is a commonly used neural network model, belonging to the category of feedforward neural networks. It is often employed for tasks such as classification and regression. The structure of an MLP typically includes multiple layers of neurons with weighted connections between them. Each neuron receives the output from the neurons in the previous layer, performs certain transformations and activations, and then passes its output to the next layer [5]. Except the input and output layers, all other layers incorporate an activation function, commonly using functions like the sigmoid or ReLU functions.

During the training of an MLP model, the backpropagation algorithm is utilized to update the weights to minimize prediction errors. The advantages of MLP models lie in their capability to handle nonlinear models and high-dimensional feature spaces, exhibiting strong learning and generalization abilities. However, drawbacks include the need for substantial amounts of training data and the potential for issues like overfitting. In the context of economics, MLP models find applications in various tasks, such as forecasting economic trends, predicting financial market behaviors, and analyzing economic data.

2.2.2. Auto Encoder

Autoencoders, which are fundamental to deep learning, are designed to extract low-dimensional features from high-dimensional input data. They primarily serve in data reconstruction and generation by learning these representations. The encoder's role is to transform input data into a compressed latent space, and the decoder then maps these latent representations back to the original data space. This mechanism enables autoencoders to abstract complex representations of the input data and perform tasks related to reconstruction and generation, making them versatile tools. The training of autoencoders often involves the backpropagation algorithm, which is crucial for refining the model parameters and reducing reconstruction errors [6]. The model's efficacy and robustness can be further enhanced by imposing additional constraints, like sparsity or denoising.

In the realm of data mining, autoencoders are employed for various purposes, including reducing data dimensionality, detecting anomalies, reconstructing data, and learning significant features. Particularly in the economic sector, they are instrumental in reducing the dimensionality of economic data, detecting irregularities in financial datasets, and extracting relevant economic features. Autoencoders' wide-ranging applications and their significant role in data mining and economic research underscore their importance in these fields.

3. A study of standardized data on merchant transaction characteristics

3.1. Neural Networks in Data Mining

Data mining in merchant transaction data offers insights into merchant behaviors and trends, essential for marketing, management, and risk management. It helps identify risks like fraud and non-compliance, improving risk management. This analysis also enhances merchant management and customer satisfaction, increasing loyalty and business. Ultimately, data mining boosts financial institutions' marketing strategies, competitiveness, and customer satisfaction.

The data mining process initiates with the collection of real merchant transaction data. This collected data undergoes a specialized preprocessing method tailored for merchant transactions, which includes steps like data integration, feature extraction, and data transformation [7]. These steps result in a refined set of merchant transaction feature data. The processed data is then subjected to hierarchical clustering using Evidence Accumulation Clustering (EAC), designed specifically for merchant transaction features. This procedure leads to the creation of the Merchant Transaction Feature Standard Database.

Delving into data collection and preprocessing, the methodology and thought process behind the creation of this database are discussed. An improved hierarchical clustering algorithm, based on EAC and tailored for merchant transaction data, is introduced. This algorithm integrates clustering concepts, using the miniBatchKmeans algorithm as the base method and enhancing it through Evidence Accumulation Clustering. Its performance and stability are validated through experiments on public datasets. Additionally, by conducting experiments on a merchant transaction dataset and comparing this algorithm with other hierarchical clustering methods, such as the BIRCH algorithm and traditional EAC algorithm, the Merchant Transaction Feature Standard Database is successfully established, and the experimental results are thoroughly analyzed.

Table 2 summarizes the key aspects of neural networks in data mining in the context of merchant transaction data analysis.

Table 2. Data Mining Process for Merchant Transaction Analysis Using Neural Networks.

Step/Method	Purpose	Outcome	Details/Techniques Used
Data Collection	Gathering real merchant transaction data.	Collection of merchant transaction data.	Real-world data gathering.
Data Preprocessing	Tailoring data for analysis: integration, feature extraction, and transformation.	Prepared merchant transaction feature data.	Data integration, feature extraction, data transformation.
Hierarchical Clustering	Clustering merchant transaction features using EAC.	Clustered data based on merchant transaction characteristics.	Evidence Accumulation Clustering (EAC).
Creation of Merchant Transaction Feature Standard Database	Establishing a standardized database for merchant transactions.	A comprehensive and standardized database for further analysis.	Use of hierarchical clustering methods.
Algorithm Validation	Testing the clustering algorithm's performance and stability.	Validated performance of the hierarchical clustering algorithm.	Experiments with public datasets, comparison with BIRCH and traditional EAC algorithms.

3.2. A data preprocessing method for merchant transaction data

This study introduces a specialized data preprocessing method for merchant transaction data, streamlining data integration, feature extraction, and transformation to improve analysis efficiency and provide deeper insights into merchant behavior for informed business decisions:

(1) Handling Missing Data: When a dataset contains missing values, data preprocessing can employ imputation methods to address these missing values, preventing them from adversely affecting subsequent economic analyses [8].

(2) Addressing Noise Issues: Data preprocessing can reduce data noise by detecting and removing outliers, duplicates, and anomalies, thereby enhancing the accuracy and robustness of economic models.

(3) Dealing with Data Integration: In practical economic research, data may originate from various sources. Data preprocessing helps integrate these diverse data sources to create a comprehensive and consistent dataset for more comprehensive economic analysis.

(4) Data Transformation: Data preprocessing involves techniques such as normalization, standardization, and discretization, which can transform economic data with different scales and ranges into a uniform format, facilitating more effective comparisons and analysis of the data.

3.3. Data Characteristics: An Overview

Due to the unique characteristics of the transaction data we obtained from real merchants, the data possesses the following features:

(1) Sensitivity and Limited Nature of Transaction Data: The transaction data contains a significant amount of merchant information, such as the merchant's location, POS machine details, etc. Additionally, it includes customer consumption information, such as credit card numbers, identity information, phone numbers, and other private data. Due to the sensitivity of this information, the data undergoes declaration and anonymization processes before being provided to us. Despite these measures, various confidentiality factors still need to be considered. Consequently, the data provided to us is limited, and our task is to analyze deeper insights from this data as features of merchant transactions.

(2) Concentration and Sparsity in Data Distribution: The sampled data from merchant transactions shows concentration in some categories and sparsity in others. For example, category C04 (department stores) has abundant transaction data due to high consumer footfall and card usage, despite fewer merchants. Conversely, category S03 (real estate) features fewer merchants with low transaction volumes and card swipes, leading to sparser data.

(3) Lack of Label Data: For privacy reasons, the data provided to us does not include real merchant data exhibiting "skimming" fraud behavior. All the data pertains to normal merchant transactions,

requiring us to simulate and create label data for "skimming" behavior. This process helps avoid the concentration of label data in real situations [9]. This lack of label data diversity can pose challenges for the training of fraud detection models.

4. Conclusion

In conclusion, this paper embarked on a comprehensive exploration of data analysis and mining techniques, focusing on their applications in economics, particularly in understanding merchant transaction characteristics. The study emphasized the crucial role of these techniques in extracting meaningful insights from vast datasets, contributing to effective decision-making in the economic domain. The incorporation of classification tasks, regression missions, and relevance analysis provided a holistic view of the diverse applications of data analysis and mining techniques. Furthermore, the introduction of neural network models, such as Multilayer Perceptron and Auto Encoder, highlighted their significance in handling complex economic data with strong learning and generalization capabilities. The examination of standardized data on merchant transaction characteristics underscored the potential benefits of employing data mining techniques in optimizing marketing, merchant management, and risk management strategies. The proposed data preprocessing method tailored to merchant transaction data demonstrated its effectiveness in handling missing data, addressing noise issues, integrating diverse data sources, and transforming data for more comprehensive economic analysis.

References

[1] Mölder, Felix, et al. "Sustainable data analysis with Snakemake." F1000Research 10 (2021).

[2] Dries, Ruben, et al. "Advances in spatial transcriptomic data analysis." Genome research 31.10 (2021): 1706-1718.

[3] Nguyen, Andy, Lesley Gardner, and Don Sheridan. "Data analytics in higher education: An integrated view." Journal of Information Systems Education 31.1 (2020): 61.

[4] Gupta, Manoj Kumar, and Pravin Chandra. "A comprehensive survey of data mining." International Journal of Information Technology 12.4 (2020): 1243-1257.

[5] Dogan, Alican, and Derya Birant. "Machine learning and data mining in manufacturing." Expert Systems with Applications 166 (2021): 114060.

[6] Romero, Cristobal, and Sebastian Ventura. "Educational data mining and learning analytics: An updated survey." Wiley interdisciplinary reviews: Data mining and knowledge discovery 10.3 (2020): e1355.

[7] Gupta, Shashi Kant, et al. "Data Mining Processes and Decision-Making Models in the Personnel Management System." Designing Workforce Management Systems for Industry 4.0. CRC Press, 2023. 85-104.

[8] Jones, Charles I., and Christopher Tonetti. "Nonrivalry and the Economics of Data." American Economic Review 110.9 (2020): 2819-2858.

[9] McLaren, John. "Racial disparity in COVID-19 deaths: Seeking economic roots with census data." The BE Journal of Economic Analysis & Policy 21.3 (2021): 897-919.