Analysis of Daily Household Transactions using the KDD Process

Generated by ChatGPT

September 26, 2023

Abstract

This research paper employs the Knowledge Discovery in Databases (KDD) process to analyze a dataset of daily household transactions. By systematically progressing through data selection, preprocessing, transformation, mining, and evaluation, the analysis uncovers insights into transaction behaviors, modes, and financial dynamics. The findings serve as a foundation for understanding and optimizing household financial management.

1 Introduction

This research paper provides an in-depth analysis of a dataset containing daily household transactions. By leveraging the Knowledge Discovery in Databases (KDD) process, the objective is to uncover meaningful patterns and insights from the dataset, which consists of various transactional behaviors.

2 Knowledge Discovery in Databases (KDD) Process

The KDD process, often synonymous with data mining, is a systematic approach to discovering valuable knowledge from vast amounts of data. The process comprises several stages:

1. **Data Selection:** This stage involves choosing the relevant dataset or subset of data from a larger pool.
2. **Data Preprocessing:** Here, the data is cleaned and transformed to address issues like missing values, noise, or inconsistencies.
3. **Data Transformation:** The data is then converted or condensed into forms suitable for mining.
4. **Data Mining:** Algorithms are applied to extract patterns, relationships, or knowledge from the processed data.
5. **Evaluation:** The final step involves interpreting and validating the discovered patterns to ensure they provide meaningful insights.

3 Methodology and Application of KDD

3.1 Data Selection

For our analysis, we chose a dataset titled "daily_household_transactions.csv". This dataset comprises various attributes related to daily household transactions, such as the mode of transaction, category, amount, and whether it was an income or expense.

3.2 Data Preprocessing

3.2.1 Handling Missing Values

```
data['Subcategory'].fillna('Not Specified', inplace=True)
data['Note'].fillna('No Note', inplace=True)
```

Explanation: Missing values can distort the results of data analysis. In our dataset, the "Subcategory" and "Note" columns had missing entries. To address this, we replaced missing values in the "Subcategory" column with "Not Specified" and those in the "Note" column with "No Note".

3.2.2 Encoding Categorical Variables

```
for column in ['Mode', 'Category', 'Subcategory', 'Income/Expense', 'Currer
    data[column] = LabelEncoder().fit_transform(data[column])
```

Explanation: Machine learning algorithms require numerical input data. Therefore, we transformed categorical columns into numerical values using label encoding. This process assigns a unique integer to each category.

3.2.3 Scaling Numerical Variables

```
data['Amount'] = StandardScaler().fit_transform(data[['Amount']])
```

Explanation: The magnitude of transaction amounts can vary widely. Scaling ensures that the range of this feature doesn't unduly influence the algorithms. We used the StandardScaler to standardize the "Amount" column.

3.3 Data Transformation

The preprocessing steps, including encoding and scaling, served as our data transformation phase, making the data ready for mining.

3.4 Data Mining and Evaluation

3.4.1 Visual Analysis

```
sns.histplot(data['Amount'], bins=50, kde=True)
sns.countplot(data=data, x='Mode')
sns.countplot(data=data, x='Income/Expense')
```

Explanation: Visual analysis aids in understanding the underlying patterns in the data. We plotted histograms and bar charts to visually inspect the distribution of transaction amounts, the frequency of different transaction modes, and the balance between income and expense entries.

4 Results

The visual analysis revealed:

- A concentration of transactions around smaller magnitudes.
- Variability in the frequency of different transaction modes.
- A higher proportion of expenses compared to incomes in the dataset.

5 Conclusion

Through the KDD process, we successfully transformed and analyzed the daily household transactions dataset. The insights gained provide a foundational understanding of transactional behaviors, which can be crucial for financial planning and management.