User Behavior Recognition

Đinh Quang Minh

08 - 2024

Introduction

- Problem Statement
- Motivation
- Main Contributions
- Key Technique
- Dataset
- Comparison Experiment
- Key challenges

Problem Statement

- **User Character Analysis:** Traditional psychological models based on question-and-answer tests may not be suitable in the context of fraud detection, as user characteristics can change with the environment and may not be accurately represented through these tests.
- **Recognition Models Based on Transaction Behavior and Interaction Behavior:**
- Current models focus only on external behavior without considering the internal psychological factors of users, which can lead to missing important aspects that influence user behavior in new situations.
- When faced with entirely new situations, models based only on known behaviors may not be accurate, leading to errors in recognition.

Motivation

- The rise of digital finance and e-commerce has significantly increased fraud risks, causing substantial financial losses.
- Traditional rule-based systems and existing data mining techniques are inadequate due to their reliance on external behaviors and the need for manual intervention.
- There is a need to incorporate psychological factors into fraud detection to improve accuracy and efficiency.

Main Contributions

- The introduction of the transaction character model that incorporates user behavior psychology into fraud detection..
- The creation of a user behavior benchmark and a mapping relationship model that links user interaction behavior with transaction behavior.

Key Technique

- **Cautiousness Character Model**
- **User Behavior Benchmark**
- **Relationship Model Between Cautiousness and Transaction Behavior**

Key Technique

I. Cautiousness Character Model: The cautiousness character model aims to describe the user's cautiousness when interacting with products before making a transaction. Three main features define this model:

- Browsing Cautiousness in Browsing History (BCH)
- Contrast of Similar Products at the Time of Purchase (CPP)
- Comparison of Similar Products in History (CPH)
- Cautiousness Score (cau)

1.BCH (Browsing Cautiousness in Browsing History):
The ratio of browsing events to the total events (browsing + purchase) for a given product category C_i

$$\text{BCH}_i = \frac{\text{visit\_type\_count}_i}{\text{purchase\_type\_count}_i + \text{visit\_type\_count}_i}$$

- $\text{visit\_type\_count}_i$ = Number of browsing events for category C_i.
- $\text{purchase\_type\_count}_i$ = Number of purchase events for category C_i.

2.Contrast of Similar Products at the Time of Purchase (CPP):

A normalized score representing the user's cautiousness, computed using a sigmoid function.

$$\mathrm{CPP}_i = \frac{\mathrm{type\_visit\_time}_i}{\mathrm{session\_time}}$$

- $\mathrm{type\_visit\_time}_i$ = Total time spent browsing similar products in category C_i during the session.
- $\mathrm{session\_time}$ = Total time of the session.

3.Comparison of Similar Products in History (CPH):

The ratio of the number of viewed products to the total number of products in a category C_i

$$\text{CPH}_i = \frac{\text{type\_good\_count}_i}{\text{type\_visit\_count}_i}$$

- $\text{type\_visit\_count}_i$ = Number of products viewed in category C_i.
- type_good_count_i = Total number of products in category C_i.

4.Cautiousness Score (cau):
A normalized score representing the user's cautiousness, computed using a sigmoid function.

$$\mathrm{cau}_i = \frac{1}{1 + e^{-3\times(\mathrm{BCH}_i+\mathrm{CPP}_i+\mathrm{CPH}_i-1.5)}}$$

1.Transaction Amount Attribute (TPR):
Measures the likelihood of spending within specified price ranges by dividing transaction amounts into quintiles and calculating probabilities.

$$\mathrm{TPR}_q = \frac{N_q}{N}$$

- N_q = Number of transactions within the q-th quintile.
- N = Total number of transactions.

2.Product Category Attribute (CGP):
Measures the probability distribution of transactions across different product categories.

$$\mathrm{CGP}_j = \frac{T_j}{T}$$

- T_j = Number of transactions in category C_j.
- T = Total number of transactions.

3.Product Price Level Attribute (GPL):

Captures user preference for different price levels within product categories.

$$\text{GPL}_{k,j} = \frac{P_{k,j}}{P_j}$$

- $P_{k,j}$ = Number of transactions in price level k for category C_j.
- P_j = Total number of transactions in category C_j.

4.Transaction Interval Attribute (TDD):
Analyzes the time intervals between consecutive transactions.

$$\mathrm{TDD} = \frac{1}{N-1}\sum_{i=2}^{N}(t_i - t_{i-1})$$

- t_i = Timestamp of the i-th transaction.
- N = Total number of transactions.

5.Interval Attributes of Same Category Transaction (CTDD): Measures the time intervals between purchases within the same product category.

$$\mathrm{CTDD}_j = \frac{1}{M_j - 1} \sum_{i=2}^{M_j} (t_{i,j} - t_{i-1,j})$$

- $t_{i,j}$ = Timestamp of the *i*-th transaction in category C_j.
- M_j = Number of transactions in category C_j.

6.Transaction Date Attribute (TIW):
Reflects whether transactions are more likely to occur on workdays or holidays.

$$\mathrm{TIW}_d = \frac{D_d}{D}$$

- D_d = Number of transactions on day type d (workday or holiday).
- D = Total number of transactions.

7.Transaction Count Attribute (STC):
Captures the number of transactions per session.

$$\mathrm{STC} = \frac{N_s}{S}$$

- N_s = Number of transactions in session s.
- S = Total number of sessions.

III. Relationship Model Between Cautiousness and Transaction Behavior

1.Distance Between Current Behavior and Behavior Benchmark (dbhv):
Measures the deviation between current transaction attributes and user's behavior benchmark.

$$d_{\mathrm{bhv}}(\mathbf{X}, \mathbf{BBV}_u) = \sqrt{\sum_{i=1}^{57} (x_i - \mathrm{bbv}_i)^2}$$

- $\mathbf{X}$ = 57-dimensional vector of current transaction attributes.
- $\mathbf{BBV}_u$ = 57-dimensional vector of user's behavior benchmark.

III. Relationship Model Between Cautiousness and Transaction Behavior

2.Distance Between Current Cautiousness and Cautiousness Benchmark (dcau):

Measures the difference between current interaction behavior and cautiousness benchmark.

$$d_{\text{cau}}(\mathbf{Y}, \mathbf{TCV}_u) = \sqrt{\sum_{i=1}^{30} (y_i - \text{tcv}_i)^2}$$

- $\mathbf{Y}$ = 30-dimensional vector of current interaction behavior's cautiousness.
- $\mathbf{TCV}_u$ = 30-dimensional vector of user's cautiousness benchmark.

3.Relationship Model Function (f):
Combines the distances with weights to determine overall fraud likelihood.

$$f(\mathbf{X}, \mathbf{Y}) = \omega_{\text{bhv}} \cdot d_{\text{bhv}} + \omega_{\text{cau}} \cdot d_{\text{cau}} + \omega_0$$

- ω_{bhv}, ω_{cau}, ω_0 = Weights determined using a least-squares generalized inverse method.

III. Relationship Model Between Cautiousness and Transaction Behavior

4.Decision Rule (g):
Classifies the transaction as fraudulent or normal based on the relationship model function and a threshold.

$$g(\mathbf{X}, \mathbf{Y}) = f(\mathbf{X}, \mathbf{Y}) - \text{threshold}$$

DATASET

DATASET Overview

1. **Original Source:** The dataset is derived from a public dataset of a domestic e-commerce platform, covering transactions over one year (from May 1, 2016, to April 30, 2017).
2. **Key Components:**
   - **User Information Dataset:** Contains information on 98,924 users.
   - **Order Dataset:** Includes 792,723 transaction records for 98,924 customers.
   - **Behavior Dataset:** Contains 6,941,441 interaction records for 93,453 users.
   - **Commodity Information Dataset:** Contains details on 99,412 products across six categories.

Dataset Derivation Process

- **User Sorting:** Users are sorted from high to low based on the number of transaction records and grouped into various categories with transaction limits of 300, 160, 100, 30, 20, 10, and 5.
- **Fraud Simulation:** Fraudsters are simulated by randomly selecting users from the same group as the target user, extracting transaction logs, and corresponding interaction logs.

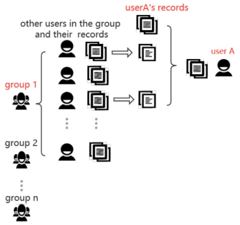

Figure 1: Experimental dataset derivation process.

Comparison Experiment

- The performance is measured using four indicators: recall rate, precision rate, accuracy rate, and F1 value.
- A baseline model (UBC) is used for comparison with the proposed UBRMTC model.
  - UBRMTC shows higher performance than UBC across most datasets, particularly in recall, precision, accuracy, and F1 value.
  - UBRMTC demonstrates stability even with fewer user transactions, while UBC shows more fluctuation.
  - The UBRMTC model's time complexity is comparable to UBC, with both having a transaction detection complexity of $O(1)$.

Comparison Experiment

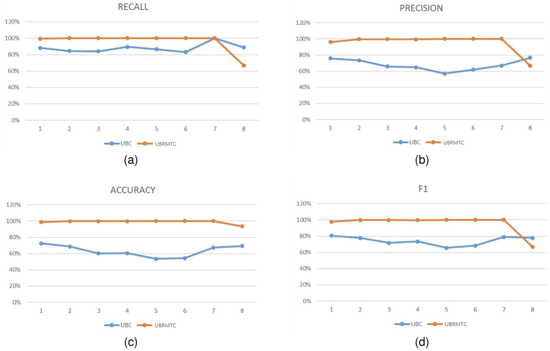

Figure 2: Results of evaluation indicators for each set: (top left) recall rate of UBC and UBRMTC for Groups 1–8, (top right) precision of UBC and UBRMTC for Groups 1–8, (bottom left) accuracy of UBC and UBRMTC for Groups 1–8, and (bottom right) F1 value of UBC and UBRMTC for Groups 1–8. (a) RECALL. (b) PRECISION. (c) ACCURACY. (d) F1.

Comparison Experiment

- UBRMTC excels by integrating interaction behavior and cautiousness, enhancing fraud detection accuracy.
- UBRMTC accurately distinguishes between normal and fraudulent transactions through cautiousness (dcau) and transaction behavior (dbhv) analysis.

			Number of algorithm execution	Time complexity
Modeling	UBRMTC	Cautiousness	3m+30	O(n)
		Transaction behaviour	7t+7	O(n)
		Threshold selection	Ct	O(n)
	UBC	Transaction behaviour	7t+7	O(n)
		Threshold selection	Ct	O(n)
Classification	UBRMTC	Transaction detection	10	O(1)
	UBC	Transaction detection	7	O(1)

Key challenges

- Integrating psychological traits into fraud detection models, which traditionally only analyze external behaviors.
- Balancing the influence of cautiousness with actual transaction data to accurately detect fraud without increasing false positives.
- Handling the diversity and variability in user behavior across different transaction categories and platforms.

Thanks