Попереднє опрацювання для нейромережних методів інтелектуального аналізу даних у задачах електронної комерції

Ю. Цимбал, к.т.н., доц.; Р. Ткаченко д.т.н., проф.; У. Поліщук, к.т.н., асист.; П. Вітинський, студ.

Національний університет «Львівська політехніка»
E-mail: yurij.tsymbal@gmail.com

Abstract. The problem of the development of methods for data preprocessing for neural network on the geometrical transformation model has been considered. The importance of using of exploratory data analysis has been shown. The results of the proposed method for solving the problem of prediction of returns in online trading have been presented.

Keywords: neural network, geometrical transformation model, Data Mining Cup.

Вступ

Одним із найпоширеніших різновидів задач інтелектуального аналізу даних є дослідження економічних процесів, зокрема, у електронній комерції. Основною характерною рисою таких задач є велика кількість накопичених даних про клієнтів, зроблені транзакції та необхідність видобування прихованих в них знань, що надасть можливість підвищувати ефективність торгівлі та збільшувати прибутки.

Збір та підготовка даних до опрацювання є, як правило, найтривалішим етапом інтелектуального аналізу даних і таким, що найгірше піддається узагальненню, оскільки найбільше залежить від конкретної задачі. При побудові моделі в першу чергу слід зважати на прояви некоректності вхідних даних, зокрема, на пропущені значення, неможливі події або нереальні величини, так звані "викиди" (outliers). Крім того, важливим є врахування різних форматів представлення та кодування вхідних даних, застосування різних одиниць виміру, а також проблем, пов'язаних із частотою збору даних і датою останнього оновлення.

Постановка задачі

Розглянемо методику попередньої обробки даних на конкретному прикладі. Постановка задачі і дані для нього наведено на інтернет-сайті всесвітнього студентського конкурсу з інтелектуального аналізу даних Data Mining Cup (DMC) 2014 [1]. Автори статті були тренерами та учасниками команд Національного університету "Львівська політехніка", що змагались у конкурсі – Uni_Lviv_Polytechnic_1 та Uni_Lviv_Polytechnic_2.

Інтернет-торгівля останнім часом набуває дедалі більшої популярності. Водночас, дотримуючись принципу "клієнт завжди правий", більшість інтернет-магазинів розвинених країн, зокрема Німеччини, надають можливість покупцям протягом певного терміну повернути придбаний товар до магазину за повну вартість. У торгівлі одягом та взуттям частота таких повернень може досягати 50 %.

Для зменшення можливих витрат при цьому, інтернет-магазин зацікавлений у побудові як найточнішого прогнозу факту "повернення/неповернення" після оформлення замовлення певного товару певним покупцем (тобто віднесення цього замовлення до одного з двох класів), або визначення ймовірності такого повернення.

Інтернет-магазин надав учасникам конкурсу DMC 2014 історичні дані про замовлення товарів упродовж квітня 2012 р. – березня 2013 р., загалом 481092 записи, що складаються з 14 атрибутів (рис. 1).

#	Name	Datatype	Comment
1	orderItemID	INT	порядковий номер замовлення товару
2	orderDate	DATE	дата замовлення
3	deliveryDate	VARCHAR	дата доставки
4	itemID	INT	ідентифікатор товару
5	size	VARCHAR	розмір
6	color	VARCHAR	колір
7	manufacturerID	INT	ідентифікатор виробника
8	price	DECIMAL	ціна
9	customerID	INT	ідентифікатор покупця
10	salutation	VARCHAR	звертання до покупця, напр. Mr. або Ms.
11	dateOfBirth	VARCHAR	дата народження
12	state	VARCHAR	місце проживання (федеральна земля Німеччини)
13	creationDate	DATE	дата створення облікового рахунку покупця в магазині
14	returnShipment	TINYINT	ознака "повернення/неповернення"

Рис. 1. Структура таблиці з навчальними даними конкурсу DMC 2014 (до попереднього опрацювання).

На основі цих даних потрібно розробити певну модель, яка за відомими значеннями перших 13 параметрів (вхідних) обчислює значення останнього (вихідного) параметра.

Зауважимо, що у таблиці з навчальними даними наявні численні пропуски у значеннях параметрів *deliveryDate* (39419 записів), *color* (143 записи) і *dateOfBirth* (48889 записів).

Для проведення тестів учасникам конкурсу DMC 2014 надано історичні дані про замовлення товарів упродовж квітня 2013 р., загалом 50078 записів, з наперед невідомими значеннями атрибуту *returnShipment*, які і потрібно передбачити.

Опис проведених досліджень

Для розв'язання поставленої задачі інтелектуальний аналіз даних проведено у декілька етапів за схемою [2]:

1. Аналіз умови задачі, яку розв'язують, формулювання цілей, які має бути досягнуто внаслідок застосування методів видобування даних, а також вибір оптимальних методів інтелектуального аналізу даних.

В якості інструменту для розв'язання задачі вирішено обрати нейромережні засоби на основі парадигми "модель геометричних перетворень" (МГП) [3], яка реалізує принципово інші концепції у по–

рівнянні із класичними нейромережами прямого поширення, такими як “back propagation”. Процес синтезу структури та навчання нейромережі є неітеративним, внаслідок чого стає можливим отримання швидкого розв’язку для задач інтелектуального аналізу даних надвеликих розмірностей. Також мережі МГП мають досить високу точність і здатність до узагальнення, є простими у використанні та забезпечують повторюваність результатів навчання завдяки відсутності початкової ініціалізації вагових коефіцієнтів випадковими значеннями.

2. Підготовка даних для автоматизованого аналізу.

Оскільки вхідні дані задачі є досить різнотипними: числовими, текстовими, неперервними та дискретними, для коректної роботи їх попередньо перекодовано у числовий формат, як єдино можливий для представлення вхідних даних нейромережі МГП.

Оскільки природа зв’язків між вхідними змінними (атрибутами) є загалом невідомою, на цьому етапі використано методи розвідувального аналізу [4]. Результатом є, зокрема, вибір найбільш важливих змінних, виявлення відхилень та аномалій та розробка початкових моделей.

У результаті аналізу виявлено досить цікаві закономірності та тенденції. Зокрема, зауважено суттєві відмінності у відсотку повернень серед покупців із західної Німеччини (від 44,4 % до 49,0 %) і східної (від 50,2 % до 50,8 %). Тому було вирішено застосувати для кодування атрибуту *state* шкалу, де значення від 1 до 16 відповідають землям, впорядкованим за зростанням цього відсотку. Відсутність значення атрибуту *deliveryDate* свідчить про неповернення товару, оскільки він і не надійшов покупцю з тих чи інших причин, що було враховано при прийнятті рішення.

Також, замість безпосереднього використання атрибутів типу “дата” було вирішено ввести нові атрибути *oc* (“*orderDate-creationDate*”), який відповідає тривалості користування покупцем цим інтернет-магазином та *do* (“*deliveryDate-orderDate*”), який відповідає проміжку часу між замовленням та доставкою товару покупцеві (рис. 2).

#	Name	Datatype	Comment
1	price	DECIMAL	ціна
2	state	VARCHAR	місце проживання (федеральна земля Німеччини)
3	salutation	VARCHAR	звертання до покупця, напр. Mr. або Ms.
4	oc	INT	різниця між датами замовлення та створення рахунку
5	do	INT	різниця між датами доставки та замовлення
6	returnShipment	TINYINT	ознака "повернення/неповернення"

Рис. 2. Структура таблиці з навчальними даними конкурсу DMC 2014 (після попереднього опрацювання)

3. Застосування методів видобування даних та побудова моделі.

4. Перевірка побудованої моделі.

Для перевірки адекватності побудованої моделі використано один з найпростіших та поширених способів перевірки, який полягає в тому, щоб із вхідних навчальних даних сформувати дві вибірки: першу застосувати для побудови моделі (тренувальна вибірка); другу – використати для перевірки побудованих моделей (тестова вибірка). За різницею точності класифікації на тренувальній та тестовій вибірках оцінюють адекватність побудованої моделі.

5. Інтерпретація моделі людиною з метою використання для прийняття рішення.

Внаслідок експериментів, проведених за методикою з п. 4, було прийнято рішення застосувати порогову функцію на виході нейромережі для зведення отриманого дійсного значення до одного з двох: 0 – “товар не повернено”, 1 – “товар повернено”.

Результати досліджень

Отримані результати класифікації замовлень товарів інтернет-магазину за ознакою “повернення/неповернення” на основі допомогою нейромережі моделі геометричних перетворень подано у табл. 1.

Таблиця 1

Результати класифікації замовлень товарів інтернет-магазину

	Тестові дані		Сума штрафних балів
	Повернення	Неповернення	
Реальні значення	25029	25049	
Правильно класифіковані значення	21013 (84,0 %)	10281 (41,0%)	
Неправильно класифіковані значення	4016 (16,0%)	14768 (59,0%)	**18784 (37,5%)**

З табл. 1 видно, що модель досить точно спрогнозувала випадки реального повернення товару, проте досить часто неправильно класифікувала замовлення з неповерненим товаром, як з поверненим.

Отриманий результат дав змогу команді Uni_Lviv_Polytechnic_1 посісти 25 місце у підсумковій таблиці з 50 учасників.

Використання нейромережі узагальненої регресії (GRNN) на попередньо опрацьованих вхідних даних за структурою рис. 2, без застосування порогової функції на виході дало команді Uni_Lviv_Polytechnic_2 результат у 21813,04125 балів (39 підсумкове місце).

Для порівняння, команда Uni_Iowa_State_1, що здобула 1 місце, отримала 14165 балів, що відповідає 28,0% неправильно класифікованих значень.

Висновки

Розроблена методика попередньої обробки даних для задач електронної комерції, зокрема для класифікації замовлень, надає можливість підвищити ефективність розв’язання завдань видобування даних великих розмірностей, представлених навчальними вибірками значних обсягів (сотні тисяч векторів).

[1]. Data Mining Cup 2014 http://www.data-mining-cup.de/en/review/goto/article/dmc-2014.html.

[2]. Барсегян А.А. и др. Анализ данных и процессов (3-е изд.). – СПб.: БХВ-Петербург, 2009. – 512 с.

[3]. Грицик В.В., Ткаченко Р.О. Нові підходи до навчання штучних нейромереж // Доповіді Національної Академії Наук України, 2002, № 11, С. 59-64.

[4]. John W. Tukey. Exploratory Data Analysis. – Addison Wesley, 1977. – 711 p.