УДК 004.9 : 004.89

А. А. Яровий
А. А. Папа
О. П. Прозор

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ВІДТОКУ КЛІЄНТІВ ТЕЛЕКОМ-КОМПАНІЇ

Вінницький національний технічний університет

Анотація

Дослідження присвячено задачі підвищення точності прогнозування відтоку клієнтів компації за допомогою використання методів машинного навчання. Запропоновано модель прогнозування відтоку клієнтів телеком-компанії, що відрізняється від відомих комбінованим застосуванням методів дерев рішень та найближчих сусідів. Реалізовано інформаційну технологію аналізу відтоку клієнтів телеком-компанії.

Ключові слова: машинне навчання, дерево рішень, метод найближчих сусідів.

Abstract

The study focuses on improving the accuracy of forecasting the outflow of customers through the use of machine learning methods. A model for forecasting the outflow of clients of a telecom company is proposed, which differs from the others by the combined use of decision tree methods and the nearest neighbors. Information technology analysis of the outflow of customers of the telecom company is implemented.

Keywords: machine learning, decision tree, method of the nearest neighbors.

Вступ

На теперішній час велика кількість компаній орієнтуються на отриманні нових та збереженні старих клієнтів, яким надаються ті чи інші послуги компанії. Однією із задач маркетингових відділів є прогнозування відтоку клієнтів.

Актуальність ідеї створення програмного продукту для прогнозування відтоку клієнтів є досить високо-пріоритетною на даний час, адже майже кожен власник бізнесу, який побудований на роботі з клієнтами та наданні тих чи інших послуг, хоче бути застрахований або як мінімум попереджений про можливий відтік клієнтів. Досить часто доводиться зустрічатися з проблемою відсутності програмних засобів для прогнозування відтоку клієнтів. Тому, доцільним є розробка інформаційної технології для аналізу відтоку клієнтів телеком-компанії.

Метою роботи є створення web-серверу та android-додатку, які будуть аналізувати та видавати інформацію у вигляді прогнозу щодо можливого відтоку клієнтів заданої компанії на основі вхідних даних.

Результати дослідження

Задача прогнозування відтоку клієнтів є задачею класифікації. Тобто, на основі відомих характеристик користувача необхідно передбачити належність його до групи тих користувачів, які підуть або залишаться. Задача класифікації є задачею навчання з учителем, тобто необхідні набори даних: навчальна та тестова вибірки.

До найпопулярніших методів машинного навчання для вирішення задачі класифікації можна віднести дерева рішень та метод найближчих сусідів [1].

Дерево рішень — елементарний класифікатор, використовується в галузі статистики та аналізу даних для прогнозних моделей. Дерево рішень містить такі елементи як «листя» і «гілки». На ребрах («гілках») дерева ухвалення рішення записані атрибути, від яких залежить цільова функція, в «листі»

записані значення цільової функції, а в інших вузлах — атрибути, за якими розрізняються випадки. Щоб класифікувати новий випадок, треба спуститися по дереву до листа і видати відповідне значення. Подібні дерева рішень широко використовуються в інтелектуальному аналізі даних. Мета полягає в тому, щоб створити модель, яка прогнозує значення цільової змінної на основі декількох змінних на вході.

Найчастіше дерево рішень є узагальненням досвіду експертів, засобом передачі знань майбутнім співробітникам або моделлю бізнес-процесу компанії. Наприклад, до впровадження масштабованих алгоритмів машинного навчання в банківській сфері завдання кредитного скорингу вирішувалася експертами. Рішення про видачу кредиту позичальникові приймалося на основі деяких інтуїтивно (або з досвіду) виведених правил, які можна представити у вигляді дерева рішень [2].

Серед переваг дерев рішень можна виділити:

- породження чітких правил класифікації, зрозумілих людині, наприклад, "якщо вік <25 та інтерес до мотоциклів, то відмовити в кредиті" (інтерпретованість моделі);
- дерево рішень можна легко візуалізувати, тобто його можна "інтерпретувати", як саму модель (дерево), так і прогноз для окремого взятого тестового об'єкта (шлях в дереві);
- швидкі процеси навчання і прогнозування;
- невелика кількість параметрів моделі;
- підтримка і числових, і категоріальних ознак.

Недоліками дерев рішень є:

- у породженні чітких правил класифікації є й інша сторона: дерева дуже чутливі до шумів у вхідних даних, вся модель може кардинально змінитися, якщо трохи зміниться навчальна вибірка (наприклад, якщо прибрати одну з ознак або додати кілька об'єктів), тому і правила класифікації можуть сильно змінюватися, що погіршує інтерпретованість моделі;
- необхідність відсікати гілки дерева (pruning) або встановлювати мінімальну кількість елементів в листі дерева або максимальну глибину дерева для боротьби з перенавчанням. Втім, перенавчання – проблема всіх методів машинного навчання;
- проблема пошуку оптимального дерева рішень (мінімального за розміром і здатного без помилок класифікувати вибірку) NP-повна, тому на практиці використовуються евристики типу жадібного пошуку ознаки з максимальним приростом інформації, які не гарантують знаходження глобально оптимального дерева;
- модель може тільки інтерполювати, але не екстраполювати. Тобто, дерево рішень здійснює константний прогноз для об'єктів, що перебувають у просторі ознак поза паралелепіпеда, що охоплює всі об'єкти навчальної вибірки.

Метод найближчих сусідів (k Nearest Neighbors, або kNN) – теж поширений та "популярний" метод класифікації, також іноді використовується в задачах регресії. На рівні з деревом рішень, він один з найбільш зрозумілих підходів до класифікації. Формально основою методу є гіпотеза компактності: якщо метрика відстані між прикладами введена досить вдало, то схожі приклади набагато частіше лежать в одному класі, ніж в різних [3]. Наприклад, якщо не знаєш, який тип товару вказати в оголошенні для Bluetooth-гарнітури, можеш знайти 5 схожих гарнітур, і якщо 4 з них віднесені до категорії "Аксесуари", і тільки один - до категорії "Техніка", то здоровий глузд підкаже для свого оголошення теж вказати категорію "Аксесуари".

Для класифікації кожного з об'єктів тестової вибірки необхідно послідовно виконати наступні операції:

- обчислити відстань до кожного з об'єктів навчальної вибірки;
- відібрати об'єкти навчальної вибірки, відстань до яких мінімальна;
- клас об'єкта, який класифікують - це клас, який найчастіше трапляється серед найближчих сусідів.

Під задачу регресії метод адаптується досить легко - на 3 кроці повертається не мітка, а число - середнє (або медіанне) значення цільового показника серед сусідів.

Примітна властивість такого підходу - його оптимальність. Це означає, що обчислення починаються тільки в момент класифікації тестового прикладу, а не заздалегідь, тільки при наявності навчальних прикладів, ніяка модель не будується. У цьому відмінність, наприклад, від раніше розглянутого дерева рішень, де спочатку на основі навчальної вибірки будується дерево, а потім відносно швидко відбувається класифікація тестових прикладів [4].

Варто відзначити, що метод найближчих сусідів - добре вивчений підхід. Для методу найближчих сусідів існує чимало важливих теорем, які стверджують, що на "нескінченних" вибірках це оптимальний метод класифікації. Автори класичної книги "The Elements of Statistical Learning" вважають kNN теоретично ідеальним алгоритмом, застосовність якого просто обмежена обчислювальними можливостями і прокляттям розмірності.

Якість класифікації / регресії методом найближчих сусідів залежить від декількох параметрів:

- кількість сусідів;
- метрика відстані між об'єктами (часто використовуються метрика Хеммінга, евклідова відстань, косинусова відстань і відстань Маньківського). Відзначимо, що при використанні більшості метрик значення ознак треба масштабувати. Умовно кажучи, щоб ознака "Зарплата" з діапазоном значень до 100 тисяч не вносила більший вклад в відстань, ніж "Вік" зі значеннями до 100;
- ваги сусідів (сусіди тестового прикладу можуть входити з різними вагами, наприклад, чим далі приклад, тим з меншим коефіцієнтом враховується його "голос").

Перевагами методу найближчих сусідів є:

- метод є швидким на не дуже великих вибірках;
- проста реалізація;
- добре вивчений теоретично;
- хороший метод для першого рішення задачі, причому не тільки класифікації або регресії, а й, наприклад, рекомендації;
- можна адаптувати під потрібне завдання вибором метрики або ядра (наприклад, ядро може задавати операцію подібності для складних об'єктів типу графів, а сам підхід kNN залишається тим же);
- достатня інтерпретація, можна пояснити, чому тестовий приклад був класифікований саме так.

Серед недоліків методу найближчих сусідів можна виділити:

- якщо в наборі даних багато ознак, то важко підібрати відповідні ваги і визначити, які ознаки не важливі для класифікації / регресії;
- залежність від обраної метрики відстані між прикладами. Вибір за замовчуванням евклідової відстані найчастіше нічим не обґрунтований. Можна відшукати хороше рішення перебором параметрів, але для великого набору даних це забирає багато часу;
- метод схильний перенавчатися;
- повільно працює на дуже великих вибірках.

Отримані результати аналізу методів машинного навчання для прогнозування відтоку клієнтів показують доцільність і високу перспективність застосування обраних. Тому, запропоновано інформаційну технологію аналізу відтоку клієнтів, що відрізняється комбінованим застосуванням методів дерев рішень та найближчих сусідів (kNN) для аналізу відтоку клієнтів, що забезпечило підвищення точності прогнозу відтоку клієнтів.

Проведено тестування системи та здійснено аналіз результатів тестування інформаційної технології аналізу відтоку клієнтів. Тестування показало повну відповідність системи поставленим задачам, а саме виконання аналізу менше ніж за 2 секунди, точність формування прогнозу відтоку клієнта більше 85%. При порівнянні розробленого продукту з аналогами, отримано кращі результати за рахунок використання комбінації методів машинного навчання.

Висновки

Здійснено аналіз методів машинного навчання для прогнозування відтоку клієнтів, за результатами якого була підтверджена доцільність та перспективність застосування комбінацій декількох методів для вирішення вказаної задачі. Розроблено модель прогнозування відтоку клієнтів телеком-компанії, що відрізняється від відомих комбінованим застосуванням методів дерев рішень та найближчих сусідів (kNN), що забезпечило підвищення точності прогнозу відтоку конкретного клієнта. Реалізовано клієнтську та серверну частини інформаційної технології аналізу відтоку клієнтів телеком-компанії.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Папа А. А. Аналіз методів машинного навчання для прогнозування відтоку клієнтів [Електронний ресурс] / А. А. Папа, В. В. Уштаніт, А. А. Яровий, О. П. Прозор // ВНТУ. – 2018. – Режим доступу до ресурсу: https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2018/paper/view/5317.
2. Analysis of churn prediction: A case study on telecommunication services in Macedonia [Електронний ресурс]. – Режим доступу: https://www.researchgate.net/publication/312573014_Analysis_of_churn_prediction_A_case_study_on_telecommunication_services_in_Macedonia
3. Машинне навчання. Типи навчання. [Електронний ресурс]. – Режим доступу: https://courses.prometheus.org.ua/courses/IRF/ML101/2016_T3/about
4. MachineLearning.ru [Електронний ресурс]. – Режим доступу: http:// www.machinelearning.ru

Яровий Андрій Анатолійович — д.т.н., професор, завідувач кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, Хмельницьке шосе, 95, e-mail: a.yarovyy@vntu.edu.ua.

Папа Андрій Андрійович — аспірант кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, Хмельницьке шосе, 95, e-mail: papa.andriy@gmail.com.

Прозор Олена Петрівна — к.пед.н., доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, Хмельницьке шосе, 95, e-mail: prozor@vntu.edu.ua

Andrii A. Yarovyi — Doctor of Science (Eng.), Professor, Head of the Computer Science Department, Vinnytsia National Technical University, Vinnytsia, Khmelnytske shose, 95, e-mail: a.yarovyy@vntu.edu.ua.

Andrii A. Papa — Postgraduate Student of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, Khmelnytske Shose, 95, e-mail: papa.andriy@gmail.com.

Olena P. Prozor — Candidate of Pedagogic Sciences, Prof. Assistant of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, Khmelnytske Shose, 95, e-mail: prozor@vntu.edu.ua