УДК 004.89

А. А. Папа
Є. О. Шемет
А. А. Яровий
Л. М. Ваховська

ПРОГНОЗУВАННЯ ВІДТОКУ КЛІЄНТІВ ЗА ДОПОМОГОЮ КОМБІНУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Розглянуто актуальність проблеми прогнозування відтоку клієнтів. Здійснено аналіз моделі прогнозування відтоку клієнтів на основі комбінування дерев рішень та методу найближчих сусідів, що використовується в основі методу беггінгу. За результатами дослідження була підтверджена доцільність та перспективність застосування моделі на основі комбінування дерев рішень та методу найближчих сусідів у реальному програмному продукті.

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

Abstract

The urgency of the problem of forecasting the outflow of customers is considered. An analysis of the customer outflow forecasting model based on a combination of decision trees and the nearest neighbors method used in the basis of the beggling method is performed. The results of the study confirmed the feasibility and prospects of applying the model based on a combination of decision trees and the method of the nearest neighbors in a real software product.

Keywords: machine learning, decision tree, nearest neighbor method, bagging, data analysis.

Вступ

У сучасному світі в більшості компаній, де збирається велика кількість даних, придатних для аналізу, використовуються методи машинного навчання та інтелектуального аналізу даних. Одним із поширених прикладів використання методів машинного навчання у реальному секторі бізнесу є задача прогнозування відтоку клієнтів. Прогнозуванням відтоку клієнтів в основному займаються телекомунікаційні компанії, банки, страхові компанії та інші. В умовах постійної конкуренції прогнозування відтоку клієнтів з метою утримання стає одним з найактуальніших напрямків у сучасному бізнесі. Як правило, існуючі дані є великими масивами зі структурованою та неструктурованою інформацією, в яких для аналізу та виявлення прихованих закономірностей широко використовується інтелектуальний аналіз даних та основані на ньому методи машинного навчання [1]. Таким чином для вирішення задачі збереження існуючих користувачів доцільно проаналізувати моделі прогнозування відтоку клієнтів на основі комбінування дерев рішень та методу найближчих сусідів, що використовується в основі методу беггінгу, які надають можливість прогнозування відтоку клієнтів.

Метою роботи є аналіз моделі прогнозування відтоку клієнтів на основі комбінування дерев рішень та методу найближчих сусідів, що використовується в основі методу беггінгу, що в подальших дослідженнях дозволить підвищити точності прогнозу відтоку клієнтів.

Результати дослідження

Першим етапом роботи з відтоком клієнтів є передчасне виявлення групи осіб, схильних до припинення користування послугами. Знаючи заздалегідь про можливість відтоку клієнта можна застосувати стратегічні рішення. Основна мета аналізу відтоку клієнтів полягає у створенні списку клієнтів, які з великою ймовірністю у найближчому майбутньому будуть перервані. Існують різні підходи до аналізу відтоку клієнтів. Більшість із них заснована на методах машинного навчання, що показують у сучасних умовах досить високу ефективність [2].

В даному випадку задачу прогнозування відтоку клієнтів доцільно розглядати спільно із задачею класифікації. Тобто, на основі відомих характеристик користувача необхідно спрогнозувати 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. Дослідники, що займаються статистикою вже давно використовують метод, який має назву «bootstrap sampling», що умовно може бути перекладено як «варіація завантажувальної вибірки». Одне із втілень такої ідеї в машинному навчанні – «bootstrap aggregating», або скорочено «bagging», тобто «об'єднання результатів при різних навантаженнях» [3]. Ідея беггінгу полягає в тому, що при відсутності великої навчальної вибірки можна створювати багато випадкових вибірок з вихідної простим вибором із заміщенням. Хоча елементи в вибірках можуть перетинатися або дублюватися на практиці, все ж результати об'єднання з багатьох вибірок виявляються точніші, ніж тільки по одній початковій. Метод так називається, оскільки він об'єднує результати прогнозування різних класифікаторів, навчених на випадкових підмножинах. Беггінг виявляється корисний тільки у випадку різних нестабільностей класифікаторів, коли малі зміни в початковій вибірці призводять до існуючих змін класифікації [4].

Розглянемо детальніше сутність самого методу беггінгу в контексті поставленої задачі. Нехай є навчальна вибірка X. За допомогою bootstrap згенеруємо з неї вибірки X_1,…,X_M. Тепер на кожній вибірці навчимо свій класифікатор a_i (x). Підсумковий класифікатор буде усереднювати відповіді всіх a_i (x) (в разі класифікації це відповідає голосуванню). Візуалізацію цієї схеми представлено на рисунку 1.

Рисунок 1 – Схема роботи композиції методів

Беггінг дозволяє знизити дисперсію (variance) при навчанні класифікатора, зменшуючи величину, яка показує ступінь відмінності помилки, при умові навчання моделі за допомогою різних даних або, інакше кажучи, являє собою запобіжну ланку для перенавчання (рис. 2). Ефективність методу досягається за рахунок того, що базові алгоритми, які пройшли навчання на різних підвибірках, виходять досить різними, і їхні помилки взаємно компенсуються при голосуванні, а також за рахунок того, що об'єкти-виключення можуть не потрапляти до деяких навчальних підвибірок [4].

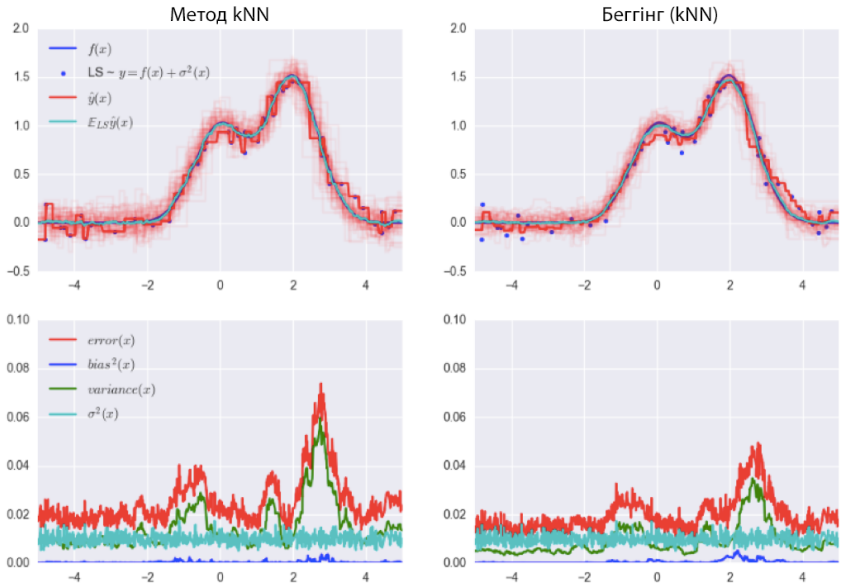

Рисунок 2 – Графіки результату роботи композиції методів: беггінгу із методом найближчих сусідів

Із графіків помітно, що помилка дисперсії набагато менше при беггінгу, як і було вказано вище. Беггінг ефективний на малих вибірках, коли виключення навіть малої частини навчальних об'єктів призводить до побудови істотно різних базових класифікаторів. У разі великих вибірок зазвичай генерують підвибірки істотно меншої довжини.

Висновки

Здійснено аналіз моделі прогнозування відтоку клієнтів на основі комбінування дерев рішень та методу найближчих сусідів, що використовується в основі методу беггінгу, за результатами якого була підтверджена доцільність та перспективність застосування підходу для вирішення вказаної задачі. Досліджено методи прогнозування відтоку клієнтів телекомунікаційної компанії, що відрізняється від відомих застосуванням моделі прогнозування відтоку клієнтів на основі комбінування дерев рішень та методу найближчих сусідів, що використовується в основі методу беггінгу, що в подальших дослідженнях дозволить підвищити точності прогнозу відтоку клієнтів.

Отримані результати дослідження показують доцільність і перспективність застосування обраного підходу для створення реального програмного продукту [5]. Отримані результати планується використати в подальшій роботі з метою підвищення якості прогнозування відтоку клієнтів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425.
2. Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. Expert Systems with Applications, 36(10), 12547-12553.
3. Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.
4. Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. Statistical science, 22(4), 477-505.
5. Andrii Papa, Yevhen Shemet, Andrii Yarovyi, Lyubov Vahovska “Development of information technology for analyzing the customer churn of a telecommunication company”. –

Information and control systems. – Vol. 2, No. 2(64), 2022. – p. 11-15. – DOI: https://doi.org/10.15587/2706-5448.2022.255861

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

Шемет Євген Олександрович — аспірант кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, Хмельницьке шосе, 95, e-mail: yevhene@gmail.com.

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.

Любов Михайлівна Ваховська – асистент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, Хмельницьке шосе, 95, e-mail: lmnechipor@gmail.com.

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

Yevhen A. Shemet — Postgraduate Student of Computer Science Department, Vinnytsia National Technical University, Vinnytsia, Khmelnytske Shose, 95, e-mail: yevhene@gmail.com.

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.

Liubov M. Vakhovska – Assistant of the Computer Science Department, Vinnytsia National Technical University, Vinnytsia, Khmelnytske shose, 95, e-mail: lmnechipor@gmail.com.