

АНСАМБЛЕВІ МЕТОДИ МАШИННОГО НАВЧАННЯ

ENSEMBLE METHODS OF MACHINE LEARNING

Ансамблеві методи – це галузь машинного навчання, яка на вирішення завдання навчає множини моделей про «слабких учнів» і поєднує їх задля отримання кращого результату.

Можна виділити три категорії ансамблевих алгоритмів [1].

1. Bagging або Bootstrap Aggregation. У цій категорії паралельно та незалежно навчаються кілька «слабких учнів», результат класифікації визначається голосуванням. Для організації майже повної незалежності дані для навчання слабких учнів вибираються з навчальної вибірки за допомогою статистичного методу bootstrap. Суть методу можна описати так: виробляється генерація вибірок розміру B з вихідного набору даних розміру N здійснюється шляхом випадкового вибору з повтореннями елементів у кожному з B разів.

2. Boosting. У цій категорії алгоритмів група «слабких учнів» навчається послідовно, кожен новий «слабкий учень» навчається на даних, які були погано класифіковані попереднім учнем. Результат передбачення визначається кожним слабким учнем, ними накладається вага моделі, та, скомбінувавши результати передбачень, визначається результат. Даний вид навчання дозволяє отримувати точні передбачення на всіх типах даних.

3. Stacking. У цій категорії алгоритмів «слабкі учні» навчені різними методами, а результати їх прогнозів навчають мета-модель, яка визначає кінцевий результат прогнозу. Властиво «слабким учнем» може бути і сама мета-модель.

Для досягнення поставленої мети було обрано два ансамблеві алгоритми машинного навчання - Random Forest та AdaBoost.

Random Forest – представник ансамблевих алгоритмів машинного навчання категорії bagging. Модель, навчена за допомогою алгоритму, є колекцією дерев вибору, кожне з яких навчається незалежно і паралельно, а результат класифікації визначається загальним голосуванням. Серед переваг цього алгоритму варто відзначити його простоту, швидкість порівняно з іншими алгоритмами бегінга та бустингу та високу точність.

AdaBoost – ансамблевий алгоритм машинного навчання категорії boosting. Схема його є ітераційним алгоритмом, у якому після кожної епохи навчання ваги «слабких учнів» індивідуально перераховуються. Ваги збільшуються для некоректно передбачених екземплярів та зменшуються для коректно передбачених. На першій ітерації слабкі учні навчаються на оригінальному наборі даних, із кожною наступною ітерацією набір даних змінюється. Результат класифікації визначається загальним голосуванням навчених «слабких учнів».

Література

1. Ensemble methods: bagging, boosting and stacking. URL: <https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205> (дата звертання 11.11.2024).