

ВИБІР АЛГОРИТМУ МАШИННОГО НАВЧАННЯ ДЛЯ ЗАСТОСУВАННЯ ПРИ ОБРОБЦІ ДАНИХ В ПІДМЕРЕЖІ ІОТ

Бугаєнко Ю.М., Курдеча В.В.

Інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, Україна

E-mail: yura.buhaenko@gmail.com

CHOICE OF MACHINE EDUCATION ALGORITHM FOR APPLICATION FOR DATA PROCESSING IN THE IOT NETWORK

The article describes the choice of machine learning algorithm for application in the proposed data processing method in the IoT subnet to further reduce the load on the information processing server.

У статті наводиться опис вибору алгоритму машинного навчання для застосування у запропонованому методі обробки даних в підмережі ІоТ для подальшого зменшення навантаження на сервер обробки інформації.

Сенсорні мережі і вузли ІоТ (Internet of Things) продукують величезні обсяги інформації, що генеруються з вбудованих або носимих розумних пристроїв і корпоративних банків даних. Отримати корисну інформацію з таких обсягів даних стає все важче, а проблема аналізу даних є однією із ключових на даному етапі розвитку Інтернету речей. У деяких випадках, інтерактивне управління та обробка даних може знадобитися в реальному часі, наприклад в областях, що мають особливе значення, наприклад системи спостереження за здоров'ям чи в системи безпеки. Але в інших випадках це не є визначаючим фактором, тому використовуються різні методи для зменшення кількості оброблюваних даних задля зменшення навантаження на обчислювальну мережу.

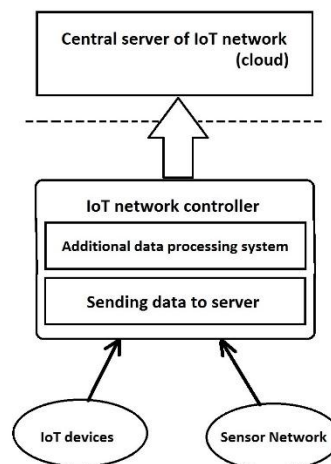


Рис. 1. Модифікована схема передачі і обробки даних в ІоТ.

У попередній роботі [1] було запропоновано вирішувати дану проблему на стороні підмережі IoT і з застосуванням алгоритмів злиття даних. В новій роботі, для досягнення універсальності системи прийняття рішень а також покращення її якісних показників, було вирішено перенести її на алгоритми машинного навчання (рис. 1).

При виборі алгоритму який слугуватиме базою для даного способу обробки даних потрібно враховувати складність алгоритму. Оскільки в таких мережах, центральний вузол має не надто великі обчислювальні можливості, то вибраний алгоритм має бути максимально простим, але при цьому достатньо точно аналізувати різні параметри в даних що надходять [2]. Насправді, будь-які підходи до машинного навчання роблять одне і теж. У нас завжди є дві основні речі: По-перше, це набір гіпотез. Це може бути набір функцій, або набір ваг для нейронів нейронної мережі, або набір всіляких вирішальних дерев. Залежить від конкретного підходу. По-друге, у нас є набір навчальних даних. Наше завдання завжди зводиться до одного: визначити, яка ж з наших моделей найбільш адекватна нашим навчальним даним.

Після огляду і порівнянні деяких алгоритмів, для застосування його в даному способі обробки даних, був обраний Наївний Байєсівський класифікатор. Байєсівський підхід дає точний чисельний критерій для вирішення цього завдання.

Наївний Байєсівський алгоритм (НБА) - це алгоритм класифікації, заснований на теоремі Байєса з припущенням про незалежність ознак. Моделі на основі НБА досить прості і вкрай корисні при роботі з дуже великими наборами даних. При своїй простоті НБА здатний перевершити навіть деякі складні алгоритми класифікації. Це і стало вирішальним фактором для вибору саме цього алгоритму.

З точки зору машинного навчання теорему Байєса можна застосувати до машинного навчання таким чином: позначимо наші дані - D , а наші гіпотези - h . Тоді нам потрібно знайти вірогідність гіпотез наших даних $p(h | D)$, яка за теорією Байєса рівна:

$$p(h | D) = \frac{p(D|h) p(h)}{p(D)}$$

Нас цікавить тільки співвідношення ймовірностей, тому ми можемо викинути з цього виразу $p(D)$ (D не залежить від h) і $p(h)$ (будемо вважати, що всі гіпотези різновірогідні; строго кажучи, це не завжди так, ми ще повернемося до цього питання, але в багатьох випадках це справедливо) [3]. Виходить, що нам треба знайти таку гіпотезу h , для якої максимально $p(D | h)$. Говорячи математичною мовою, ми повинні для кожної гіпотези обчислити її апостеріорну ймовірність і вибрати гіпотезу, для якої ця ймовірність максимальна.

Розглянемо переваги і недоліки Байєсівського класифікатора більш детально:

Позитивні сторони:

- Класифікація, в тому числі багатокласова, виконується легко і швидко. Це дозволяє прогнозувати ймовірність для безлічі значень цільової змінної.

- Коли допущення про незалежність виконується, НБА перевершує інші алгоритми, такі як логістична регресія (logistic regression), і при цьому вимагає менший обсяг навчальних даних.

- НБА краще працює з категорійними ознаками, ніж з безперервними. Для безперервних ознак передбачається нормальний розподіл, що є досить сильним допущенням.

- Класифікація в режимі реального часу. НБА дуже швидко навчається, тому його можна використовувати для обробки даних в режимі реального часу.

Негативні сторони:

- Якщо в тестовому наборі даних є певне значення категорійної ознаки, яке не зустрічалося в навчальному наборі даних, тоді модель присвоїть нульову ймовірність цього значення і не зможе зробити прогноз. Це явище відоме під назвою «нульова частота». Дану проблему можна вирішити за допомогою згладжування. Ще одним обмеженням НБА є припущення про незалежність ознак. В реальності набори повністю незалежних ознак зустрічаються вкрай рідко.

Висновки. В даній статті був запропонований спосіб обробки даних на стороні підмережі IoT який базується на Наївному Байєсівському критерії. Даний спосіб має декілька вагомих переваг, головними з яких є зменшення, за рахунок його використання обсягів даних, що в свою чергу зменшуватиме навантаження на сервер обробки інформації, та безпроводну мережу. Також суттєвою перевагою даного способу є те, що його імплементація може проходити без додавання додаткових фізичних пристроїв до системи Internet of Things.

Література

1. Бугаєнко Ю.М., «Модифікована структура мережі IoT для зменшення навантаження на мережу», ст 32-38, Київ 2016
2. Fan Tongke, “Smart Agriculture Based on Cloud Computing and IOT,” Journal of Convergence Information Technology (JCIT), vol8., issue2, no 2, pp. 1-5, Jan 2013
3. Metechko I. Наивный байесовский классификатор [Електронний ресурс] / Igor Metechko – Режим доступу до ресурсу:
<https://sohabr.net/habr/post/262155/?version=109183>.