Кокозей А. Ю.,
Криворізький Національний університет
Шаповалова Н. Н.,
ст. викладач, Криворізький Національний університет

ДОСЛІДЖЕННЯ АСПЕКТІВ СТРАТЕГІЙ СЕМПЛІНГУ

Розроблено модуль для генерації вибірок даних з використанням різних стратегій семплінгу.

В задачах машинного навчання якість моделей сильно залежить від вихідних даних, але дані дуже рідко бувають ідеальними. Отже, перед початком побудови моделей, необхідно провести певні міри, щодо перетворення даних у більш інформативні та зручні для роботи. Такі міри називають попередньою обробкою даних [1]. Найважливіші етапи попередньої обробки даних включають: агрегацію (об'єднання двох або більше ознак в одну), зменшення розмірності, нормалізацію ознак (перетворення значень ознак у абсолютно інший набір значень за допомогою простих функцій), семплінг даних та інші.

Семплінг даних – це метод дослідження множини даних, що використовується для вибору, маніпулювання та аналізу репрезентативної підмножини для виявлення закономірностей та тенденцій. Репрезентативність семплу (вибірки) – це здатність вибіркової сукупності відтворювати основні характеристики генеральної сукупності. Репрезентативність досягається за рахунок правильного формування семплу, який, за принциповими для дослідження параметрами, має відтворювати генеральну сукупність об'єктів. Таким чином, семплінг дає можливість працювати з невеликою, керованою кількістю даних, щоб швидше будувати та запускати аналітичні моделі, при цьому отримуючи точні результати.

Семплінг може бути особливо корисним для наборів даних, які занадто великі для ефективного аналізу. Ідентифікація та аналіз репрезентативного зразка є більш ефективними, зокрема економічно, ніж аналіз усієї сукупності даних. Важливим фактором є розмір семплу даних та можливість введення помилки вибірки.

Помилка вибірки – це відхилення результатів, отриманих за допомогою вибіркового спостереження від справжніх даних генеральної сукупності. Помилки вибірки можна розділити на два класи: це помилки реєстрації та помилки репрезентативності.

До помилок реєстрації відносять випадкові та статистичні помилки. Випадкові помилки – це статистичні похибки, властиві вибірковому методу. Такі помилки зменшуються при зростанні обсягу семплу. Систематична помилка залежить від різних факторів, що впливають на дослідження, зміщують результати дослідження в певну сторону та залежать від якості вихідної генеральної сукупності.

Важливу роль у семплінгу грає розмір семплу. У деяких випадках невелика вибірка зможе бути достатньо репрезентативною та відобразити найважливішу інформацію про набір даних. В інших, використання більшої вибірки може збільшити ймовірність точного представлення даних у цілому, навіть якщо збільшений розмір вибірки може перешкоджати простоті маніпулювання та інтерпретації.

Існує багато різних методів для отримання зразків з даних, відповідний обирається залежно від набору даних та ситуації. Методи семплінгу можна поділити за двома типами. Ймовірнісний семплінг – це такий вид пошуку підмножини, при якому кожний елемент генеральної сукупності має ненульову ймовірність потрапити у вибірку. При неімовірнісний семплінгу у вихідній множині існує хоча б один елемент який може не потрапити у вибірку взагалі, що зробить підмножину не репрезентативною, тому об’єкти обираються не випадково [2].

Розглянемо види ймовірнісного семплінгу. Простий ймовірнісний семплінг – це випадковий вибір предметів з усієї сукупності. Систематичний семплінг є різновидом випадкової вибірки, впорядкованої за будь-якою ознакою: перший елемент відбирається випадково, потім, з кроком n відбирається кожен k-ий елемент семплу. Пропорційний семплінг – такий, що у вибірці представлені різні групи об’єктів за певним параметром в тих самих пропорціях, що і серед генеральної сукупності. У кластерному семплінгу одиницями семплу виступають не об'єкти, а кластери, сформовані за певним параметром. Стратифікований семплінг дозволяє на основі загального фактору створювати підмножини набору даних, а семпли збираються випадковим чином з кожної підгрупи.

При семплінгу, що базується на неімовірнісному підході, вибірку даних визначають на основі аналітичних суджень. Такий вид відокремлення вибірки використовується при зборі інформації шляхом соціологічних опитувань і не може бути застосованим до існуючого набору даних.

ВИСНОВКИ

Семплінг – метод дослідження множини шляхом аналізу її підмножин. Ця методика допомагає зменшити дані під час початкового та кінцевого аналізу даних. Це може бути надзвичайно корисно, якщо обробка всієї сукупності даних є дорогою або забирає багато часу. Семплінг працює, коли семпл є репрезентативним для всього набору даних. Отже, якщо вибірка буде обрана невдало, це буде безпосередньо відображено у кінцевому результаті. Тому існує безліч стратегій, які допомагають виконувати семплінг якісно, залежно від потреби та ситуації. Таким чином, розробка програмного забезпечення, яке автоматизує різні стратегії семплінгу, є актуальною задачею.

ЛІТЕРАТУРА

1. Айвазян С. А. Прикладна статистика: основи моделювання та попередня обробка даних / С. А. Айвазян, Л. Д. Мешалкин, И. П. Єнюков. – М: Фінанси і статистика, 1987.
2. Deming W. Some Theory of Sampling / W. Deming., 2010. – 640 с. – (Dover Publications).

Юшкевич.С.В ,
Державний університет «Житомирська Політехніка»
к.к.п, доцент Вакалюк.Т.І,
Державний університет «Житомирська політехніка»

ЗАГРОЗИ ЯКІ ВИНИКАЮТЬ ПРИ ВПРОВАДЖЕННІ ШТУЧНОГО ІНТЕЛЕКТУ

Комп'ютерні технології з'явились відносно недавно і здавалось би за короткий час їх існування та розвитку ,вони не повинні були б сильно змінитись. Але зрозуміло ,що це не так – XX і XXI століття стали надзвичайно продуктивними саме в розвитку науки і не останнє місце в цій системі займають комп'ютерні технології.