УДК 519.7

РІЗНОМАНІТНІ СТРАТЕГІЇ СЕМПЛІНГУ В УМОВАХ НЕЗБАЛАНСОВАНОСТІ КЛАСІВ

Студ. С.С. Хижняк , гр. БМО1-17
Науковий керівник доц. О.Л. Блохін
Київський національний університет технологій та дизайну

Мета і завдання. Дослідити процес аналізу вибірки, що містить великі масиви інформації. Охарактеризувати основні стратегії семплінгу в умовах незбалансованості класів.

Об'єкт та предмет дослідження. Процес та методи семплінгу, зокрема коли вибірка містить занадто малу або занадто велику частку прикладів деякого класу.

Методи та засоби дослідження. Використовуються такі загальнонаукові методи дослідження, як аналіз і синтез та індукція і дедукція, також теоретичний метод опису.

Наукова новизна та практичне значення отриманих результатів. На сьогоднішній день дуже мало вітчизняних науковців приділяють увагу цій темі, тому важливо дослідити та охарактеризувати процес семплінгу.

Результати дослідження. При зборі і консолідації даних в сховищах даних можуть накопичуватися величезні масиви інформації. Коли виникає необхідність виконати аналіз аналітик стикається з проблемою вибору: використовувати для аналізу всі наявні дані чи вибрати зі всієї безлічі наявних даних деяку підмножину і на основі аналізу цієї підмножини спробувати зробити висновки про властивості досліджуваного процесу або об'єкта.

Обидва підходи мають свої переваги і недоліки. Використання всієї безлічі наявних даних дозволить забезпечити максимальну достовірність результатів аналізу, оскільки при цьому буде врахована вся доступна інформація. Але на практиці цей підхід реалізувати проблематично, оскільки обсяг аналізованих даних може виявитися настільки великим, що час і обчислювальні ресурси, необхідні для аналізу, вийдуть за всі мислимі рамки. В результаті переваги, отримані в результаті аналізу даних, будуть знецінені.

Головна перевага другого підходу полягає в тому, що аналізується відносно невелика підмножина даних. При цьому часові і обчислювальні витрати невисокі, що особливо важливо для оперативного управління, коли управлінські рішення, які генеруються на основі результатів аналізу, потрібно отримувати практично в режимі реального часу. Недолік підходу в тому, що, оскільки при аналізі використовується тільки частина доступної інформації, вельми вірогідне зниження достовірності результатів аналізу.

Існують різні стратегії вилучення підмножин спостережень з вихідних множин даних. Комплекс таких методів отримав назву «семплінг» (sampling). Основний недолік семплінгу полягає у ймовірності отримання некоректного результату аналізу якщо вибірка не відображатиме в достатній мірі початкову сукупність.

Підмножина даних, що використовується для аналізу, має бути репрезентативною : вибрана таким чином, щоб результати її аналізу не сильно відрізнялися від результатів, які були б отримані при аналізі всієї множини доступних даних. Для забезпечення репрезентативності вибірки семплінг використовує спеціальні методи і алгоритми відбору, що сприяють забезпеченню її інформаційної насиченості.Найпростіший метод семплінгу - рівномірний випадковий семплінг

(вихідна сукупність розбивається на підгрупи, а вибірка проводиться випадковим чином з цих підгруп). Головний його недолік полягає в тому, що, якщо вихідна сукупність не є гомогенною, тобто містить різного розміру групи, в кожній з яких дані мають різний розподіл, це може привести до отримання зміщеною вибірки. Кращих результатів вдається домогтися, якщо робити вибірку кожної групи, незалежно від інших груп. Для цього використовуються інші методи семплінгу, такі як стратифікаційний і кластерний.

Стратифікаційний семплінг виконується в два етапи: спочатку здійснюється стратифікація - угруповання елементів вихідної сукупності у відносно однорідні підгрупи, які називаються стратами або шарами; потім в кожному з шарів виділяється вибіркова частка, що є пропорційною складу всієї початкової сукупності. Таким чином, при використанні пропорційного розподілу вибірка, отримана в результаті стратифікаційного семплінгу, виявляється зменшеною копією вихідної сукупності.

Нерідко виникають ситуації, коли в наборі даних частка прикладів деякого класу занадто мала (цей клас будемо називати міноритарним, а інший, сильно представлений, - мажоритарних). Коли важливі з точки зору розв'язуваної задачі об'єкти або події представлені не дуже великим числом спостережень, це не дозволяє виконати їх достовірний аналіз. Відновлення балансу класів може проходити двома шляхами. У першому випадку видаляють деяку кількість прикладів мажоритарного класу (undersampling), у другому - збільшують кількість прикладів міноритарного (oversampling).

Найпростіша стратегія видалення прикладів мажоритарного класу з набору даних -випадкове видалення прикладів мажоритарного класу (Random Undersampling), коли розрахована кількість мажоритарних прикладів випадково вибирається і видаляється. Приклади з мажоритарного класу можуть віддалятися не тільки випадковим чином, але і за певними правилами. Серед таких стратегій:

1. Пошук зв'язків Томека (Tomek Links)
2. Правило зосередженого найближчого сусіда (Condensed Nearest Neighbor Rule)
3. Односторонній семплінг (One-side sampling, one-sided selection)
4. Правило «очищення» сусіда (neighborhood cleaning rule)

Інший підхід - збільшення числа прикладів міноритарного класу. Найпростіший метод - це дублювання прикладів міноритарного класу. Залежно від того, яке співвідношення класів необхідно, вибирається кількість випадкових записів для дублювання. Такий підхід до відновлення балансу не завжди може виявитися найефективнішим, тому був запропонований спеціальний метод збільшення числа прикладів міноритарного класу - алгоритм SMOTE (Synthetic Minority Oversampling Technique).

Висновки. Семплінг дозволяє пришвидшити час аналізу даних, проте щоб результати аналізу залишались такими, як при аналізі всієї множини, потрібно застосувати той метод семплінгу, від якого дані вибірки постраждають найменше.

Ключові слова. Аналіз, семплінг, вибірка, алгоритм, клас.

ЛІТЕРАТУРА

1. Паклин Н. Б., Орешков В. И. Биэнес аналитика: от данных к энаниям -СПб.: Питер, 2013. -704 с.

2. Различные стратегии сэмплинга в условиях несбалансированности классов. // [Електронний ресурс]. – Режим доступу: https://basegroup.ru/ community/articles/ imbalance-datasets

3. Сэмплинг. // [Електронний ресурс]. – Режим доступу: https://basegroup.ru/deductor/ function/algorithm/sample