DOI: 10.34185/1991-7848.itmm.2023.01.077

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ З ВИКОРИСТАННЯМ WEKA EXPLORER

Гнатушенко Вік.В., Дейнека Б.М.

Український державний університет науки і технології. Україна

Технології інтелектуального аналізу даних потрібні в першу чергу спеціалістам, що ухвалюють важливі рішення, - керівникам, аналітикам, експертам, консультантам. Прибуток компанії більшою мірою визначається якістю цих рішень - точністю прогнозів, оптимальністю вибраних стратегій, і від якості цих рішень залежить розвиток компанії. Класичні методики виявляються малоефективними для багатьох практичних завдань, оскільки неможливо точно описати реальність за допомогою невеликого числа параметрів моделі, або розрахунок моделі займає дуже багато часу і обчислювальних ресурсів. Тому використання технології, яка автоматично видобувала б із даних нові нетривіальні знання у формі моделей, залежностей, законів тощо, гарантуючи при цьому їхню статистичну значущість, є актуальною задачею.

Технології аналізу даних, що базуються на застосуванні класичних статистичних підходів, мають низку недоліків. Відповідні методи ґрунтуються на використанні усереднених показників, на підставі яких важко з’ясувати справжній стан справ у досліджуваній сфері. Окрім того, стандартні статистичні методи відкидають (нехтують) нетипові спостереження — так звані піки та сплески. Проте окремі нетипові значення можуть становити самостійний інтерес для дослідження, характеризуючи деякі виняткові, але важливі явища. Навіть сама ідентифікація цих спостережень, не говорячи про їх подальший аналіз і докладний розгляд, може бути корисною для розуміння сутності досліджуваних об’єктів чи явищ. Як показують сучасні дослідження, саме такі події можуть стати вирішальними щодо майбутнього поводження та розвитку складних систем.

Ці недоліки статистичних методів спонукали до розвитку нових методів дослідження складних систем, що базуються на нелінійній динаміці, теорії катастроф, фрактальній геометрії тощо. В реальних ситуаціях задача побудови моделей на підставі навчання ускладнюється тим, що в наборах даних можуть бути приклади, в яких значення деяких атрибутів невідомі – наприклад, дані про вимірювання не були записані або були втрачені. Також часто трапляються некоректні класифікації через помилки або похибки в даних, необхідних для побудови моделі класифікації.

В основу технологій інтелектуального аналізу даних покладено концепцію шаблонів (патернів), що відбивають певні фрагменти багатоаспектних зв'язків у множині даних, характеризуючи закономірності, притаманні підвибіркам даних, які можна компактно подати у зрозумілій людині формі. Шаблони відшукують методами, що виходять за межі апріорних припущень стосовно структури вибірки та вигляду розподілів значень аналізованих показників. Важлива особливість цієї технології полягає в нетривіальності відшукуваних шаблонів. Це означає, що вони мають відбивати неочевидні, несподівані регулярності у множині даних, складові так званого прихованого знання.

WEKA - це бібліотека алгоритмів машинного навчання для вирішення проблем інтелектуального аналізу даних на реальних даних. WEKA також забезпечує середовище для розробки багатьох алгоритмів машинного навчання. Вона має набір інструментів для виконання різних завдань з видобутку даних, таких як класифікація даних, кластеризація даних, регресія, вибір атрибутів, часте видобування наборів елементів тощо. Всі ці завдання можна виконати за файлом sample.ARFF, який є у сховищі WEKA, або користувачі можуть підготувати свої файли даних. Використовуючи Java-бібліотеку WEKA, ви можете написати код, який аналізує ваші дані, і «на льоту» вносити всі необхідні корективи, замість того щоб чекати, поки хтось отримає ваші дані зі сховища, конвертує в формат WEKA і прожене їх через WEKA Explorer . Зразки файлів .arff - це набори даних, що мають вбудовані історичні дані, зібрані дослідниками. Очевидно, що він далекий від ідеалу. Оскільки WEKA є Java-додатком, то воно включає в себе Java -бібліотеки, яка може викликатися з іншої програми, що виконується на стороні сервера.

Модель класифікації можливо побудувати за навчальним набором даних (training set). Цей підхід використовує відомі дані для аналізу зв'язків значень атрибутів навчальних екземплярів даних. Таким чином, коли з'являється новий екземпляр даних (набір значень атрибутів) невідомого класу, потрібно провести аналіз значень атрибутів за допомогою побудованої моделі класифікації та визначити відповідний клас. При побудові моделі класифікації зазвичай використовують набір даних, який ділиться на дві частини. Перша частина (60-80% або 2/3 даних) використовується як навчальний набір для побудови моделі. Після цього тестові дані класифікуються за допомогою побудованої моделі та порівнюються з їх дійсними класами. Такий підхід дозволяє оцінити точність побудованої моделі класифікації. Тестова перевірка моделі класифікації дозволяє уникнути зайвого перенавчання моделі. Оскільки модель класифікації будується для класифікації некласифікованих екземплярів, при перевірці її оптимальності використовується тестовий набір

даних. Таким чином, гарантується, що побудована модель класифікації зможе з досить високою ймовірністю визначити клас ще некласифікованого екземпляру. Перевірку моделі треба провести на тестовому наборі даних «1/3» і оцінити, наскільки результати класифікації відрізняються від тестових класів.

Таким чином використання WEKA для інтелектуального аналізу даних, з необхідністю аналітичної обробки надвеликих об'ємів інформації, що накопичується в сучасних сховищах даних, дає можливість візуалізації, класифікація, кластеризація та асоціація методів математичної статистики і машинного навчання для вирішення завдань предметної галузі.

Література

1. Дубук В.І., Коцун В.І. Людино-машинний інтерфейс: Навч.-ме-тод. посіб. у формі брошури для студ. ВНЗ галузі знань 12 «Інфор-маційні технології» спеціальностей 121 «Інженерія програмного за-безпечення», 122 «Комп'ютерні науки та інформаційні техноло-гії» - Львів: Європейський університет, Львівська філія, 2018 - 70 с.
2. Дубук В.І., Чорний М.В. Розробка програмного забезпечення з графічним людино-машинним інтерфейсом в інформаційно-аналі-тичній системі оцінки ринку електричної енергії // Моде-лювання та інформаційні технології. Зб. наук. пр. ІПМЕ НАН України. - Вип. 82. - К.: 2018. - С. 33 – 40.
3. Томас Єрл, Ваджид Хаттак, Пол Булер Основи Big Data: Концепції, алгоритми та технології/Пер.з англ. Анатолія Гладуна;За наук.ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.

INTELLECTUAL DATA ANALYSIS USING WEKA EXPLORER

Hnatushenko Viktoriia, Deineka Bohdan

Abstract. The issue of intelligent data analysis was studied, the disadvantages and advantages of using weka explorer were described.

Keywords: WEKA, intellectual analysis, statistical methods, templates, attributes.

References

1. Dubuk V.I., Kotsun V.I. Liudyno-mashynnyi interfeis: Navch.-me-tod. posib. u formi broshury dlia stud. VNZ haluzi znan 12 «Infor-matsiini tekhnolohii» spetsialnostei 121 «Inzheneriia prohramnoho za-bezpechennia», 122 «Kompiuterni nauky ta informatsiini tekhnolo-hii» - Lviv: Yevropeiskyi universytet, Lvivska filiia, 2018 - 70 s.
2. Dubuk V.I., Chornyi M.V. Rozrobka prohramnoho zabezpechennia z hrafichnym liudyno-mashynnym interfeisom v informatsiino-anali-tychnii systemi otsinky rynku elektrychnoi enerhii // Mode-liuvannia ta informatsiini tekhnolohii. Zb. nauk. pr. IPME NAN Ukrainy. - Vyp. 82. - K.: 2018. - C. 33 – 40.
3. Tomas Yerl, Vadzhyd Khattak, Pol Buler Osnovy Big Data: Kontseptsii, alhorytmy ta tekhnolohii/Per.z anhl. Anatoliia Hladuna;Za nauk.red. Oleksiia Naidy. Dnipro: «Balans Biznes Buks», 2018. 320 s.