ВИЯВЛЕННЯ ЗА МІКРОДАНИМИ ФАКТОРІВ ВПЛИВУ НА ПРИЙНЯТТЯ РІШЕНЬ

Чертов Олег Романович,
доктор технічних наук, професор,
завідувач кафедри прикладної математики,
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

Зазвичай перед прийняттям відповідального рішення що пересічна людина, що особа, наділена владою, намагаються отримати консультацію чи пораду від фахівців у відповідній галузі. Проте очевидно, що якість консультаційної допомоги, яка надається, є суттєво суб'єктивною, бо спирається, насамперед, на власний досвід експерта. Добре відомим є парадокс «упередження вцілілого» (survivorship bias), що вперше був описаний ще під час Другої світової війни математиком Абрагамом Волдом (Abraham Wald), який займався вивченням розташування пробоїн, що отримували бомбардувальники, які повернулися з бойових завдань. Критично важливим виявилося дослідження всієї вибірки, а не лише її відібраної систематичним чином частини (скажімо, тільки ті літаки, що повернулися на базу) [1, с. 260].

Останніми роками значного поширення набуває спосіб надання результатів різноманітних статистичних спостережень через мікродані – певні вибірки первинних даних про респондентів. Достатньо згадати найбільш масштабний з таких проектів IPUMS-International [2], у межах якого вже зібрано та відкрито для доступу дослідникам більше 1 млрд персональних записів з 443 переписів 98 країн.

Маючи мікродані, що містять детальну інформацію про респондентів, потрібно відшукати приховані закономірності та залежності, що можуть допомогти відповісти на запитання: якими чинниками керуються люди під час процесу прийняття своїх важливих рішень (наприклад, чи мати дитину, переїхати в інше місце або залишитися тощо).

Завдання, близькі до пошуку факторів впливу на прийняття рішень, розглядаються у трьох різних напрямах досліджень з пошуку: контрастних наборів (contrast set mining) [3], початкових шаблонів (emerging pattern mining) [4] та виділених груп (subgroup discovery) [5], які наразі проводяться незалежно одне від одного, використовують різні алгоритми статистичного (машинного) навчання та застосовуються для розв'язання різних типів задач.

У своїх попередніх роботах автор разом із колегами розробив алгоритм пошуку факторів впливу на основі кластеризації [6]. Цей алгоритм нагадує системи рекомендацій, оскільки в результаті він дає набір правил (рекомендацій), які можуть спрямовувати певну соціальну групу до бажаного стану. Наприклад, якщо потрібно збільшити відсоток людей, що залишаються жити та працювати у сільській місцевості, для молодих людей

потрібно покращити доступ до Інтернет, а для більш старших – доступ до якісних навчальних закладів для дітей і до якісної медичної допомоги.

У цій роботі пропонується використовувати асоціативні правила замість кластеризації.

Будь-яке асоціативне правило можна представити як дві множини, пов'язані операцією імплікації: $A \rightarrow B$, тобто якщо має місце умова A, то внаслідок виконується B. Зазвичай кожне асоціативне правило характеризують підтримкою (support), тобто відносною кількістю випадків, які містять як умову, так і наслідок, та вірогідністю (confidence), тобто відношенням кількості випадків, що містять умову і наслідок, до кількості випадків, що містять тільки умову. Чим більше значення мають ці характеристики, тим краще.

Основна ідея роботи полягає у тому, щоб розбити початковий набір даних на дві множини: ті, записи, що містять бажані властивості, та ті, що їх не мають. Наприклад, у контексті пошуку факторів впливу на підвищення народжуваності можна виділити сім'ї з однією-двома маленькими дітьми та сім'ї з аналогічними соціально-демографічними рисами, але без дітей. Усі складові в умовах асоціативних правил можна розділити на дві групи: інваріантні (які важко чи неможливо відносно швидко змінити, наприклад, стать, вік, національність) та варіативні. Будуючи контрастні множини асоціативних правил, можна знаходити умови, які відповідають впливовим чи, принаймні, суттєво взаємопов'язаним факторам. Наприклад, аналіз даних перепису населення США 2010 р. у штаті Каліфорнія показав, що для групи молодих сімей латиноамериканського походження, де дружина має лише шкільну освіту, наявність автомобіля є суттєвим фактором для прийняття рішення про народження дитини (вірогідність перевищувала 75%). В усіх інших групах сімей цей чинник не мав великого значення.

Список використаних джерел

1. Mangel M., Samaniego F. Abraham Wald's work on aircraft survivability. *Journal of the American Statistical Association*. 1984. № 79 (386). P. 259–267.

2. Integrated Public Use Microdata Series International. Minnesota Population Center. URL: https://international.ipums.org/international/ (дата звернення: 10.11.2019).

3. Bay S. D., Pazzani M. J. Detecting change in categorical data: Mining contrast sets. *Proceedings of the 5th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM. 1999. P. 302–306.

4. Liu Q., Dong G. A contrast pattern based clustering quality index for categorical data: *Proceedings of the 9th IEEE International Conference on Data Mining*. IEEE. 2009. P. 860–865.

5. Atzmueller M. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2015. № 5 (1). P. 35–49.

6. Chertov O., Aleksandrova M. Fuzzy clustering with prototype extraction for census data analysis. *Soft Computing: State of the Art Theory and Novel Applications*. Springer, 2013. P. 289–313.