УДК 519.6

К.т.н., доцент Чертов О.Р., студентка Александрова М.В.

Національний технічний університет України
«Київський політехнічний інститут»

АЛГОРИТМ ПОШУКУ ЗАКОНОМІРНОСТЕЙ В ДАНИХ ПЕРЕПИСУ НАСЕЛЕННЯ

Abstract

Oleg Chertov, assoc. prof., PhD; Marharyta Aleksandrova, student

Pattern search algorithm for census data

This paper concerns the problem of Data Mining methods use for census data analysis. Authors proposed a novel clustering-based technique. Described Influence search algorithm allows determining factors that can affect people decision-making process.

Вступ

Інформація, отримана в результаті аналізу даних перепису, є надзвичайно важливою в багатьох галузях людської діяльності, як то: економіка, політика, соціологія тощо. Однією з особливостей сучасного стану справ в статистиці населення є представлення інформації не тільки в агрегованому, а й в первинному вигляді. Відповідні дані часто стають доступними широкому колу дослідників з усього світу через такі проекти як IPUMS-International [1].

Для аналізу переписних даних активно застосовуються статистичні методи, зокрема, регресійний, дисперсійний та нелінійний аналізи, непараметричні підходи [2]. Однак, використання таких методів вимагає від дослідника знання, або, принаймні, припущення про існування певної закономірності в масиві даних.

На відміну від статичних підходів, методи інтелектуального аналізу даних (Data Mining) дозволяють знаходити приховані закономірності в масивах даних. Серед найбільш поширених методів Data Mining можна відзначити наступні: кластеризація (класифікація), пошук асоціативних правил, дерева розв’язків, нейронні мережі, метод «найближчого сусіда» [3].

В роботах [4, 5] розглянуто побудову системи пошуку асоціативних правил в просторових базах даних перепису, а також вимоги, що висуваються до подібних систем.

На відміну від зазначених робіт, автори вважають, що методи інтелектуального аналізу даних можуть застосовуватись для роботи не тільки із агрегованими, але й із первинними даними перепису. Особливої уваги заслуговують нові підходи, що використовують методи Data Mining в якості окремого етапу обробки переписних даних.

Постановка задачі

Метою даної роботи є розробка алгоритму для визначення важелів впливу на уклад життя людей. Особливий інтерес представляє знаходження можливостей стимулювати процес прийняття бажаного рішення щодо таких базових питань як *чи варто народжувати дитину*, *чи переїздити до іншого міста*, *чи починати навчання*.

Алгоритм пошуку впливів

Для розв'язання поставленої задачі авторами було запропоновано алгоритм пошуку впливів, що базується на кластеризації. Алгоритм складається з таких кроків.

1. Виділити із початкової множини записів дві групи N_1 та N_2. Перша група повинна містити записи про респондентів, які володіють тією характеристикою, наявність якої є предметом дослідження, а друга – навпаки. Також на зазначені групи можуть накладатися додаткові умови, які, зазвичай, є природними обмеженнями з точки зору проблемної галузі.
2. Визначити ті атрибути, які можуть потенційно впливати на наявність обраної характеристики. Також необхідно окремо виділити атрибути для кластеризації, тобто такі, що є числовими, або можуть бути порівняні за допомогою чисел.
3. Кластеризувати групу N_1.
4. Виходячи із специфіки поставленої задачі, визначити інваріантні для обох груп параметри та їх границі в кожному з отриманих кластерів.
5. Використовуючи отримані межі, виділити із групи N_2 прототипи кластерів групи N_1.
6. Порівняти відповідні характеристики кластерів та їх прототипів, зробити висновки.

Для проведення кластеризації автори пропонують використовувати субтрактивний алгоритм кластеризації (*subtractive clustering algorithm*) [6].

Він є відносно швидким та належить до групи *off-line* методів кластеризації. На відміну від *on-line* методів, він не додає по одній точці множини для проведення чергової ітерації процедури знаходження центрів кластерів, а працює із усією множиною одночасно. Це дозволяє успішно обробляти статичні та об'ємні за своєю природою переписні дані.

Застосування запропонованого алгоритму

Запропонований алгоритм було застосовано для аналізу даних перепису декількох штатів США із метою визначення факторів, що можуть стимулювати підвищення народжуваності.

Для експерименту були взяті дані перепису штатів Каліфорнія, Техас та Гавайї за 2000 рік. Оскільки метою експерименту було визначення факторів підвищення народжуваності, в якості групи N_1 автори взяли множину респондентів, які мають одну або дві дитини віком до 2-х років, а до групи N_2 були віднесені сім'ї без дітей. Це надає можливість прослідкувати процес переходу сім'ї від стану бездітної до сім'ї із маленькою дитиною (дітьми), а отже визначити, які саме фактори можуть сприяти цьому переходу.

Для забезпечення достовірності отриманих результатів, а також для зменшення впливу негативних факторів, на групи були накладені додаткові обмеження: сім'ї повинні бути повними; обидва батьки повинні бути здоровими та мати найбільш сприятливий для заведення дитини вік. Після проведення додаткових досліджень було визначено, що жителі Каліфорнії найчастіше заводять дітей в віці 24-38 років для чоловіків, та 22-37 для жінок, мешканці штату Гавайї в 23-28 та 21-37 років, в Техасі відповідні вікові межі становлять 22-36 та 20-34.

Для проведення кластеризації використовувались такі параметри: вік та освіта батьків, загальний дохід батька. Дохід матері не враховувався, оскільки значна частина жінок після народження дитини йде у декретну відпустку. В результаті застосування субтрактивного алгоритму було отримано 3 кластери для штату Каліфорнія та по 4 кластери для штатів Гавайї і Техас. В кожному штаті були виділені три стандартні кластери. Перший відповідає меншій віковій групі, його представники мають найменший рівень доходів та освіти. Другий кластер містить респондентів старшого віку із найвищим рівнем освіти та доходів. Третій займає проміжне положення. Додатково було виділено кластер із представниками, що мають тільки шкільну освіту та належать до наймолодшої вікової групи (Гавайї) або в яких мати належить до молодшої вікової групи, а батько — до середньої (Техас).

В якості інваріантного для обох груп параметру було обрано вік батьків, оскільки значення інших параметрів, таких як дохід та наявність власного житла, можуть змінюватися в залежності від соціальної політики уряду.

Висновки

Після аналізу характеристик отриманих кластерів та їх атрибутів було зроблено такі висновки. Наявність власного житла та збільшення рівня доходів позитивно впливають на бажання людей заводити дитину. Проте, надання матеріального допомоги молодим парам, особливо тим, які мають високий рівень освіти, скоріше за все не сприятиме народженню дитини, оскільки найбільш освічені люди заводять дітей після 30 років. Також каліфорнійці старшої вікової групи мексиканського походження більш схильні до заведення дитини, ніж вихідці із Західної Європи. Аналогічна ситуація має місце в наймолодшій віковій групі штату Техас.

Література

1. Minnesota Population Center, University of Minnesota. Integrated Public Use Microdata Series International. [Електронний ресурс]. — Режим доступу: https://international.ipums.org/international
2. U.S. Census Bureau. Statistical Quality Standard E1: Analyzing Data [Електронний ресурс]. — Режим доступу: http://www.census.gov/quality/standards/standarde1.html
3. *Berson A., Smith S, Thearling K.* An Overview of Data Mining Techniques [Електронний ресурс]. — Режим доступу: http://www.stat.ucla.edu/~hqxu/stat19/DM-Techniques.pdf
4. Mining census and geographic data in urban planning environments / *D. Malerba, F. Lisi, A. Appice, F. Sblendorio* // Atti della Terza Conferenza Nazionale su Informatica e Pianificazione Urbana e Territoriale (INPUT 2003), June, 5—7, 2003 : proceedings. — Firenze : Alinea Editrice, 2003. — P. 36—50.
5. *Klosgen W., May M.* Census data mining — an application // Proc. 6th European Conf. Principles of Data Mining, Knowledge Discovery (PKDD'02). — 2002. — P. 65—79.
6. Generation of fuzzy rules with subtractive clustering / *A. Priyono, M. Ridwan, A. J. Alias et al.* // Jurnal Technology. — 2005. — Vol. 43 (D). — P. 143—153.