УДК 31.311.2

Олександр Григорович ХАНІН

кандидат фізико-математичних наук,
доцент кафедри алгебри та математичного аналізу,
Східноєвропейський національний університет імені Лесі Українки
E-mail: a.hanin@bigmir.net

МЕТОД X^2- КЛАСТЕРИЗАЦІЇ В ЗАДАЧАХ МАРКЕТИНГУ

Ханін, О. Г. Метод χ^2- кластеризації в задачах маркетингу [Текст] / Олександр Григорович Ханін // Економічний аналіз: зб. наук. праць / Тернопільський національний економічний університет; редкол.: О. В. Ярощук (голов. ред.) та ін. – Тернопіль: Видавничо-поліграфічний центр Тернопільського національного економічного університету «Економічна думка», 2016. – Том 26. – № 1. – С. 38-42. – ISSN 1993-0259.

Анотація

Необхідність кластеризації даних виникає в багатьох практичних задачах, зокрема задачах маркетингу. Групування в певному розумінні близьких об'єктів дозволяє підходити до вивчення кожної групи з єдиної позиції, наприклад, пропонувати єдину маркетингову стратегію. Існує чимало методів кластеризації, кожен з яких може давати свої результати, відмінні від інших. Таким чином, вважається, що кластеризація є лише методом попереднього, розвідувального аналізу даних. Нами пропонується метод кластеризації, який ґрунтується на порівнянні емпіричного та теоретичного розподілів за допомогою критерію узгодженості χ2, що дозволяє здійснювати, незалежно від розподілу генеральної сукупності, статистично обґрунтоване групування багатовимірних даних, тобто проводити кластеризацію з певною, наперед заданою ймовірністю похибки. Метод реалізований програмно в середовищі Delphi.

Ключові слова: *теоретичний та вибірковий розподіли; біноміальний розподіл; розподіл Стьюдента; критерій узгодженості χ2; довірчий інтервал; інструмент Excel «Пошук розв'язків»; кластерний аналіз; маркетинг.*

Oleksandr Hryhorovych KHANIN

PhD in Physical and Mathematical Sciences,
Associate Professor,
Department of Algebra and Mathematical Analysis,
Eastern European National University named after Lesya Ukrainka
E-mail: a.hanin@bigmir.net

X^2- CLUSTERING METHOD IN THE PROBLEMS OF MARKETING

Abstract

The need for data clustering occurs in many practical problems,in particular in the problems of marketing. The similar objects grouping allows to study each group with a unified position, for example, we can propose a unified marketing strategy. There are many clustering methods. Each of these methods can produce the results that are different from others. Thus, it is believed that the clustering method is the only preliminary, exploratory data analysis. We propose a method of clustering on the basis of a comparison of empirical and theoretical distributions using χ2 consistency test. This test allows, regardless of the population distribution, statistically valid grouping of multidimensional data. That is, we can realize clustering with some prescribed probability of error. The method is implemented in Delphi software environment

Keywords: *theoretical and sample distribution; binomial distribution; Student distribution; χ2 consistency criterion; confidence interval; Excel tool"Search of solutions"; cluster analysis; marketing.*

JEL classification: C120

Вступ

У задачах маркетингу часто виникає необхідність кластеризації даних, тобто автоматичного розбиття певної множини об'єктів (населення, товарів, підприємств, регіонів тощо) на підмножини (групи), які об'єднують схожі, у певному розумінні, об'єкти. Існує чимало методів кластеризації, які відрізняються як способом обробки даних, так і обраним ступенем схожості об'єктів, що групуються в один кластер [1]. Деякі з них реалізовані у відомих програмах обробки статистичних даних, таких, як SPSS Statistics чи STATISTICA. Так чи інакше, цим методам властивий один і той самий недолік: це методи розвідувального аналізу [2, с. 129], які дають можливість вивчити структуру даних, але не роблять жодних статистичних висновків. Запропонований нами метод, який може застосовуватися в багатьох випадках, є ідейно та обчислювально простим та дозволяє, незалежно від розподілу генеральної сукупності, формулювати певні статистичні висновки, що робить результат кластеризації статистично обґрунтованим.

Мета статті

Мета статті – на основі запропонованої нами раніше методології порівняння двох вибіркових розподілів за допомогою критерію χ^2 [3] запропонувати метод кластерного аналізу даних, що мають довільний розподіл, який дозволить на певному рівні значущості приймати статистично обґрунтовані рішення про належність об'єкта певному кластеру.

Виклад основного матеріалу дослідження

У ході будь-якого методу багатовимірної кластеризації виникає необхідність певного нормування даних по кожному виміру для того, щоб кожен з вимірів був однаково врахований при обчисленні відстані між об'єктами, тобто фактично обирається однаковий масштаб по кожному з вимірів (якщо не стоїть задача надати якомусь з них більший пріоритет). Наш метод передбачає попереднє нормування даних так, щоб вектор вимірів по кожному об'єкту становив статистичний розподіл відносних частот, тобто кожен вимір знаходився в межах від 0 до 1, а їх сума дорівнювала 1. Таким чином, відбуватиметься кластеризація об'єктів за статистичним розподілом певних ознак. Наприклад, нехай розглядаються результати маркетингового дослідження по кожному регіону України, яке полягає у виборі респондентами одного з 4 запропонованих факторів вибору як найбільш пріоритетного:

Таблиця 1. Результати маркетингового дослідження: кількість респондентів по регіонах України, які обрали певний фактор як найбільш пріоритетний

Регіони	Разом	Фактор 1	Фактор 2	Фактор 3	Фактор 4
Автономна Республіка Крим та м. Севастополь	579	24	182	107	266
Області					
Вінницька	319	69	95	8	147
Волинська	530	26	109	19	376
Дніпропетровська	1182	51	147	46	938
Донецька	847	25	295	74	453
Житомирська	169	12	75	6	76
Закарпатська	159	15	50	6	88
Запорізька	1044	82	604	49	309
Івано-Франківська	315	15	107	25	168
Київська та м. Київ	3617	220	940	360	2097
……					
Тернопільська	305	58	66	6	176
Харківська	973	75	292	61	545
Херсонська	367	57	157	22	130
Хмельницька	331	141	97	17	76
Черкаська	405	64	170	17	154
Чернівецька	228	26	96	13	93
Чернігівська	284	33	51	17	183

Для кожного регіону, поділивши кількість респондентів, які обрали певний фактор, на загальну кількість опитаних, отримаємо відносну частоту, яку будемо інтерпретувати як оцінку ймовірності, яка випадково вибрана у певному регіоні, того, що людина обере як найбільш пріоритетний певний фактор.

Таблиця 2. Розподіл відносних частот за факторами вибору (пріоритетами) за регіонами України

Разом	Фактор 1	Фактор 2	Фактор 3	Фактор 4
Автономна Республіка Крим та м. Севастополь	0,04	0,31	0,19	0,46
Області				
Вінницька	0,22	0,30	0,02	0,46
Волинська	0,05	0,21	0,04	0,71
Дніпропетровська	0,04	0,12	0,04	0,79
Донецька	0,03	0,35	0,09	0,53
Житомирська	**0,07**	**0,44**	**0,04**	**0,45**
Закарпатська	0,09	0,32	0,04	0,55
Запорізька	0,08	0,58	0,05	0,30
Івано-Франківська	0,05	0,34	0,08	0,53
Київська та м. Київ	0,06	0,26	0,10	0,58
……..				
Тернопільська	0,17	0,64	0,01	0,17
Харківська	0,19	0,22	0,02	0,58
Херсонська	0,08	0,30	0,06	0,56
Хмельницька	0,16	0,43	0,06	0,36
Черкаська	0,43	0,29	0,05	0,23
Чернівецька	0,16	0,42	0,04	0,38
Чернігівська	0,11	0,42	0,06	0,41

Було б доцільно об'єднати в один кластер такі регіони, розподіл відносних частот за факторами вибору в яких з певною наперед заданою ймовірністю збігається з деяким еталонним, теоретичним розподілом. Однак у нашій ситуації такий розподіл принципово невідомий, оскільки ми маємо справу лише з емпіричними даними.

В основу методу кластеризації покладемо ідею порівняння емпіричних $\chi 2$-розподілів [3]. На першому кроці кластеризації регіонів виберемо будь-який регіон як еталонний, наприклад Житомирську область (у таблицях виділена жирним шрифтом). Побудуємо для еталонного регіону по кожному фактору довірчий інтервал надійності 95 % для теоретичної ймовірності, що випадково опитана в цьому регіоні людина вибере як найбільш пріоритетний саме вказаний фактор (надійність може бути й іншою, але наперед заданою). Об'єм вибірки по еталонній (Житомирській) області за таблицею 1 становить $n=169$. Будемо по черзі розглядати вибір випадково взятою людиною кожного фактору як найбільш пріоритетного як «успіх», а решти – як «невдачу». Наприклад, вибір «Фактору1» вважатимемо «успіхом». Тоді ми матимемо справу з біноміальним розподілом, для теоретичної ймовірності якого легко побудувати двосторонній асимптотичний довірчий інтервал будь-якої надійності [4, с. 400-401]. Так вибіркова оцінка невідомого стандартного відхилення теоретичної ймовірності «успіху» .

$$s_n = \sqrt{\frac{w_{усп}(1-w_{усп})}{n}} \approx 0{,}002,$$

де $w_{усп} \approx 0{,}07$ – відносна частота «успіху» (див. таблицю 2).

Тоді права межа довірчого інтервалу становить

$$w_{усп} + t_{0{,}95} \cdot s_n \approx 0{,}108,$$

де $t_{0{,}95} \approx 1{,}96$ – відповідний квантиль двостороннього розподілу Стьюдента з $n-1=168$ ступенями вільності(в Excel-2010 його можна знайти за допомогою функції СТЬЮДЕНТ.ОБР.2Х(0,05;168)), ліва становить

$$w_{усп} - t_{0{,}95} \cdot s_n \approx 0{,}032.$$

Тобто з надійністю 95 % теоретична ймовірність p, що навмання вибрана у Житомирській області людина обере в якості найбільш пріоритетного «Фактор1», знаходиться в інтервалі $(0{,}032;\ 0{,}108)$.

Так само побудуємо довірчі інтервали надійності 95 % для теоретичних ймовірностей, які відповідають іншим факторам.

Таблиця 3. Межі довірчих інтервалів надійності 95 % для теоретичних ймовірностей, що відповідають факторам вибору для Житомирської (еталонної) області

Фактори вибору	Ліва межа довірчого інтервалу	Права межа довірчого інтервалу
Фактор1	0,032	0,108
Фактор2	0,365	0,515
Фактор3	0,010	0,070
Фактор4	0,375	0,525

$\chi 2$– відстань між емпіричним та теоретичним розподілом знаходиться за формулою [5, с. 454]:

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{r}\frac{(\nu_i - np_i)^2}{p_i}, \quad (1)$$

де r – кількість груп, на які розбиті дані (в нашому випадку – кількість факторів вибору, тобто $r=4$);

p_i – теоретичні ймовірності. У нашому випадку це ймовірності, що навмання вибрана людина в еталонній (Житомирській) області обре як найбільш пріоритетний *і-ий* фактор ($i=1,2,3,4$). Ці ймовірності нам невідомі, але ми встановили довірчі інтервали, в яких вони знаходяться (див. таблицю 3);

ν_i– вибіркові частоти кожної групи (в нашому випадку – кількість респондентів по іншому (не еталонному) регіону, які обрали *і-ий* фактор як найбільш пріоритетний. Див. таблицю 1);

n – об'єм вибірки (в нашому випадку n дорівнює загальній кількості опитаних респондентів по іншому (не еталонному) регіону. Див. таблицю 1).

Таким чином, ми будуємо довірчі межі певної надійності для теоретичного (еталонного) розподілу, за допомогою яких на першому кроці будемо порівнювати за критерієм χ^2 емпіричні розподіли по решті регіонів з розподілом по еталонному регіону.

В один кластер з еталонним об'єднаємо регіони, емпіричні розподіли за факторами вибору в яких значуще не відрізняються за критерієм $\chi 2$ від теоретичного розподілу для еталонного регіону (Житомирської області). Оскільки нам невідомі точні значення теоретичних ймовірностей p_i еталонного регіону (тобто теоретичний розподіл), замінимо їх на такі значення з довірчих відрізків (тобто з довірчих інтервалів, разом із їх кінцями), які зроблять значення χ^2 у виразі (1) найменшим з можливих, тим самим при перевірці гіпотези про узгодженість розподілів ми мінімізуємо похибку 1-го роду. Знайти мінімум виразу (1) по p_i (i=1,2,3,4) на декартовому добутку довірчих відрізків можна, використавши, наприклад, такий інструмент Excel, як «Пошук розв'язків» [3]. Зауважимо, що довірчі відрізки не повинні містити нульові значення. Якщо це трапилося, тобто значення відносних частот для деяких категорій дуже малі, варто об'єднати ці категорії з іншими.

Якщо отриманий мінімум є більшим за критичне значення (квантиль розподілу, оберненого до χ^2 з $r-1=4-1=3$ ступенями вільності, який відповідає ймовірності 0,95.ВExcel-2010 його можна знайти за допомогою функції ХИ2.ОБР.ПХ(0,05;3)), то розглянутий регіон не входить в один кластер з еталонним, у протилежному випадку регіони об'єднуються в один кластер. Після порівняння кожного з решти регіонів з еталонним закінчується формування першого кластера.

Наступним кроком як еталонний береться регіон, який не потрапив у перший кластер, і аналогічним чином формується другий кластер. Процес кластеризації повторюється, поки кожен регіон не потрапить у певний кластер.

Висновки та перспективи подальших розвідок

Запропонований метод дає можливість за допомогою критерію узгодженості χ^2 в багатьох практичних задачах, зокрема задачах маркетингу, швидко проводити статистично обґрунтований кластерний аналіз довільно розподілених даних з визначеною наперед ймовірністю похибки кластеризації.

Список літератури

1. Нейский, И. М. Классификация и сравнение методов кластеризации [Електронний ресурс] / И. М. Нейский. – Режим доступу: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf.
2. Пилипчук, А. В. Организация фирменных торговосбытовых систем в агропромышленном комплексе Беларуси / А. В. Пилипчук. – Минск: Ин-т системных исследований в АПК НАН Беларуси, 2011. – 178 с.
3. Ханін, О. Г. Методологічні особливості застосування критерію узгодженості χ^2 в практичних задачах економіки, соціології та маркетингу / О. Г. Ханін // Економічний аналіз. – 2015. – Том 22. – № 1. – С. 67-70.
4. Сигел, Э. Практическая бизнес-статистика / Э. Сигел. – М.: Вильямс, 2002. – 1056 с.
5. Крамер Г. Математические методы статистики / Г. Крамер. – М.: Мир, 1976. – 648 с.

References

1. Neiskiy, I. M. *Classification and comparison of the clustering methods*. Retrieved from: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf.
2. Pylypchuk, A. V. (2011). *Organization of sales and marketing systems in the agricultural sector of Belarus*. Minsk: The Institute of System Research in Agroindustrial Complex NAS of Belarus.
3. Khanin, O. G. (2015). Methodological features of $\chi 2$ consistency criteria in the practical problems of Sociology, Economics and Marketing. *Economic analysis*, 22, 1, 67-70.
4. Siegel, A. (2011). *Practical Business Statistics.* Moscow: Wilyams.
5. Kramer. H. (1999). *Mathematical Methods of Statistics*. Moscow: Mir.

Стаття надійшла до редакції 07.12.2016 р.