УДК 519.2

ЗАТОСУВАННЯ КЛАСТЕРНОГО АНАЛІЗУ ДЛЯ ОЦІНКУ РИЗИКІВ У АВТОСТРАХУВАННІ

Чистякова Т.Ю.

Науковий керівник: док. фіз.-мат. наук, професор Авраменко О.В.

Анотація. *У даній статті розглядається процедура застосування кластерного аналізу для класифікації клієнтів страхової компанії та їх автомобілів за ймовірністю настання страхового випадку, яка у подальшому оцінюється страхувальником. Кластери сформувалися з урахуванням 4-х кількісних показників і виявили клієнтів компанії та їх автомобілів, які схожі за цими показниками і мають приблизно однакову ймовірність потрапляння до ситуації страхового випадку.*

Ключові слова: ієрархічний кластерний аналіз, метод К-середніх, дендрограма.

Abstract. *This publication examines the procedure for applying cluster analysis for the classification of clients of an insurance company and their cars in the probability of occurrence of an insured event, which is subsequently assessed by the policyholder. Clusters were formed taking into account 4 quantitative indicators and found the clients of the company and their cars that are similar in these indicators and have approximately the same probability of getting into the situation of the insured event.*

Keywords: hierarchical cluster analysis, K-mean method, dendrogram.

Постановка проблеми. У статистичних дослідженнях групування первинних даних є основним прийомом розв’язування задачі класифікації, а значить і основою всієї подальшої роботи із зібраною інформацією. При наявності декількох ознак (вихідних або узагальнених) задача класифікації може бути вирішена методами кластерного аналізу, які відрізняються від інших методів багатовимірної класифікації відсутністю навчальних вибірок, тобто апріорної інформації про розподіл генеральної сукупності, яка представляю собою вектор X [1].

Аналіз досліджень і публікацій. У наукових дослідженнях кластеризація при правильному застосуванні дозволяє навіть відкривати нові перспективні напрямки [2]. Яскравим прикладом є періодична таблиця елементів: безперечною заслугою Д.Менделєєва є те, що у 1986 році він поділив 60 відомих на той час елементів на кластери або періоди за схожими

характеристиками. Вивчення причин об’єднання елементів у явно виражені кластери визначило пріоритети наукових досліджень на роки вперед. І лише через 50 років засобами квантової фізики вдалося науково обґрунтувати такий поділ [3]. На сьогоднішній день з’являється величезне коло задач, до яких доцільно застосовувати кластерний аналіз для виявлення груп схожих об’єктів. Комп’ютерні статистичні пакети дозволяють дуже швидко виконувати кластеризацію, вибравши необхідний метод [4].

Мета статті: застосувати процедуру ієрархічного кластерного аналізу для цілей аналізу ризиків у автострахуванні, а саме виокремлення класів автомобілів та їх володарів, кожен із яких відповідатиме певній групі ризику.

Виклад основного матеріалу (результатів) дослідження.

Постановка задачі. Розглянемо ряд спостережень із даними клієнтів страхової компанії, які звернулися на протязі одного тижня до компанії з метою планового страхування своїх автомобілів. Сформуємо вихідний файл даних із наступною інформацією про автомобілі та їх володарів:

- Перша змінна – марка автомобіля;
- Друга змінна – оцінкова вартість автомобіля на поточну дату;
- Третя змінна – вік водія;
- Четверта змінна – стаж водія;
- П’ята змінна – вік водія.

Маємо задачу багатовимірної класифікації. Метою даного аналізу є виокремлення класів автомобілів та їх володарів, кожен із яких відповідатиме певній групі ризику. Спостереження, які попадуть в одну групу ризику, характеризуються однаковою ймовірністю настання страхового випадку, яка у подальшому оцінюється страхувальником.

Використання кластерного аналізу для розв’язання даної задачі найбільш ефективне. У загальному випадку кластерний аналіз передбачений для об’єднання деяких об’єктів у кластери таким чином, щоб у один клас попадали максимально схожі об’єкти, а об’єкти різних кластерів максимально

відрізнялися один від одного. Кількісний показник схожості розраховується заданим способом на основі даних, які характеризують об’єкти.

Спочатку запишемо таблицю з даними 25 клієнтів страхової компанії, яка складається із перерахованих вище 5-ти показників.

Таблиця 1. Початкові дані клієнтів страхової компанії

№	Марка автомобіля	Варітість автомобіля (дол.США)	Вік водія	Стаж водія	Вік автомобіля
1	Renault Logan	6000	41	10	8
2	Daewoo Lanos	3995	56	30	9
3	ВАЗ 2106	2000	31	13	14
4	Renault Logan	4000	37	10	10
5	Mitsubishi Lancer	5000	42	7	8
6	Opel Astra	7000	50	30	11
7	Volkswagen Sharan	5500	52	27	18
8	ВАЗ 2103	1000	34	7	25
9	ВАЗ 2107	1400	40	3	18
10	ВАЗ 2101	800	29	5	31
11	Иж-2125	600	46	23	35
12	Volvo 760	3000	32	10	30
13	Peugeot 206	4000	42	1	12
14	Mercedes-Benz C-Клас	4000	25	5	24
15	Mitsubishi Galant	2900	37	19	20
16	Volkswagen Passat	900	42	23	37
17	Ford Transit	2200	23	5	21
18	Skoda Kodiaq	40000	42	7	0
19	Mercedes-Benz C-Клас	7000	32	14	18
20	Nissan Almera	4500	24	3	10
21	Ford Transit Connect	4000	27	5	11
22	Volkswagen Caddy	4900	29	7	14
23	ЗАЗ-965	400	62	35	42
24	Mazda 6	8000	37	15	6
25	Богдан 2111	3500	38	20	5

Всі кластерні алгоритми потребують оцінки відстаней між кластерами або об’єктами, і ясно, що при розрахунку відстаней потрібно знати масштаб вимірювання.

Оскільки різні вимірювання використовують абсолютно різні типи шкал (гроші, вік, стаж), то дані потрібно стандартизувати (нормалізувати), так що кожна змінна буде мати середнє 0 і стандартне відхилення 1.

Виконаємо нормалізацію усіх 5-ти змінних таблиці 1 у статистичному пакеті PASW Statistics. Всі подальші розрахунки кластерного аналізу також будемо проводити у цьому пакеті, так як розрахунку кластерного аналізу для великого масиву даних дуже громісткі.

Крок 1. Ієрархічний кластерний аналіз

На першому кроці зясуємо, чи утворюють автомобілі «природні» кластери, які можуть бути осмислені.

Найбільш важливим результатом, отриманим у результаті деревовидної (ієрархічної) кластеризації, є ієрархічне дерево. Тому у Графики вибираємо вертикальну дендрограму, як графічний варіант представлення результату кластеризації.

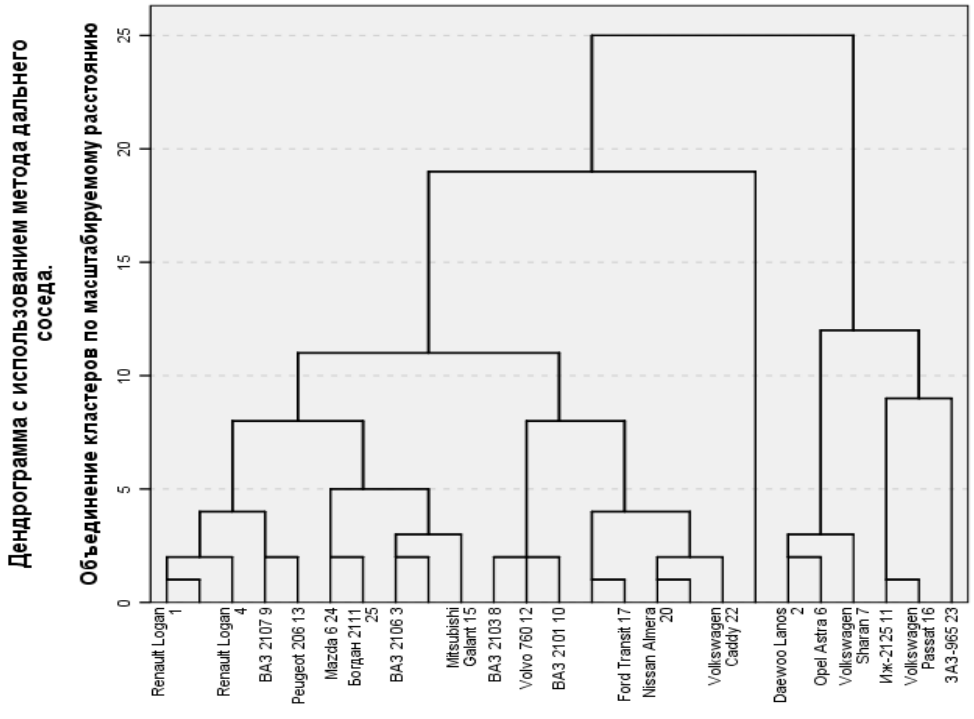

Рис. 1. Дендрограма результатів ієрархічного кластерного аналізу

Як тільки починаємо рухатися вниз, автомобілі, які «тісніше притискуються один з одним» об’єднуються і формують кластери. Кожен вузол діаграми, наведений на рисунку1, являє собою об’єднання двох або більше кластерів, положення вузлів на вертикальній осі визначає відстань, на якому були об’єднані відповідні кластери.

Відштовхуючись від візуального представлення результатів, можна зробити припущення, що автомобілі утворюють чотири природніх кластери.

Крок 2. Кластеризація методом К-середніх

У результаті ієрархічного кластерного аналізу нами було зроблено припущення, що автомобілі утворюють чотири природніх кластери. Перевіримо дане припущення, зробивши розбиття початкових даних методом К-середніх на 4 кластери, і перевіримо значущість відмінностей між отриманими групами.

У перший кластер виділився один об’єкт. Він дуже різко відрізняється від інших по ціновій політиці і по віку авто, так як автомобіль новий. Власник цього автомобіля являє собою клієнта, який характеризується низькою ймовірністю потрапляння до ситуації о страхового випадку.

Марка автомобіля	Вартість автомобіля	Вік водія	Стаж водія	Вік автомобіля	Кластер
Skoda Kodiaq	40000	42	7	0	1

У другий кластер виокремились автомобілі у досить високій ціновій політиці із середніми показниками віку автомобілів та високими показниками віку водії та їх стажу. Тобто це стабільна група клієнтів страхової компанії, водії якої мають великий стаж, а автомобілі знаходяться у гарному стані. І, відповідно, клієнти цієї групи характеризуються малою ймовірністю потрапляння до ситуації страхового випадку.

Марка автомобіля	Вартість автомобіля	Вік водія	Стаж водія	Вік автомобіля	Кластер
Daewoo Lanos	3995	56	30	9	2

Opel Astra	7000	50	30	11	2
Volkswagen Sharan	5500	52	27	18	2

У четвертий кластер виокремились об’єкти середньої цінової категорії, із відносно невисокими показниками стажу водіїв та середнім віком автомобілів. Це найбільший кластер, який представляє найбільш розповсюджений випадок клієнта страхової компанії. Клієнти цієї групи характеризуються середньою ймовірністю потрапляння до ситуації страхового випадку.

Марка автомобіля	Вартість автомобіля	Вік водія	Стаж водія	Вік автомобіля	Кластер
Renault Logan	6000	41	10	8	4
ВАЗ 2106	2000	31	13	14	4
Renault Logan	4000	37	10	10	4
Mitsubishi Lancer	5000	42	7	8	4
ВАЗ 2103	1000	34	7	25	4
ВАЗ 2107	1400	40	3	18	4
ВАЗ 2101	800	29	5	31	4
Volvo 760	3000	32	10	30	4
Peugeot 206	4000	42	1	12	4
Mercedes-Benz C-Клас	4000	25	5	24	4
Mitsubishi Galant	2900	37	19	20	4
Ford Transit	2200	23	5	21	4
Mercedes-Benz C-Клас	7000	32	14	18	4
Nissan Almera	4500	24	3	10	4
Ford Transit Connect	4000	27	5	11	4
Volkswagen Caddy	4900	29	7	14	4
Mazda 6	8000	37	15	6	4
Богдан 2111	3500	38	20	5	4

І останній кластер по ступеню ризику – це третій кластер. Автомобілі цього блоку характеризуються дуже великим віком, тобто вони дуже старі і

знаходяться у стані, близькому до аварійного. Відповідно вони характеризуються низькою вартістю. Клієнти цієї групи характеризуються високою ймовірністю потрапляння до ситуації страхового випадку.

Марка автомобіля	Вартість автомобіля	Вік водія	Стаж водія	Вік автомобіля	Кластер
Иж-2125	600	46	23	35	3
Volkswagen Passat	900	42	23	37	3
ЗАЗ-965	400	62	35	42	3

Висновки та перспективи подальших пошуків у напрямі дослідження. Отже, за допомогою кластерного аналізу нами було зроблено розбиття автомобілів та їх володарів за величиною ймовірності потрапляння до ситуації страхового випадку. Кластери сформувалися з урахуванням 4-х кількісних показників і виявили клієнтів компанії та їх автомобілів, які схожі за цими показниками і мають приблизно однакову ймовірність потрапляння до ситуації страхового випадку. Аналогічні дослідження можна проводити для аналізу ризиків не тільки у автострахуванні, а також у інших видах страхування різних об’єктів.

Список використаної літератури

1. *Дубров А.М., Мхитарян В.С., Трошин Л.И.* Многомерные статистические методы: Учебник. М.:Финансы и статистика. – 2003. – 352 с.
2. *Лупан І.В., Авраменко О.В., Акбаш К.С.* Компютерні статистичні пакети: навчально-методичний посібник. – 2-е вид. – Кіровоград: «КОД» 2015. – 236 с.
3. *Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.* Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – Спб.: БХВ-Петербург, 2007. – 384 с.
4. *Оленко А.Я.* Компютерна статистика: Навчальний посібник. – К.:Видавничо-поліграфічний центр «Київський університет». – 2007. – 174 с.