УДК: 519.237.8(045)

Т.І.Олешко, д.т.н., професор, Національний авіаційний університет
Н.В.Ратушна, асистент, Національний авіаційний університет

ВИКОРИСТАННЯ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ ДЛЯ КЛАСИФІКАЦІЇ АЕРОПОРТІВ ЗА ВИДАМИ НЕАВІАЦІЙНОЇ ДІЯЛЬНОСТІ

В статті розглянуті основні методи кластерного аналізу які можна використати для класифікації аеропортів за видами неавіаційної діяльності

Ключові слова: аеропорт, неавіаційна діяльність, класифікація, кластерний аналіз.

В статье рассмотрены основные методы кластерного анализа которые можно использовать для классификации аэропортов по видам неавиационной деятельности.

Ключевые слова: аэропорт, неавиационная деятельность, классификация, кластерный анализ.

Постановка проблеми. В прикладних економічних дослідженнях неавіаційної діяльності аеропортів виникає наукова задача класифікації об'єктів неавіаційної діяльності аеропортів певного регіону для подальшої оцінки її ефективності та грошового потоку і ранжування аеропортів за відповідною шкалою. Вирішення цієї задачі дає особі, що приймає рішення, можливість обґрунтування черговості, повноти і складу заходів для підвищення ефективності неавіаційної діяльності типових аеропортів в різних регіонах.

Аналіз останніх досліджень і публікацій. В своїх працях питання щодо ефективного функціонування аеропортів їх класифікації за різними ознаками та прогнозування їх подальшого розвитку вивчали такі вчені: Кулаєв Ю.Ф., Полянська Н.Е., Парій В.М., Омельяненко С.Л. Голубєв І.С., Костроміна О.В., та ін.[3,4].

Мета статті. Використання методів кластерного для виділення типових аеропортів, що розрізняються за видами неавіаційної діяльності у класах.

Виклад основного матеріалу. Метою вирішення задачі класифікації є виділення типових аеропортів, що розрізняються за видами неавіаційної діяльності у класах. При цьому слід мати на увазі, що спрощені інформаційні матриці відображають види неавіаційної діяльності і є інформаційним портретом аеропорту.

Кластерний аналіз містить у собі набір різних алгоритмів класифікації. Він допомагає вирішити загальне питання, як організувати дані в явні структури, тобто розгорнуті таксономії. Фактично, кластерний аналіз є не

стільки звичайним статистичним методом, скільки "набором" різних алгоритмів "розподілу об'єктів за кластерами". [1]

У загальному випадку класифікація є способом виділення підмножин аеропортів, які належать до одного класу об’єктів та більш подібні між собою, ніж об'єкти, які належать до інших класів. Класифікації потрібні настільки, наскільки вони дозволяють замінити множину елементів аеропортів, кожний з яких у якомусь ступені відрізняється від будь-якого іншого узагальненим класом, що містить деякі узагальнені значення щодо видів неавіаційної діяльності. Якщо який-небудь клас, що поєднує множину аеропортів, стійкий у часі і просторі, то він звичайно одержує власне ім'я і стає образом множини його часткових проявів.

Таким чином, у результаті класифікації вихідна розмаїтість аеропортів зменшується при мінімальній втраті змістовної інформації про неавіаційну діяльність в аеропортах. Ідеальною є класифікація, при якій за деяким кінцевим набором видів неавіаційної діяльності будь-який аеропорт може бути однозначно віднесений конкретно до одного класу. Формально це можливо, якщо множина аеропортів строго дискретна.

Якщо множини аеропортів строго безперервні, тобто для якого-небудь елементу в околиці будь-якого радіуса завжди знайдеться елемент, що належить тій же множині, то їхня однозначна класифікація неможлива. Класи в такому варіанті можуть виділяти деякі найбільш ймовірні сполучення значень ознак, але при цьому завжди будуть існувати перехідні ситуації.

Формально, максимальне число класів, які можна виділити на множині, прямо пов'язане з його ентропією чи розмаїтістю і дорівнює $2 \cdot H$. Це представлення дуже близьке до поняття числа ступенів свободи у статистиці, що пов'язується з обсягом вибірки N:

$$df = \log_2 N + 1 \qquad (1)$$

Число ступенів свободи визначає максимальну розмаїтість, яку може містити обмежена вибірка. Таким чином, число статистично обґрунтованих класів аеропортів не може бути більшим за число ступенів свободи.

Очевидно, корисно розрізняти генетичні та фізіономічні класифікації. Перші будуються на основі порівняння «подібності – розходження» фізично зрозумілих видів неавіаційної діяльності, що визначають розмаїтість станів аеропортів, другі – на основі «подібності – розходження» яких-небудь вимірних видів неавіаційної діяльності, що спостерігаються. Якщо ці ознаки дійсно визначають важливі функціональні властивості об'єкта класифікації, то фізіономічна класифікація неминуче в тому чи іншому ступені буде відображати не тільки фізіономічну подібність, але і спорідненість. Однак збіг генетичної і фізіономічної класифікації в загальному випадку не обов'язковий.

У переважній більшості випадків аеропорти можуть поділятися на класи різними способами. Вибір способу часто визначається практичними вимогами, що пред’являються до класифікації. Приймаючи неминучість

множинності класифікацій, необхідно звернути увагу на необхідність максимально чіткого обґрунтування і пояснення правил класифікації, що застосовуються до множини видів неавіаційної діяльності. Тільки на цій основі можна забезпечити їхню відтворюваність і порівнюванність.

В остаточному підсумку, в основі будь-якої класифікації так чи інакше закладені метрика і спосіб групування конкретних об'єктів класифікації. Метрика визначає спосіб виміру «подібності – розходження» порівнюваних об'єктів. Спосіб групування визначає правила, за якими класифіковані об'єкти об'єднуються в групи подібних чи класи. Після того як визначена основна схема оцінки відстані між класифікованими об'єктами, природно перейти до розгляду методів класифікації.

Ми ставимо за мету класифікувати аеропорти за видами неавіаційної діяльності, щоб змістовно описати розходження між ними. Для вирішення цієї задачі одними із найприйнятніших алгоритмів є деревоподібна кластеризація, метод K середніх та двухвходова кластеризація.

Метою деревоподібної кластеризації є об'єднання аеропортів у досить великі кластери, використовуючи деяку міру подібності чи відстані між інформаційними характеристиками неавіаційної діяльності. Типовим результатом такої кластеризації є ієрархічне дерево.

Коли дані мають просту "структуру" у термінах кластерів, подібних між собою, тоді ця структура, швидше за все, повинна бути відображена в ієрархічному дереві різними областями. У результаті успішного аналізу методом групування з'являється можливість знайти кластери (області) та інтерпретувати їх.

Метод деревоподібної кластеризації використовує при формуванні кластерів відстані між видами неавіаційної діяльності спрощеної інформаційної матриці аеропорту. Ці відстані можуть визначатися в одновимірному чи багатовимірному просторі видів неавіаційної діяльності, які виступають їх змінними–ознаками. При проведенні деревоподібної кластеризації в статистичному пакеті “Statistica v.6.0” однією із задач був вибір метрики для обчислення відстаней між об’єктами. [2]

Для оцінки розходження чи обчислення відстані між класами аеропортів щодо ознак видів неавіаційної діяльності застосовують цілий ряд метрик:

1) відстань Мінковського:

$$R(x,y)=\left(\sum_{i=1}^{m}(x_i-y_i)^p\right)^{\frac{1}{r}}, \qquad (2)$$

де: $R(x,y)$ – відстань між точками x і y; x і y - змінні, що описують множину від 1 до m; p – ступінь різниці від 1 до k; r – ступінь кореня із сум різниць у ступені p пар порівнюваних змінних від 1 до k. Зазвичай k не перевищує 3. Параметр p відповідає за поступове зважування різниць за окремими координатами (в нашому випадку – видами неавіаційної

діяльності), параметр r відповідає за прогресивне зважування великих відстаней між аеропортами. Якщо обидва параметри - r і p дорівнюють 2, то ця відстань збігається з метрикою - відстань Евкліда.

2) Евклідова відстань:

$$R(x,y)=\left\{\sum_{i=1}^{m}(x_i-y_i)^2\right\}^{\frac{1}{2}} \tag{3}$$

3) Квадрат евклідової відстані:

$$R(x,y)=\sum_{i=1}^{m}(x_i-y_i)^2 \tag{4}$$

4) Відстань Манхетен–сіті:

$$R(x,y)=\sum_{i=1}^{m}\left|x_i-y_i\right| \tag{5}$$

5) Кореляція Пірсона:

$$R(x,y)=1-r_{xy} \tag{6}$$

де r_{xy} – кореляції між двома аеропортами за ознаками i – видів неавіаційної діяльності.

У залежності від співвідношень r і p метрики відображають простори різної кривизни щодо лінійного простору Евкліда. При застосуванні відстані Манхетен-сіті віддалені точки виявляються ближче, ніж у метриці Евкліда. Навпаки, у просторі квадратичної чи кубічної метрики Евкліда, коли $p=2$ і $p=3$ відповідно, а $r=1$, віддалені точки виявляються далі, ніж у звичайній метриці Евкліда. Звідси випливає просте правило застосування цих метрик:

1) якщо розподіл значень змінних близький до нормального, то це оптимальна метрика Евкліда;

2) якщо розподіл значень змінних має дуже великий ексцес, то варто застосовувати відстань Манхетен–сіті, а в межі при дуже великому позитивному ексцесі відстань з $p=1$, при $r>1$;

3) якщо розподіл даних має дуже великий негативний ексцес і тим більше близький до рівномірного, то оптимальною є відстань Мінковського з $p>1$ і $r=1$.

Об'єднання або метод деревоподібної кластеризації використовується при формуванні кластерів неподібності або відстані між об'єктами. Ці відстані можуть визначатися в одновимірному або багатовимірному просторі. Найбільш прямий шлях обчислення відстаней між об'єктами у багатовимірному просторі полягає в обчисленні евклідових відстаней. Якщо ви маєте двох-або тривимірний простір, то ця міра є реальною геометричною відстанню між об'єктами у просторі (ніби відстані між об'єктами виміряні рулеткою). Однак алгоритм об'єднання не "піклується" про те, чи є надані для цієї відстані справжніми або деякими іншими похідними мірами відстані, що

більш важливо для дослідника, і завданням дослідників є підібрати правильний метод для специфічних застосувань. Модуль Кластерний аналіз дозволяє обчислювати різні типи відстаней, крім того, користувач може обчислити матрицю відстаней незалежно і використовувати її безпосередньо у процедурі об'єднання.

Метод К середніх істотно відрізняється від таких агломеративних методів, до яких відноситься деревоподібна кластеризація (об'єднання). Ми вже маємо гіпотези щодо числа кластерів за результатами бальної оцінки аеропортів та деревоподібної кластеризації інформаційних портретів аеропортів за видами неавіаційної діяльності. Алгоритм методу К середніх вирішує задачу утворення такого числа кластерів, щоб їх склад був настільки різний, наскільки це можливо. У загальному випадку метод К середніх будує рівно К різних кластерів, розташованих на найбільш можливих відстанях один від одного.

З обчислювальної точки зору можна розглядати цей метод як дисперсійний аналіз "навпаки". Програма пакету “Statistica v.6.0” починає з К випадково обраних кластерів, а потім змінює приналежність об'єктів до них так, щоб мінімізувати мінливість усередині кластерів та максимізувати мінливість між кластерами.

Даний спосіб аналогічний методу "дисперсійний аналіз (ANOVA) навпаки" у тому сенсі, що критерій значимості в дисперсійному аналізі порівнює міжгрупову мінливість із внутрішньогруповою при перевірці гіпотези про те, що середні значення в групах відрізняються одна від одної. У кластеризації методом К середніх програма переміщує об'єкти з одних груп (кластерів) в інші для того, щоб одержати найбільш значимий результат при проведенні дисперсійного аналізу (ANOVA).

Звичайно, коли результати кластерного аналізу методом К середніх отримані, можна розрахувати середні для кожного кластера на кожнім кроці, щоб оцінити наскільки кластери відрізняються один від одного. Значення F-статистики, отримані для кожного кроку, є іншим індикатором того, наскільки добре крок дискримінує кластери. Відповідно до цього методу було отримано кластеризацію аеропортів за видами неавіаційної діяльності

Висновки. В результаті кластерного аналізу аеропорти України вдалося розбити на чотири класи, що розрізняються за видами неавіаційної діяльності, які в них представлені.

1. *А. Бююль., П. Цефель.* Анализ статистических данных и восстановление скрытых закономерностей [Текст]. – Санкт-Петербург: ООО «ДиаСофтЮП», 2004г. – 608с.
2. *В. Боровиков*. Популярное введение в программу STATISTICA [Текст]. - .М:, 2000г. – 269с.
3. *Кулаєв Ю.Ф.* Економіка цивільної авіації України [Текст].- К.: Фенікс, 2004р. - 667 с.
4. *Полянська Н.Е.* Організація комерційної роботи на повітряному транспорті [Текст].- К.: НАУ, 2006 р. – 396 с.

Поступила 13.09.2010р.