УДК 658.012

І. П. ГАМАЮН, д-р техн. наук, проф., НТУ «ХПІ»;
О. М. БЕЗМЕНОВА, асп., НТУ «ХПІ»

ФОРМУВАННЯ ПОКАЗНИКІВ СХОЖОСТІ МІЖ ОБ'ЄКТАМИ, ЩО ХАРАКТЕРИЗУЮТЬСЯ ПАРАМЕТРАМИ, ВИМІРЮВАНИМИ В РІЗНИХ ШКАЛАХ ВИМІРУ

Як складова частина задачі кластеризації (класифікації об'єктів) розв'язується задача оцінювання ступеня подібності об'єктів, що описуються ознаками, вимірюваними в різних шкалах, а саме, у шкалі найменувань (номінальні ознаки) і в кількісній шкалі. Запропоновано коефіцієнт ступеня подібності об'єктів, описуваних номінальними ознаками. За наявності ознак, що відносяться до різних шкал, пропонується використовувати розроблений у статті комбінований коефіцієнт ступеня схожості. Зроблено висновки про властивості запропонованих показників ступеня схожості.

Ключові слова: кластеризація, класифікація об'єктів, шкала найменувань, номінальні ознаки, коефіцієнт ступеня подібності, комбінований коефіцієнт ступеня схожості.

Вступ. Добре відомо, що в суспільстві, керованому інформаційними технологіями, знання – найістотніший актив у будь-якій організації. У найрізноманітніших галузях аналіз інформації, що описує складні системи, застосовується техніка кластеризації. Дж. Хартиган [1] дав прекрасний огляд багатьох опублікованих досліджень, що містять результати, отримані методами кластерного аналізу. Наприклад, у галузі медицини кластеризація захворювань, лікування захворювань або симптомів захворювань приводить до широко використовуваної таксономії. У області психіатрії правильна діагностика кластерів симптомів, таких як параноя, шизофренія і так далі, є вирішальною для успішної терапії. У археології за допомогою кластерного аналізу дослідники намагаються встановити таксономію кам'яних знарядь, похоронних об'єктів тощо. Відомі широкі застосування кластерного аналізу в маркетингових дослідженнях. Загалом, усякий раз, коли необхідно класифікувати «гори» інформації до придатних для подальшої обробки груп, кластерний аналіз виявляється дуже корисним і ефективним. Використання методів інтелектуального аналізу в технічній діагностиці дозволяє виявляти дефекти і несправності на ранніх етапах та усувати відмови в процесі технічного обслуговування, що підвищує надійність і ефективність експлуатації, а також дає можливість експлуатації технічних систем відповідального призначення.

Кластерний аналіз є не стільки звичайним статистичним методом, скільки «набором» різних алгоритмів «розподілу об'єктів по кластерах». Існує точка зору, що на відміну від багатьох інших статистичних процедур, методи кластерного аналізу використовуються у більшості випадків тоді, коли ви не маєте яких-небудь апріорних гіпотез відносно класів, але все ще

знаходитеся в описовій стадії дослідження. Слід розуміти, що кластерний аналіз визначає «найбільш можливо значимий розв'язок» [1].

Успішне застосування інтелектуального аналізу даних у таких вельми помітних областях, як електронна комерція, маркетинг та роздрібна торгівля призвела до його застосування в інших галузях і секторах. Серед цих галузей лише розкривається охорона здоров'я. Охорона здоров'я навколишнього середовища все ще залишається інформаційно багатою, але бідною на знання галуззю [2]. Медична діагностика розцінюється як важливе, поки що складне завдання, яке треба виконувати точно і ефективно. Автоматизація цієї системи була б надзвичайно вигідна. Нажаль лікарі не мають компетенції у кожній вузькій спеціалізації і, окрім того, існує нестача людських ресурсів в певних місцях. Тому, автоматична медична система діагностики, ймовірно, була б надзвичайно вигідна, зводячи усіх їх разом. Відповідна комп'ютерна інформаційна система і/або системи забезпечення прийняття рішень може допомогти в досягненні клінічних випробувань за менші кошти. Існує безліч даних, доступних в рамках системи охорони здоров'я. Тим не менш, існує недостатня кількість ефективних інструментів аналізу, направлених на виявлення прихованих взаємозв'язків та тенденцій в даних.

Проблема оцінювання ступеня схожості об'єктів. Важливим у цьому відношенні є питання, пов'язане з вибором системи параметрів, що описують стан об'єктів, а також показників ступеня схожості. Досить опрацьованими є методи оцінювання ступеня зв'язку між параметрами (див., наприклад, [3, 4]). Задачі класифікації об'єктів є, в деякій мірі, двоїстими по відношенню до групування параметрів. Якщо об'єкти описуються ознаками, що носять кількісний характер, то ступінь їх подібності найчастіше оцінюється за допомогою різного роду мір, що базуються на відстані Евкліда. Однак у таких галузях науки, як медицина, психологія, соціологія і т. п. значна частина параметрів має якісний характер і вимірюється в дискретних шкалах – шкалі найменувань і порядковій шкалі. При цьому об'єкт може характеризуватися параметрами з різних шкал.

Далі припускається, що досліджувана множина об'єктів може описуватися системою ознак, описуваних з використанням як кількісних, так і якісних шкал. Якщо кожен об'єкт описується кількісними властивостями (ознаками), то він може бути представлений як точка в багатогвимірному просторі, а схожість з іншими об'єктами визначатиметься як відповідна відстань.

При класифікації використовуються різні міри відстані між об'єктами. Для кількісних ознак оцінювання ступеня подібності між об'єктами можна здійснювати на основі якої-небудь модифікації відстані Евкліда. Оскільки ознаки мають різні розмірності, попередньо необхідно здійснювати їх нормування з метою отримання безрозмірних величин, змінюваних в одному діапазоні для забезпечення можливості їх співвіднесення. У той же час наявність якісних ознак вимагає використання інших підходів при оцінюванні ступеня схожості.

Коефіцієнт ступінь схожості об'єктів для номінальних ознак. Нехай є система з n об'єктів O_1, O_2, …, O_n, описуваних m ознаками, з них ознаки $X^{\text{н}}_1$, $X^{\text{н}}_2$, …, $X^{\text{н}}_{m_{\text{н}}}$ вимірюються в кількісній шкалі, а ознаки $X^{\text{к}}_1$, $X^{\text{к}}_2$, …, $X^{\text{к}}_{m_{\text{к}}}$ – у шкалі найменувань ($m_{\text{н}} + m_{\text{к}} = m$). Будемо вважати, що ця система описується матрицею $\boldsymbol{X}$ розміром $n \times m$, що має наступний вигляд:

$$\boldsymbol{X} = \begin{array}{c} \begin{matrix} X^{\text{н}}_1 & X^{\text{н}}_2 & \dots & X^{\text{н}}_{m_{\text{н}}} & X^{\text{к}}_1 & X^{\text{к}}_2 & \dots & X^{\text{к}}_{m_{\text{к}}} \end{matrix} \\ \begin{pmatrix} x^{\text{н}}_{11} & x^{\text{н}}_{12} & \dots & x^{\text{н}}_{1m_{\text{н}}} & x^{\text{к}}_{11} & x^{\text{к}}_{12} & \dots & x^{\text{к}}_{1m_{\text{к}}} \\ x^{\text{н}}_{21} & x^{\text{н}}_{22} & \dots & x^{\text{н}}_{2m_{\text{н}}} & x^{\text{к}}_{21} & x^{\text{к}}_{22} & \dots & x^{\text{к}}_{2m_{\text{к}}} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ x^{\text{н}}_{n1} & x^{\text{н}}_{n2} & \dots & x^{\text{н}}_{nm_{\text{н}}} & x^{\text{к}}_{n1} & x^{\text{к}}_{n2} & \dots & x^{\text{к}}_{nm_{\text{к}}} \end{pmatrix} \end{array} \begin{matrix} O_1 \\ O_2 \\ \dots \\ O_n \end{matrix}$$

Якщо обмежитися тільки номінальними ознаками, то як міру відмінності між двома об'єктами O_k і O_l можна використовувати аналог відстані Хеммінга:

$$H(O_k, O_l) = \sum_{j=1}^{m_{\text{н}}} (x^{\text{н}}_{kj} = x^{\text{н}}_{lj}), \quad j = \overline{1, m_{\text{н}}} .$$

Пропонується як міру схожості використовувати нормований показник h_{kl} виду

$$h_{kl} = 1 - \frac{1}{m_{\text{н}}} H(O_k, O_l), \quad k, l = \overline{1, m} .$$

Легко довести, що побудований таким способом показник ступеня схожості має такі властивості:

- $0 \le h_{kl} \le 1$;
- $h_{kl} = 1$ тоді і тільки тоді, коли для всіх значень j з діапазону 0, 1, …, $m_{\text{н}}$ виконується співвідношення $x^{\text{н}}_{kj} = x^{\text{н}}_{lj}$, тобто у разі абсолютного збігу відповідних значень всіх номінальних ознак, що описують об'єкти O_k і O_l (зокрема, цьому значенню дорівнюватиме ступінь подібності будь-якого об'єкта з самим собою);
- $h_{kl} = 0$ тоді і тільки тоді, коли для всіх значень j з діапазону 0, 1, …, $m_{\text{н}}$ виконується співвідношення $x^{\text{н}}_{kj} \ne x^{\text{н}}_{lj}$.

Таким чином, показник h_{kl} є коефіцієнтом схожості, аналогічним до деякої міри модулю коефіцієнта кореляції, який використовується для оцінки та порівняння ступеня лінійних зв'язків між кількісними ознаками.

Комбінований коефіцієнт ступеня схожості. Нехай величина δ_{kl} оцінює ступінь схожості між об'єктами O_k і O_l, що описуються $m_к$ кількісними ознаками, причому будемо вважати, що вона має властивості, аналогічні властивостям коефіцієнта h_{kl}.

У якості агрегованої міри схожості між об'єктами O_k і O_l, що характеризуються кількісними та номінальними ознаками, пропонується використовувати величину, яка визначається за формулою:

$$\varsigma_{kl} = \frac{m_к \delta_{kl} + m_н h_{kl}}{m_к + m_н}.$$

Висновки. Для оцінювання ступінь схожості об'єктів, описуваних ознаками, вимірюваними в шкалах різних типів, запропоновано формувати коефіцієнт ступеня схожості за допомогою комбінування коефіцієнтів ступіня схожості, побудованих для кожного з типів ознак окремо. Для об'єктів, описуваних номінальними ознаками, розроблений коефіцієнт ступеня схожості, властивості якого аналогічні властивостям коефіцієнта кореляції. Цей показник дозволяє не тільки оцінювати ступінь схожості об'єктів, а й, будучи розподіленим на інтервалі від 0 до 1, забезпечує можливість порівняння ступенів схожості навіть у тому випадку, коли пари об'єктів описуються різними наборами ознак.

Список літератури: 1. *Hartigan J. A.* Clustering Algorithms / *J. A. Hartigan*. – New York : Wiley, 1975. – 351 p. **2.** *Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni*. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction / *Soni Jyoti, Ansari Ujma, Dipesh, Sharma, Soni Sunita* // International Journal of Computer Applications. – Volume 17. – No. 8, March 2011. **3.** *Елисеева И. И.* Статистические методы измерения связей / *И. И. Елисеева*. – Л. : Изд-во Ленингр. гос. ун-та, 1982. – 136 с. **4.** *Елисеева И. И.* Группировка, корреляция, распознавание образов / *И. И. Елисеева, В. О. Рукавишников*. – М. : Статистка, 1977. – 144 с.

Bibliography (transliterated): 1. Hartigan, J. A. Clustering Algorithms. New York: Wiley, 1975. Print. **2.** Jyot, Soni, et al. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction." International Journal of Computer Applications 17.8 (2011). Print. **3.** Eliseeva, I. I. Statisticheskie metody izmerenija svjazej. Leningrad: Izd vo Leningr. gos. un-ta, 1982. Print. **4.** Eliseeva, I. I., and V. O. Rukavishnikov. Gruppirovka, korreljacija, raspoznavanie obrazov. Moscow: Statistka, 1977. Print.

Надійшла (received) 05.02.2014