УДК 004.9: 004.423.24, 004.78:025.4.036

С. В. Бєлай, В. Е. Лісіцин

МОДИФІКАЦІЯ МЕТОДУ *k*-СЕРЕДНІХ КЛАСТЕРНОГО АНАЛІЗУ У ЗАДАЧАХ ПРОГНОЗУВАННЯ КРИЗОВИХ ЯВИЩ СОЦІАЛЬНО-ЕКОНОМІЧНОГО ПОХОДЖЕННЯ

Розглянуто особливості використання механізму атрибутивних відстаней об'єктів у геоінформаційних системах. Проаналізовано класичні методи кластеризації за технологією k-середніх, виявлено їх недоліки, запропоновано модифікований підхід до використання в електронних картах геоінформаційних систем

К л ю ч о в і с л о в а: атрибутивні відстані, методи кластеризації, кризові явища, геоінформаційна система, електронна карта.

Постановка проблеми. Сьогодення яскраво демонструє чисельні кризові явища соціально-економічного походження. Громадяни в різних країнах світу заради кращого життя готові на будь-які радикальні дії, що є за межею закону. За таких умов органам державної влади та силам охорони правопорядку необхідно мати інструментарій для аналізування та прогнозування розвитку кризових явищ у регіонах держави. З цією метою в розвинених країнах широко застосовують різноманітні методи збирання, статистичного аналізу і обробляння даних про протестну активність населення. Створюють кодовані та індексовані бази даних, що містять інформацію стосовно соціально-економічних явищ.

Деякі з таких баз даних є відкритими і загальнодоступними для використання, наприклад, інформація про всі (незалежно від тематики та чисельності) протестні дії на території України [1], зібрана за допомогою міжнародного фонду "Відродження". Також розробляються програмні комплекси аналізування та прогнозування кризових ситуацій. У більшості випадків вони є закритими для загального доступу. Прикладом відкритої системи є американська комп'ютерна система "Наутилус", яка здатна за даними засобів масової інформації розробляти прогнози щодо розвитку подій у зазначеному регіоні [2]. Провідні держави також створюють інформаційно-аналітичні системи для пошуку тематичних текстів та аналізу текстової інформації, такі як "RCO КАОТ" [3, с. 237–238], "Галактика-ZOOM" [4, с. 37], "Convera Retrieval Ware" [5] та ін.

У виборі універсальної інформаційно-аналітичної системи відіграє важливу роль наявність чи відсутність вихідних кодів програмних продуктів, закритість більшості подібних інформаційно-аналітичних систем відповідно до національних інтересів держав-розробників, а також імовірність навмисного закладення свідомої похибки в програмний продукт.

Наведені факти актуалізують дослідження з розроблення сучасних методів аналізування та прогнозування розвитку кризових явищ соціально-економічного походження в Україні.

Метою статті є проведення аналізу поняття "атрибутивні відстані", розгляд існуючого методу кластеризації на базі алгоритму *k*-середніх, аналіз його недоліків та запропонування альтернативного підходу.

Виклад основного матеріалу. З метою прогнозування кризових явищ у сучасному суспільстві необхідно мати простий, але ефективний механізм, який дозволив би проводити аналіз стану суспільної обстановки, виходячи зі статистики інформаційних потоків у Інтернет-середовищі стосовно протестних настроїв населення та показників рівня життя населення в регіонах держави.

Найбільш придатною платформою для розроблення зазначеного механізму є геоінформаційні системи, які надають можливість поєднання математичних та статистичних методів оцінювання і прогнозування соціальних подій із сучасними технологіями збирання даних, нанесення і оброблення геопросторової інформації на електронну карту. Платформою розробки була вибрана геоінформаційна система "Інструмент" [6], на базі якої розроблена модель "Аналітика".

Для розрахунку статистичних показників подій, що поєднуються у групи за певними ознаками та значеннями атрибутів, необхідно порівнювати об'єкти, нанесені на карту. Крім того, постає питання групування подій навколо деяких центрів у багатовимірному просторі атрибутів. Завдяки присутності геопросторової та часової складової у даних, питання отримання статистичних характеристик не може бути зведене до простого розрахунку дисперсії та середнього, для цього потрібен комплексний підхід. Задача щодо такого порівняння може виникнути у випадках:

– вибору регіонів із найбільш загрозливим соціальним станом та найшвидше зростаючою статистикою протестних подій;

– порівняння соціального стану у найбільш загрозливих регіонах з іншими та створення критеріїв оцінювання розвитку ситуації на визначеній території;

– створення груп (кластерів) протестних подій на основі значень їх атрибутів та віднесення цих кластерів до того чи іншого рівня соціальної загрози.

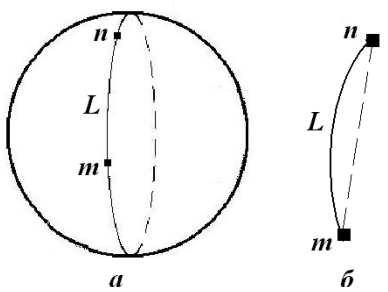

Рис. 1. Відстань L між двома точковими подіями m та n: *а* – на поверхні земної кулі; *б* – після трансформування з відповідної проекції у площину аркуша електронної карти

Розглянемо деякі особливості даних, які використовуються у моделі. Події, що наносять на карту, описуються як точкові об'єкти векторного шару. На електронній карті вони визначаються двома (X, Y) або трьома (X, Y, Z) координатами. У системі координат карти можуть бути розраховані відстані між довільно вибраними парами таких подій (рис. 1).

Для двох точок m та n, що описуються у прямокутній системі координат на площині аркуша парами координат $(x_m,\ y_m)$ та $(x_n,\ y_n)$, визначається евклідова відстань:

$$LE_{m,n} = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}. \quad (1)$$

Більш цікавим є створення груп об'єктів не тільки за значеннями геопросторових координат. Важливішим є питання групування об'єктів за значеннями у реляційній таблиці векторного шару або сервера MS-SQL. У цьому контексті простежимо декілька можливих алгоритмів розрахунку відстаней, що базуються на значеннях атрибутів.

Атрибути	*x1*	*x2*	*x3*	*x4*	*x5*
Об'єкт i	*10*	*0.7*	*100*	*10*	*1*
Об'єкт j	*8*	*0.4*	*100*	*5*	*1*

Рис. 2. Реляційна таблиця двох об'єктів з п'ятьма числовими атрибутами для кожного з них

Розглянемо алгоритм порівняння двох об'єктів, атрибути яких мають виключно числові типи (рис. 2).

Для формулювання міри схожості двох об'єктів i та j, які містять k числових атрибутів $x_{1\ldots k}$, можна застосувати евклідову відстань у вигляді

$$LE_{i,j} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2}. \quad (2)$$

У тому разі, коли атрибути об'єктів різняться ступенем важливості, у формулу (2) додають вагові коефіцієнти $w_{1\ldots k}$:

$$LE_{i,j} = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \ldots + w_k(x_{ik} - x_{jk})^2}. \quad (3)$$

Окрім того, у багатьох випадках алгоритми отримання відстаней між атрибутами двох об'єктів повинні враховувати однорідність одиниць виміру для цих атрибутів. Інакше кажучи, необхідно, щоб у формулах (2) та (3) були величини одного порядку. Тобто доцільно привести значення атрибутів до одного рівня. Це можливо, наприклад, за допомогою min-max нормалізації [7]. Для цього за допомогою мінімального (new_min_A) та максимального (new_max_A) значень визначають новий діапазон, у якому змінюватиметься атрибут A після перетворення. Наприклад, ми ставимо мету, щоб після перетворень усі значення атрибута A знаходилися у діапазоні (0…1). Тоді $(\text{new\_min}_A) = 0$, а $(\text{new\_max}_A) = 1$. Існуючі до перетворення межі змінювання діапазону значень для атрибута A позначимо як $\min_A$ та $\max_A$. Тоді min-max нормалізація відобразіть значення x_k атрибута A у нове значення x'_k таким чином:

$$x'_k = \frac{x_k - \min_A}{\max_A - \min_A}(\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A. \quad (4)$$

Якщо $\text{new\_min}_A = 0$, а $\text{new\_max}_A = 1$, то формула (4) спрощується:

$$x'_k = \frac{x_k - \min_A}{\max_A - \min_A}. \quad (5)$$

Тепер розглянемо атрибути бінарного типу. У реляційній таблиці векторного шару або сервера MS-SQL такий атрибут може приймати тільки два значення, наприклад, 1 або 0. До речі, більшість атрибутів для подій, що додаються на карту у моделі "Аналітика", – склад учасників, тактика їх дій та тип події, можуть бути описані за допомогою саме бінарного типу. Як оцінити схожість двох подій, що описуються за допомогою тільки бінарних атрибутів? Для цього порівнюють кожну відповідну пару атрибутів для двох зазначених об'єктів. Визначається так звана жаккардова відстань (Jaccard distance). Щоб обчислити її, треба попередньо підрахувати чотири параметри [8]:

– C_{11} – кількість усіх пар бінарних атрибутів, значення у яких дорівнює 1 у першому об'єкті та водночас дорівнює 1 у другому об'єкті;

– C_{10} – кількість усіх пар бінарних атрибутів, значення у яких дорівнює 1 у першому об'єкті та водночас дорівнює 0 у другому об'єкті;

– C_{01} – кількість усіх пар бінарних атрибутів, значення у яких дорівнює 0 у першому об'єкті та водночас дорівнює 1 у другому об'єкті;

– C_{00} – кількість усіх пар бінарних атрибутів, значення у яких дорівнює 0 у першому об'єкті та водночас дорівнює 0 у другому об'єкті.

Тоді загальна кількість комбінацій для усіх пар атрибутів складає:

$$C = C_{10} + C_{01} + C_{11} + C_{00}. \quad (6)$$

Тепер визначимо жаккардову відстань між об'єктами *i* та *j*:

$$LJ_{i,j} = \frac{C_{10} + C_{01}}{C_{11} + C_{10} + C_{01}}. \quad (7)$$

У виразі (7) відсутній параметр C_{00}. Справа у тому, що ця формула для оцінювання відмінності двох об'єктів справедлива у тому разі, колі бінарні атрибути асиметричні [7], тобто враховується тільки значення 1, яке визначає наявність ознаки. У структурі моделі "Аналітика" бінарні атрибути, що описують протестну подію, мають саме таку асиметричну сутність. У статистичних розрахунках нас цікавить тільки наявність тієї чи іншої ознаки у класифікаторі подій. Якщо ж обидва значення 1 та 0 однаково важливі для опису атрибута, то у формулу розрахунку відстані додається параметр C_{00}:

$$LJ_{i,j} = \frac{C_{10} + C_{01}}{C_{11} + C_{10} + C_{01} + C_{00}}. \quad (8)$$

Ще один тип атрибутів у даних моделі "Аналітика" для групування об'єктів та оцінювання їх відмінності, це впорядковані (ordinal) атрибути. Наприклад, маємо опис рівня соціально-економічного стану у регіоні у вигляді низки вербальних параметрів: "не визначений", "нормальний", "напружений", "загрозливий", "критичний". Впорядкуємо параметри, приписавши кожному з них числовий номер (ранг), який зростає у разі збільшення ступеня загрози: "не визначений" – 1; "нормальний" – 2; "напружений" – 3; "загрозливий" – 4; "критичний"– 5.

Ранг для *i*-го атрибута об'єкта змінюється від 1 до M_k, де M_k – кількість вербальних параметрів, що описують цей атрибут; k – номер атрибута (стовпчика) у реляційній таблиці (із обов'язковим припущенням того факту, що всі атрибути в таблиці мають впорядкований тип). Оскільки для іншого атрибута кількість M_k параметрів може відрізнятися, доцільно, перш ніж порівнювати об'єкти, провести min-max нормалізацію для рангів впорядкованих атрибутів. Для цього у формулі (5) відобразимо ранги у нові значення Z_{ik} на інтервалі [0…1] :

$$Z_{ik} = \frac{r_{ik} - 1}{M_{ik} - 1}. \quad (9)$$

Після цього можна застосовувати формули (3) або (4) для того, щоб порівняти між собою два об'єкти *i* та *j*.

У тому випадку, коли у реляційній таблиці містяться атрибути різних типів, необхідно за допомогою нормалізації привести усі значення до безрозмірного загального вигляду, а потім використовувати рівняння (3 – 9).

Відповідно до такої методики за допомогою редактора ГІС "Інструмент" створюються декілька "зразкових" регіонів, вибірка подій у яких відповідає поняттю типової соціально-економічної ситуації. Це можуть бути регіони, в яких упродовж минулого часу вже траплялися найзапекліші прояви протестів. Ситуації у інших регіонах порівнюватимуся із цими зразками. Якщо у вибраному інтервалі часу соціально-економічний стан у регіоні, що аналізується, близький до стану у "зразковому" регіоні, то можна говорити про ідентичність ситуацій у цих двох регіонах.

Групи подій із схожими значеннями атрибутів у цьому контексті називатимемо кластерами. Існує багато алгоритмів формування кластерів у багатовимірному просторі ознак. Для моделі "Аналітика" може бути вибраний метод *k*-середніх (*k*-means clustering). Розглянемо його сутність.

Почнемо з визначення та створення k порожніх кластерів, на які буде поділена вся множина об'єктів, що підлягають обробці. На цьому кроці зі всієї сукупності довільно вибирають k об'єктів. Кожен з них призначається одному із створених кластерів. Таким чином після завершення цього етапу у наявності маємо k кластерів, кожен з яких містить тільки один об'єкт. Відповідний об'єкт є центром такого кластера (рис. 3).

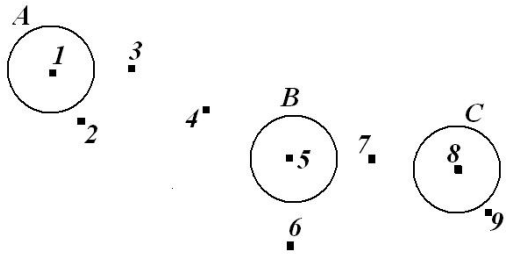

Рис. 3. Перший етап обробки. $k = 3$ – кількість кластерів, на які буде поділена уся множина об'єктів. Об'єкти 1, 5 та 8 – центри новоутворених кластерів

На наступному кроці для всіх об'єктів, що залишились і не були додані у жодну групу, розраховують відстані до центрів новостворених кластерів. Кожний такий об'єкт призначається кластеру, відстань до центра якого найменша. Після закінчення другого етапу межі кластерів змінюються, а координати центрів перераховують з урахуванням доданих об'єктів (рис. 4).

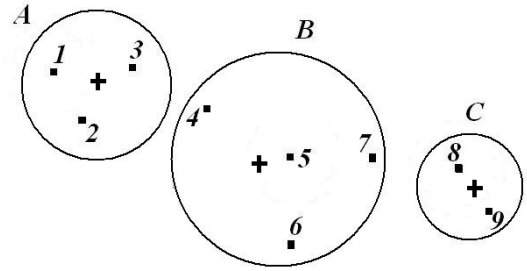

Рис. 4. Другий етап кластеризації. Кожний об'єкт належить одній з трьох груп. Нові центри кластерів позначені хрестиками

У цьому прикладі мова йде про координати центрів, що визначаються парою значень (X, Y) на площині. Для розрахунку координат центра такого кластера слід розглянути k об'єктів, що формують його як матеріальні точки із масами m_i. Сумарна маса M всіх об'єктів визначається як $\sum_{i=0}^{k} m_i$. Тоді координати (X_c, Y_c) центра кластера надаються у вигляді центра мас точкових матеріальних об'єктів на площині:

$$\left.\begin{aligned} X_c &= \frac{\sum x_i m_i}{M} \\ Y_c &= \frac{\sum y_i m_i}{M} \end{aligned}\right\}. \qquad (10)$$

Такий підхід можна застосовувати і в тому випадку, коли кластери створюються не у двовимірній геометричній площині, а у багатовимірному просторі атрибутів. Кожен атрибут визначає додаткову координату. Величина маси точкового матеріального об'єкта у цьому випадку зазвичай дорівнює 1 (якщо не враховуються вагові коефіцієнти).

На практиці постає низка питань, пов'язаних із ефективністю застосування такого методу та інтерпретацією отриманих результатів. Як кожен кластер, що створюється за методом *k*-середніх, пов'язаний із загальним соціально-економічним станом у регіоні? Що взагалі описують об'єкти, які

об'єднані за критеріями мінімальної атрибутивної відстані, і як ефективно на рівні вихідних даних керувати процесом такої кластеризації? Усі ці питання підводять до одного – які назви слід дати отриманим кластерам і як співвіднести ці назви із можливими соціально-економічними станами у регіоні?

Крім того, кластеризація за методом *k*-середніх потребує потужних комп'ютерних ресурсів у тому випадку, коли кількість об'єктів, що обробляються, занадто велика. Такі обмеження класичного методу *k*-середніх призводять до необхідності розроблення модифікованого алгоритму із більш наочними результатами групування протестних подій на карті. Перш ніж розглянути його, введемо таке поняття, як шаблонна ситуація, або шаблон.

Припустимо, що соціально-економічний стан у регіоні характеризується кількома вербальними категоріями або характеристиками рівня складності оперативної обстановки у сфері охорони громадського порядку [9]: звичайний стан, складний стан, кризовий стан, надзвичайний стан.

Для опису кожного такого стану з бази даних експерт вибирає чи штучно створює об'єкт-подію, яка за змістом атрибутів найбільш відповідає поточній категорії. За допомогою атрибутів вибраного об'єкта створюється шаблон. Зростання або збереження на високому рівні упродовж часу кількості подій у регіоні, віднесених до певного шаблону, свідчить про тенденцію формування у цьому регіоні певного стану. Наприклад, шаблон для опису критичного суспільного стану у відповідному регіоні може містити пари атрибутів та їх значень, які наведено у таблиці.

Т а б л и ц я

Можливі пари атрибутів-значень для шаблону загрозливого стану соціально-економічної ситуації в регіоні

Назва атрибута	Тип атрибута	Значення атрибута
Протести з приводу незаконних дій посадовців	бінарний	1
Протести, пов'язані із затримкою заробітної плати	бінарний	1
Протести супроводжуються деструктивними та насильницькими діями	бінарний	1
Протести супроводжуються перекриттям доріг	бінарний	1
Протести супроводжуються мітингами та маршами	бінарний	1
Кількість учасників	ціле число	200

Кожна подія, що міститься у базі даних, за критеріями мінімальної відстані між відповідними атрибутами послідовно порівнюється із кожним створеним шаблоном. Для кожного шаблону вже визначена певна категорія соціально-економічного стану, тому за результатами всього одного порівняння подія буде відразу віднесена до необхідної категорії соціально-економічного стану. Тобто значно спрощується інтерпретація результатів: кожен кластер, отриманий у результаті, пов'язаний із певним шаблоном і відповідає одному із можливих станів. Зникає потреба у багатьох ітераціях та коригуванні координат центра кожного кластера, для кожної події здійснюється тільки один цикл розрахунків.

Висновки

1. Платформою для розроблення засобів моніторингу кризових явищ була вибрана ГІС "Інструмент", на базі якої розроблена відповідна модель "Аналітика", що поєднала в собі методи збирання даних, оцінювання та прогнозування кризових явищ. Це дає можливість поєднати геопросторові і статистичні методи та додатково візуально аналізувати ситуації на електронній карті.

2. У моделі "Аналітика" обґрунтовано вибір атрибутивних відстаней між об'єктами, що аналізуються у вигляді евклідової метрики з поправкою на *n*-мірний простір атрибутів.

3. Аналіз класичного методу кластеризації *k*-середніх виявив значні проблемні питання його використання у моделі "Аналітика". Розроблений модіфікований метод кластерного аналізу *k*-середніх

дозволяє віднести подію до необхідної категорії суспільного стану у регіоні та здійснити первинний прогноз ситуації.

4. Подальші розробки доцільно присвятити перевірці коректності моделі "Аналітика" на статистичних даних протестних подій в Україні та вдосконаленню механізму моніторингу кризових явищ в українському суспільстві.

Список використаних джерел

1. Протесты, победы и репресии в Украине: результаты мониторинга, октябрь 2009 – сентябрь 2010 [Текст]. – К. : Центр исследования общества, 2011. – 64 с.

2. Суперкомпьютер способен предсказывать будущее [Электронный ресурс]. – Режим доступа : http://www.segodnya.ua/news/14287696.html (дата обращения: 12.12.13). – Название с экрана.

3. Розробка форм і способів інформаційної боротьби при виконанні внутрішніми військами Міністерства внутрішніх справ України службово-бойових завдань [Текст] : звіт про науково-дослідну роботу; кер. В. І. Воробйов / Акад. ВВ МВС України. – Х. : 2009. – 312 с.

4. Ландэ, Д. Инструментарий аналитика. Корпоративные решения. Мониторинг информации [Электронный ресурс]. – Режим доступа : http://poiskbook.kiev.ua/art/telecom0410/telecom0410.pdf (дата обращения: 12.12.13). – Название с экрана.

5. Новейшие сетевые технологии [Электронный ресурс]. – Режим доступа : http://www.ant.kiev.ua/convera/convera%20RV.html (дата обращения: 12.12.13). – Название с экрана.

6. Дробаха, Г. А. Створення просторових даних для електронних карт геоінформаційної системи внутрішніх військ МВС України. [Текст] / Г. А. Дробаха, Л. В. Розанова, В. Е. Лісіцин. – Х. : Акад. ВВ МВС України, 2012. –200 с.

7. Han Jiawei, Data minig. Concepts and techniques. Third edition. [Текст] / Jiawei Han, Micheline Kamber, Jian Pei / Morgan Kaufmann Publishers, MA, USA, 2012, 703 p.

8. Myatt Glenn J., Making sense of data I. Second Edition. [Текст] / Glenn j. Myatt, Wayne P. Johnson / WILEY, New Jersey, USA, 2014, 235 p.

9. Бацамут, В. М. Оцінювання стану оперативної обстановки у сфері охорони громадського порядку [Текст] : монографія / В. М. Бацамут, С. В. Бєлай. – Х. : Акад. ВВ МВС України, 2013. – 155 с.

Стаття надійшла до редакції 24.11.2014 р.