ДОСЛІДЖЕННЯ. РОЗРОБКИ. ПРОЕКТИ

УДК 332.13:346.57

Сергій Бєлай
старший науковий співробітник
Національної академії Національної гвардії України, к.військ.н.

Володимир Лісіцин
докторант; науковий співробітник науково-дослідного центру
Національної академії Національної гвардії України

ВИКОРИСТАННЯ МОДИФІКОВАНОГО МЕТОДУ K-СЕРЕДНІХ КЛАСТЕРНОГО АНАЛІЗУ В ЗАДАЧАХ ДЕРЖАВНОГО УПРАВЛІННЯ З ПРОГНОЗУ КРИЗОВИХ ЯВИЩ СОЦІАЛЬНО-ЕКОНОМІЧНОГО ХАРАКТЕРУ

У статті доведено актуальність проведення досліджень щодо розроблення сучасних методів аналізу та прогнозу розвитку кризових явищ соціально-економічного характеру в Україні. В якості основи для розроблення державного механізму моніторингу кризових явищ запропоновано використовувати геоінформаційні системи. Здійснено аналіз класичного методу кластеризації за технологією k-середніх та виявлено його недоліки. Розроблено та запропоновано модифікований підхід методу кластеризації за технологією k-середніх для використання у задачах державного управління з прогнозу кризових явищ соціально-економічного характеру.
Ключові слова: *методи кластеризації, кризові явища, державне управління, прогноз, геоінформаційні системи.*

Sergey Belay, Vladimir Lisitsin

APPLYING THE MODIFIED K-MEANS METHOD OF CLUSTER ANALYSIS IN THE TASKS OF SOCIAL-ECONOMICAL CRISIS FORECASTING FOR PUBLIC ADMINISTRATION

The actuality of research for modern methods of analysis and forecasting of social-economical crisis in Ukraine is proved in article. Geographical informational systems are proposed as an instrument for monitoring. Classic method of k-means for cluster analysis is used and its shortcomings are reviewed. Modified method of k-means cluster analysis is proposed for the tasks of social-economical crisis forecasting for public administration.
Keywords: *clustering, crisis, public administration, forecast, geographical informational systems.*

Сергей Белай, Владимир Лисицин

ИСПОЛЬЗОВАНИЕ МОДИФИЦИРОВАННОГО МЕТОДА K-СРЕДНИХ КЛАСТЕРНОГО АНАЛИЗА В ЗАДАЧАХ ГОСУДАРСТВЕННОГО УПРАВЛЕНИЯ ПО ПРОГНОЗУ КРИЗИСНЫХ ЯВЛЕНИЙ СОЦИАЛЬНО-ЭКОНОМИЧЕСКОГО ХАРАКТЕРА

В статье доказана актуальность проведения исследований по разработке современных методов анализа и прогноза развития кризисных явлений социально-экономического характера в Украине. В качестве основы для разработки государственного механизма мониторинга кризисных явлений предложено использовать геоинформационные системы. Осуществлен анализ классического метода кластеризации по технологии k-средних и выявлены его недостатки. Разработан и предложен модифицированный подход метода кластеризации по технологии k-средних для использования в задачах государственного управления по прогнозу кризисных явлений социально-экономического характера.
Ключевые слова: *методы кластеризации, кризисные явления, государственное управление, прогноз, геоинформационные системы.*

Постановка проблеми

Умови сьогодення яскраво демонструють появу численних кризових явищ соціально-економічного характеру. Громадяни в різних країнах світу за ради кращого життя готові йти на крайні радикальні дії, перетинаючи межу закону. За таких умов органам державної влади та силам охорони правопорядку необхідно мати інструментарій для аналізу та прогнозу розвитку кризових явищ в регіонах держави.

З цією метою в розвинених країнах широко застосовуються різноманітні методи збору, статистичного аналізу і комп'ютерної обробки даних про соціально-економічний стан в регіонах держави. Створюються бази даних, що містять ознаки кризових соціально-економічних явищ, кодовані й індексовані з метою прискорення доступу й аналізу.

Аналіз останніх досліджень і публікацій

Деякі з таких баз даних, відкриті й загальнодоступні для використання, як, наприклад, інформація про всі (в тому числі на соціально-економічному підгрунті) протестні дії на території України [1], зібрана за допомогою міжнародного фонду «Відродження». Також розробляються програмні комплекси аналізу та прогнозу кризових явищ. В більшості випадків такі системи є закритими для загального доступу, як приклад відкритої є американська комп'ютерна система «Наутилус», яка здатна на підставі даних засобів масової інформації розробляти прогнози щодо розвитку подій в зазначеному регіоні [2]. Також провідними державами створюються 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 RetrievalWare» [5] та інші.

Основними причинами неможливості використання подібних інформаційно-аналітичних систем є складний та потужний їхній механізм, що потребує значний потужностей для обробки, відсутність вихідних кодів програмних продуктів, що унеможливлює оперативне самостійне корегування продукту, закритість більшості подібних інформаційно-аналітичних систем відповідно до національних інтересів держав-розробників, а також імовірність закладення свідомої похибки в програмний продукт з метою диверсії.

Таким чином, переліченні факти безперечно актуалізують дослідження щодо розроблення сучасних методів аналізу та прогнозу розвитку кризових явищ соціально-економічного характеру в Україні.

Мета

Метою статті є проведення аналізу структури методу кластеризації k-середніх, виявлення і усунення його недоліків та здійснення його модифікації.

Виклад основного матеріалу

Гіпотезою дослідження є припущення, що протестна активність населення в регіонах держави визначається загостренням кризових явищ соціально-економічного характеру. Тому з метою прогнозування кризових явищ у сучасному суспільстві органам державної влади необхідно мати простий, але ефективний державний механізм, який дозволив би проводити аналіз стану суспільної обстановки, виходячи зі статистики інформаційних потоків, які циркулюють у Інтернет-середовищі щодо протестних настроїв населення на соціально-економічному підґрунті та показників рівня життя населення в регіонах держави.

Найбільш придатною платформою до розроблення зазначеного державного механізму є геоінформаційні системи (ГІС), які надають можливість об'єднання математичних методів оцінювання та прогнозування соціально-економічних подій в об'єднанні з сучасними технологіями збору, нанесення і обробки геопросторових даних на електронну карту. Платформою розробки була обрана геоінформаційна система «Інструмент» [6], на базі якої розроблена модель «Аналітика», що поєднала в собі методи збору, оцінювання та прогнозу кризових явищ.

Для розрахунку статистичних показників подій, що поєднаються у групи за певними ознаками та значеннями атрибутів, нам необхідно проводити порівняння об'єктів, нанесених на електронну карту під час збирання даних для моделі «Аналітика». Крім того, аналітика інтересуватиме питання групування (або розкиду) подій навколо деяких центрів у багатовимірному просторі атрибутів. Завдяки присутності геопросторової та часової складової у даних, що аналізуються, питання отримання статистичних характеристик не може бути зведене до простого розрахунку дисперсії та середнього. Потребується більш комплексний підхід. Задача щодо такого порівняння може виникнути у таких випадках, як, наприклад:

- вибір у якості базових усіх регіонів із найбільш загрозливим соціально-економічним станом та найшвидше зростаючою статистикою протестних подій певного типу;
- порівняння соціально-економічног стану в базових (найбільш загрозливих) та інших регіонів, а також створення критеріїв оцінки для 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.

Для вирішення зазначеної задачі за допомогою редактора ГІС «Інструмент» створюємо декілька «зразкових» регіонів, вибірка подій у яких найбільш повно 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». Спрощено механізм порівняння можна сформулювати так: якщо упродовж обраного інтервалу часу соціально-економічний стан у регіоні, що аналізується, близький до стану у «зразковому» регіоні, то говоритиме про ідентичність ситуацій у цих двох регіонах.

Групи подій із схожими значеннями атрибутів називатимемо надалі кластерами. Існує багато алгоритмів формування кластерів у багатовимірному просторі ознак. Для моделі «Аналітика» може бути обрано метод так званих k-середніх (k-means clustering). Розглянемо його сутність.

Процес розпочинається с визначення та створення k порожніх кластерів, на які буде поділена уся множина об'єктів, що підлягають обробці. Також на цьому кроці зі всієї сукупності довільно обираються k об'єктів. Кожен з них призначається одному зі створених кластерів. Таким чином, по завершенні цього етапу існує k кластерів, кожен з котрих містить тільки по одному з об'єктів. Відповідний об'єкт становиться центром такого кластера (рис.1).

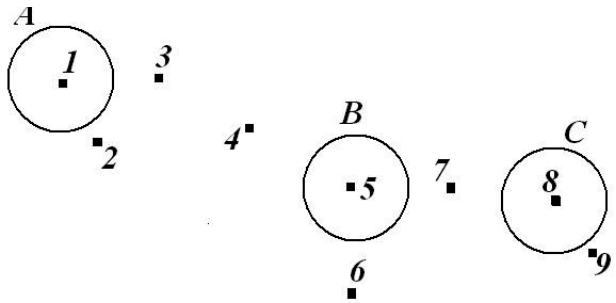

Рис. 1. Перший етап обробки. K = 3 − кількість кластерів, на які буде поділена уся множина об'єктів. Об'єкти за номерами 1, 5 та 8 стають центрами новостворених кластерів.

На наступному кроці для усіх об'єктів, що залишились та не були додані у жодну групу, виконується розрахунок відстаней до центрів новостворених кластерів. Кожний такий об'єкт призначається тому кластеру, відстань до центру якого найменша. Після закінчення другого етапу межі кластерів змінюються, а координати центрів перераховуються з урахуванням доданих об'єктів (рис. 2).

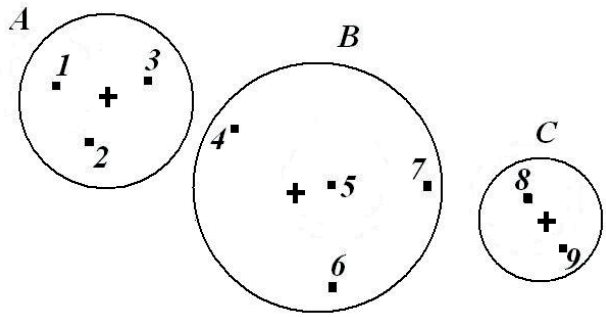

Рис. 2. Другий етап кластеризації. Кожний об'єкт належить одній з трьох груп. Нові центри кластерів позначені хрестиками.

У цьому прикладі мова йде про координати центрів, що визначаються парою значень (X, Y) на площині. Для розрахунку координат центру такого кластера розглядатимемо k об'єктів, що створюють його як матеріальні точки із масами m_i. Сумарна маса M всіх об'єктів визначається як $\sum_{i=0}^{k} m_i$. Тоді шукатимемо координати $\left(\tilde{O}_{\tilde{n}}, Y_c\right)$ центру кластера у вигляді центру мас точкових матеріальних об'єктів на площині.

$$\left.\begin{aligned} X_c &= \frac{\sum x_i m_i}{M} \\ Y_c &= \frac{\sum y_i m_i}{M} \end{aligned}\right\} \qquad (1)$$

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

На практиці виникає низка питань, пов'язана з ефективністю застосування такого методу та інтерпретацією отриманих результатів. Дійсно, у вихідних параметрах моделі визначається тільки загальна кількість кластерів, до яких будуть віднесені події, що аналізуються. На цьому етапі вплинути на склад результуючих кластерів можливо тільки за допомогою фільтрації атрибутів, які застосовуватимуться в алгоритмі. Йдеться про атрибути, що визначають реляційну таблицю бази даних із протестними подіями.

Але що ми отримуємо у результаті? 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 можна спрощено звести 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 потрібує потужних комп'ютерних ресурсів у тому випадку, коли кількість об'єктів, що обробляються дуже значна. Це пов'язано із багатьма циклами перерахування центрів кластерів і зміни їх складу.

Такі обмеження класичного методу k- середніх призвели до необхідності розроблення модифікованого алгоритму, із більш наочними результатами групування протестних подій соціально-економічного походження на карті. Перш ніж розглянути його, введемо таке поняття як шаблонна ситуація, або просто шаблон.

Припустимо, що соціально-економічний стан у регіоні характеризується кількома вербальними категоріями, тому з метою градації пропонується ввести чотири рівня загроз (звичайний, ускладнений, кризовий, надзвичайний). За принципом достатності вважається, що даної кількості рівнів буде достатньо для класифікації стану соціально-економічної безпеки. Використання більшої кількості рівнів є недоцільним, тому що це буде ускладнювати процес опису поточної оперативної ситуації, а також застосування самого механізму особою, яка буде приймати рішення.

Для опису кожного такого стану з бази даних протестних подій експерт обирає або штучно створює об'єкт-подію, яка, на його думку, за змістом атрибутів найбільше відповідає поточної категорії. За допомогою атрибутів вибраного об'єкту створюється шаблон. До речі, для більш повного опису стану експертом може бути створено водночас декілька таких шаблонів із різними множинами атрибутів. Тоді для аналізу соціально-економічного стану в обраному регіоні необхідно підрахувати за допомогою інструментів геопросторових запитів кількість подій, віднесених до певного шаблону, за послідовні інтервали часу. Зростання або збереження на високому рівні кількості таких подій упродовж часу говорить про тенденцію до формування в регіоні певного стану.

Для опису певної категорії соціально-економічного стану може бути створена будь-яка кількість шаблонів. Після цього поточна подія, що міститься у базі даних, послідовно порівнюється із кожним створеним шаблоном що створено. Порівняння здійснюється за критерієм мінімальної відстані між відповідними атрибутами події та шаблону. Оскільки для кожного шаблону вже визначена певна категорія соціально-економічного стану, то за результатами всього одного порівняння подія буде відразу віднесена до необхідної категорії соціально-економічного стану в регіоні. Тобто, не потрібно у багатьох ітераціях виконувати перерахунок координат центру кожного кластеру − для кожної події здійснюється тільки один цикл.

У моделі «Аналітика» користувач має можливість вибрати алгоритм виконання кластеризації об'єктів карти − класичний метод k- середніх, або модифікований метод, що заснований на застосуванні шаблонів.

Висновки

Таким чином, на основі викладеного вище можливо зробити такі висновки.

1. В якості платформи для розроблення державного механізму моніторингу кризових явищ була обрана геоінформаційна система «Інструмент» та розроблена відповідна модель «Аналітика», що поєднала в собі методи збору, оцінювання та прогнозу кризових явищ соціально-економічного характеру. Використання геоінформаційної системи «Інструмент» надає можливість технічного поєднання зазначених методів та додаткового візуального аналізу ситуації на електронній карті, що значно спрощую роботу аналітичних підрозділів органів державного управління.

2. Аналіз класичного методу кластеризації k-середніх виявив значні проблемні питання його використання

в моделі «Аналітика». Розроблений модіфікований метод кластерного аналізу к-середніх дозволяє віднести подію до необхідної категорії соціально-економічного стану в регіоні та здійснити первинний прогноз ситуації.

4. Напрямки подальших наукових розвідок будуть спрямовані на перевірку коректності моделі «Аналітика» зі статистикою протестних подій соціально-економічного походження в регіонах Україні та вдосконалення державного механізму моніторингу кризових явищ соціально-економічного походження.

Література.

1. Протесты, победы и репресии в Украине: результати моніторингу, октябрь 2009 − сентябрь 2010. − К. : Центр исследования общества, 2011. − 64 с.

2. Суперкомпьютер способен предсказывать будущее / Сайт агентства новостей «Сегодня» [Электронный ресурс]. − http://www.segodnya.ua/news/14287696.html (дата обращения: 12.12.13).

3. Розробка форм і способів інформаційної боротьби при виконанні внутрішніми військами Міністерства Внутрішніх Справ України службово-бойових завдань / Отчет про научно-исследовательскую роботу. − Харьков, Академия внутренних войск МВС Украины, 2009. − 312 с.

4. Ландэ Д. Инструментарий аналитика / Корпоративные решения // Мониторинг информации, ТЕЛЕКОМ. − 4, 2010. − С. 36−41. [Электронный ресурс]. − http://poiskbook.kiev.ua/art/telecom0410/telecom0410.pdf (дата обращения: 12.12.13).

5. Новейшие сетевые технологии. [Электронный ресурс]. − http://www.ant.kiev.ua/convera/convera%20RV.html (дата обращения: 12.12.13).

6. Дробаха Г. А., Створення просторових даних для електронних карт геоінформаційної системи внутрішніх військ МВС України / Г. А. Дробаха, Л. В. Розанова, В. Е. Лісіцин. − Х. : Академія внутрішніх військ, 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.