density values of clutters with neighbor ranks. If this data is supplied, the implemented software may be used for calculation of practically-significant assessments of population density spatial distribution.

References

1. GeoImage – Population maps – GEOpop [Electronic resource]: GeoImage Corporation official site – Mode of access: http://www.geoimage.fr/images/ stories/products/ geopop.pdf – Title from the screen.
2. Гайдышев И. Анализ и обработка данных. Специальный справочник/ Игорь Гайдышев. – СПб.: Питер, 2001. – 752 с. – ISBN 5-318-00220-X.
3. Бояринов А.И. Методы оптимизации в химической технологии/ А.И. Бояринов, В.В. Кафаров – М.: „Химия“, 1975. – 576 с.
4. Візіком – геоінформаційні системи, радіопланування, GPS навігатори, LBS проекти [Електронний ресурс]: офіційний сайт ЗАТ „Візіком“ – Режим доступу: http://www.visicom.ua – Назва з екрану.
5. 3D модели для планирования и оптимизации беспроводных сетей связи. – К.: ВИЗИКОМ, 2010. – 15 с.

Розглянуто алгоритм отримання кількісних оцінок просторового розподілу щільності населення на основі рангових даних. Алгоритм грунтується на методі Гаусса-Зайделя з „гнучкими“ обмеженнями. Сформульовано вимоги, які пред’являються до вихідних даних, достатні для отримання однозначного рішення. Представлено результати розрахунків у середовищі MATLAB. Показано, що результат не залежить від початкових щільностей, але визначається мінімальною кількісною різницею між щільностями клатерів із сусідніми рангами.

Ключові слова: *клаттер, щільність населення, ранг, оптимізація, обмеження*

Рассмотрен алгоритм получения количественных оценок пространственного распределения плотности населения на основе ранговых данных. Алгоритм основывается на методе Гаусса-Зайделя с „гибкими“ ограничениями. Сформулированы требования, предъявляемые к исходным данным, достаточные для получения однозначного решения. Представлены результаты расчётов в среде MATLAB. Показано, что результат не зависит от начальных плотностей, но определяется минимальным количественным различием между плотностями клаттеров с соседними рангами.

Ключевые слова: *клаттер, плотность населения, ранг, оптимизация, ограничения*

Рекомендовано до публікації докт. техн. наук Б.С. Бусигіним. Дата надходження рукопису 28.02.11

УДК 004.8

М.В. Назаренко,
Л.В. Саричева, канд. фіз.-мат. наук, доц.

Державний вищий навчальний заклад „Національний гірничий університет“, м. Дніпропетровськ, Україна, e-mail: sarycheval@nmu.org.ua

АЛГОРИТМ КЛАСТЕРИЗАЦІЇ НА ОСНОВІ НЕЧІТКИХ МНОЖИН

M.V. Nazarenko,
L.V. Sarycheva, Cand. Sc. (Phys.-Math.) Assoc. Prof.

State Higher Educational Institution “National Mining University”, Dnipropetrovsk, Ukraine, e-mail: sarycheval@nmu.org.ua

CLUSTERING ALGORITHM BASED ON FUZZY SETS

Запропоновано математичну модель і метод кластеризації, що враховує інтуїтивне уявлення про групування даних, не накладаючи апріорних припущень щодо структури даних. Метод кластеризації FuzzyCluster розроблено на основі нечіткого опису, здатного функціонувати в умовах апріорної невизначеності щодо структури даних, а також такого, що враховує інтуїтивне уявлення про групування даних. Порівняння результатів кластеризації алгоритмом FuzzyCluster з алгоритмами k-means та c-means на п’яти стандартних наборах даних показує його переваги.

Ключові слова: *кластеризація, нечітка множина, міра близькості, функція належності, нечітке відношення*

Вступ. Кластеризація даних – процес групування елементів даних на класи так, що елементи в одному класі є якомога близькими, а елементи різних класів є настільки різнорідними, наскільки це можливо.

У жорсткій кластеризації дані розділені на окремі кластери, де кожен елемент даних належить одному з кластерів. У нечіткій кластеризації елементи даних можуть належати до більш ніж однієї групи і з кожним елементом множини пов’язана функція належності до кожного кластеру. Вона вказує на силу зв’язку між цим елементом даних і конкретною групою. Нечітка кластеризація є процесом присвоєння цих мір належності та їх використання для визначення складу кожного з кластерів.

Основні відомі алгоритми кластеризації (наприклад, модифікації алгоритмів K-Means, Expectation Maximization, реалізовані, у тому числі, у Microsoft Analysis Services 2005) накладають обмеження на гео-

метрію отримуваних кластерів, зокрема, вимагаючи можливості охоплення кожним кластером окремого опуклого простору. Таке обмеження накладається припущеннями таких алгоритмів про існування центрів кластерів (K-Means) або функції щільності ймовірності для кожного кластера з відповідним значенням математичного очікування і дисперсією (Expectation Maximization). У результаті, ці алгоритми не в змозі адекватно розбити на кластери невипуклі простори, ще складніше розбиття вкладених структур.

Але необхідність такого розбиття виникає, наприклад, для адекватної та інформативної кластеризації даних, які мають вкладену форму. Також більшість алгоритмів погано працює у випадку, коли один кластер значно більше за інших, і вони знаходяться близько один від одного.

Цю проблему вирішує описуваний у даній статті алгоритм кластеризації на основі нечітких відносин, що дозволяє групувати в кластери елементи, між якими є послідовність „близьких“ один до одного елементів, що також відповідає інтуїтивному уявленню про угруповання.

Даний алгоритм не потребує великої кількості вхідних параметрів. Необхідний параметр кластеризації в даному алгоритмі – *α*-рівень, $\alpha \in [0;1]$. Така властивість надає алгоритму, на основі нечітких відносин, більше переваг, порівняно з іншими алгоритмами на основі нечіткої логіки (Fuzzy C-Means та Алгоритм Густафсона-Кессель).

Постановка задачі. Мета даної статті – розробка та тестування алгоритму кластеризації на основі нечітких множин. За основу взято алгоритм FuzzyRelationsClustering [1].

Алгоритм FuzzyCluster. Розроблений алгоритм кластеризації на основі нечітких множин складається з наступних етапів:

1. Налаштування параметрів алгоритму:
   a) вибір необхідної метрики;
   b) оцінка помилки кластеризації;
   c) вибір α-рівня, при якому досягається мінімальна кількість неправильно розпізнаних об’єктів.
2. Кластеризація даних із заданими параметрами.

Основний етап – кластеризація даних здійснюється в результаті наступних кроків:

1. Задаються вхідні дані: матриця „об’єкти-ознаки” та значення $\alpha \in [0;1]$ – параметр кластеризації.

2. За допомогою заданої метрики обчислюється функція міри подібності *j*-ого объекту з *i*-им.

3. Обчислюються функції схожості *k*-го та *l* –го об’єкту відносно *i*-го, двох об’єктів відносно всіх об’єктів кластеризації.

4. Рекурсивно обчислюється значення функції $\mu^{(n)}(i,j)$.

5. На основі $\mu^{(n)}(i,j)$ визначається відношення еквівалентності R_α, яке розбиває множину X на кластери.

Вхідні дані для алгоритму подаються після попереднього застосування мінімаксного нормування.

Для підрахунку помилки кластеризації було розроблено алгоритм, що порівнює об’єкти за критерієм належності до певного кластеру.

Математична модель алгоритму. Нехай X – метричний простір і $d: X \to R$ певна на ньому метрика, $(X_1,...,X_n) \subset X$ – послідовність елементів з X.

Ми припускаємо далі, що

$$\forall i \in \{1,..,n\},\ \exists j \in \{1,..,n\} : X_i \neq X_j. \quad (1)$$

З умови (1) виходить, що $\forall i \in \{1,...,n\}$ справедливо

$$\max\{d(X_i, X_k) \mid k \in \{1,..n\}\} > 0. \quad (2)$$

Таким чином, для кожного індексу і ми можемо визначити функцію, що описує міру подібності *j*-ого елемента послідовності з *i*-им елементом

$$\xi_i : \{1,...,n\} \to [0,1]$$

$$\xi_i(j) := 1 - \frac{d(X_i, X_j)}{\max\{d(X_i, X_k) \mid k \in \{1,...,n\}\}}, \quad (3)$$

Значення результатів обчислення даної функції будуть у межах [0 ; 1]. Тому ця матриця – нечітка множина, що характеризує міру подібності *j*-ого елемента послідовності з *i*-им елементом.

Для кожного індексу *i* визначимо функцію, що описує міру подібності *k*-ого й *l*-ого елемента відносно *i*-ого елемента

$$\varsigma_i : \{1,...,n\}^2 \to [0,1]$$

$$\varsigma_i(k,l) := 1 - |\xi_i(X_k) - \xi_i(X_l)|. \quad (4)$$

Ця матриця має трьохвимірний вигляд, значення її елементів знаходяться на інтервалі [0 ; 1]. Вона містить дані про подібність кожної з пар об’єктів відносно інших об’єктів кластеризації.

Визначимо функцію, що описує міру подібності будь-яких двох елементів послідовності щодо всіх елементів послідовності

$$\mu : \{1,...,n\}^2 \to [0,..1]\ ; \quad (5)$$

$$\mu(i,j) := \min\{\varsigma_k(i,j) \mid k \in \{1,...,n\}\}.$$

Дана функція є результатом перетину нечітких множин.

Значення отриманої в результаті матриці входять у діапазон [0; 1], причому значення на головній діагоналі дорівнюють одиниці

$$\mu(i,i) = 1,\ \forall i \in \{1,...,n\}.$$

Так як $\varsigma_k(i,i) = 1 - |\xi_k(X_i) - \xi_k(X_i)| = 1$

для всіх *k*, *i*, то результат функції, що описує міру подібності будь-яких двох елементів послідовності щодо

всіх елементів послідовності, дорівнює $\mu(i,i) := \min\{\varsigma_k(i,i) \mid k \in \{1,...,n\}\} = 1$.

Для k=1, 2,…,n визначимо за допомогою рекурсії функції

$$\mu^{(k)} : \{1,...,n\}^2 \to [0,1]$$

$$\begin{cases} \mu^{(1)}(i,j) := \mu(i,j) \\ \mu^{(k)}(i,j) := \max\left\{ \min\left\{ \begin{matrix} \mu^{(k-1)}(i,s), \\ \mu^{(k-1)}(s,j) \end{matrix} \right\} \mid s \in \{1,...,n\} \right\} \end{cases} \quad (6)$$

У даній формулі використовується t-норма (t-норми, або тріангуляторні норми, реалізують логічні операції "І", "АБО", "НЕ").

Для знаходження функції $\mu^{(k)}(i,j)$ використовується операція „АБО (I)”, що реалізовується знаходженням максимуму з перетину (мінімуму) нечітких множин.

У даній формулі використовується поняття t-норма. t-норми, або тріангуляторні норми реалізують логічні операції „І“, „АБО“, „НЕ“, а також операції взяття мінімуму, максимуму над нечіткими множинами. Для знаходження функції $\mu^{(k)}(i,j)$ використовується операція „АБО (I)”, що реалізовується знаходженням максимуму з перетину (мінімуму) нечітких множин.

Як і у функції (5), значення головної діагоналі дорівнюють одиниці $\mu^{(k)}(i,i) = 1, \forall i,k$, і так само матриця симетрична

$$\mu^{(k)}(i,j) = \mu^{(k)}(j,i), \forall i,j,k .$$

Для $\alpha \in [0,1]$ визначимо на множині $\{X_1,...,X_n\}$ бінарне відношення $R_\alpha \subset \{X_1,...,X_n\}^2$ у такий спосіб

$$(X_i, X_j) \in R_\alpha :\Leftrightarrow \mu^{(n)}(i,j) \geq \alpha \quad (7)$$

Тобто виконаємо зріз нечіткої множини. У результаті отримаємо множину альфа рівня, що буде містити елементи, приналежність яких вище або дорівнює заданому порогу альфа.

Таким чином, відношення еквівалентності R_α розбиває множину $\{X_1,...,X_n\}$ на непересічні класи еквівалентності. Два елементи X_i, X_j входять в один клас еквівалентності тоді й тільки тоді, коли значення функції $\mu^{(n)}$ від цих елементів достатньо велике, що еквівалентно існуванню послідовності пар елементів $(X_i, X_{j_1}), (X_{j_1}, X_{j_2}), ..., (X_{j_r}, X_j)$, для яких значення функції μ велике. По визначенню μ означає близькість елементів кожної пари один до одного. Тобто, два елементи входять в один клас еквівалентності тоді й тільки тоді, коли між ними є послідовність попарно близьких один до одного елементів.

Тестування алгоритму FuzzyCluster. Для тестування розробленого алгоритму застосовувалась функція підрахунку помилки кластеризації (у процентному відношенні число неправильно розпізнаних об’єктів до загальної кількості об’єктів). Порівняння результатів кластеризації алгоритмом FuzzyCluster, алгоритмом k-means та c-means наведено в таблиці.

У результаті тестування виявлено залежність алгоритму від кількості ознак об’єктів. Алгоритм чудово працює з просторовими (трьохвимірними) даними, кластеризуючи їх безпомилково (0% неправильно розпізнаних об’єктів). Це надає змогу застосовувати даний алгоритм для аналізу зображень.

Таблиця

Помилки кластеризації алгоритмами FuzzyCluster і k-means

Дані	**Розмір даних (число об’єктів × атрибути)**	**Кількість кластерів**	**Помилка кластеризації, %**		
			k-means (пакет STATISTICA)	**Fuzzy C-Means (пакет Matlab)**	**FuzzyCluster**
Елементи таблиці Менделєєва	48×4	4	33	27	25
Дані про три сорти вина	178×13	3	3,4	31,5	11,2
Графічні дані (трьохвимірні) – літери „С Р”	206×3	2	41,7	0	0
Графічні дані (трьохвимірні) – вкладені кластери	600×3	2	50	50	0
Іриси Фішера	150×4	3	11	10,6	15

Висновки. Розроблений алгоритм кластеризації має такі переваги:

1. Відсутність необхідності в апріорних припущеннях щодо структури даних (вид і параметри розподілу ймовірності по кластерах, центрів щільності, числа кластерів).

2. Зрозуміла інтерпретація результатів розбиття по кластерах: елементи входять в один кластер, коли між ними є послідовність близьких один до одного елементів.

3. Відсутність обмежень на геометрію кластерів.

4. Час виконання алгоритму мало залежить від числа ознак вхідних об’єктів.

Істотним недоліком алгоритму є великий час виконання, що характеризується n^4 порядком від числа елементів. Однак, цю проблему можна вирішити оптимізацією роботи з матрицями та вилученням неінформативних ознак.

Реалізований алгоритм кластеризції на основі нечітких відносин має право на існування, завдяки своїй унікальній властивості – використанню нечітких множин та відношень, що надає змогу кластеризувати дані незалежно від їх структури та форми.

Список літератури

1. Кластеризация на основе нечетких отношений. Алгоритм FuzzyRelationClustering. [Електронний ресурс]. – Режим доступу: http://www.spellabs.ru/download/FuzzyRelationClustering.doc. – Назва з екрану.
2. Б. Дюран, Кластерный анализ. / Б. Дюран, П. Оделл. Пер. с англ. Е.З. Демиденко. Под ред. А.Я. Боярского. – М.: „Статистика“, 1977. – 128 с.
3. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. / Загоруйко Н.Г. –Новосибирск.: ИМ СО РАН, 1999, – 270 с.
4. Интеллектуальный анализ данных с помощью кластеризации. [Електронний ресурс]. – Режим доступу: http://www.kg.ru/?page_id=183. – Назва з екрану.
5. Алгоритмы кластерного анализа. [Електронний ресурс]. – Режим доступу : http://www.dea-analysis.ru/clustering-5.htm. – Назва з екрану.
6. Алгоритмы кластеризации на службе Data Mining. [Електронний ресурс]. – Режим доступу : http://www.basegroup.ru/library/analysis/clusterization/datamining/. – Назва з екрану.

Предложена математическая модель и метод кластеризации, учитывающий интуитивное представление о группировке данных, не накладывая априорных предположений о структуре данных. Метод кластеризации FuzzyCluster разработан на основе нечеткого описания, способного функционировать в условиях априорной неопределенности относительно структуры данных, а также такого, который учитывает интуитивное представление о группировке данных. Сравнение результатов кластеризации алгоритмом FuzzyCluster с алгоритмами k-means и c-means на пяти стандартных наборах данных показывает его достоинства.

Ключевые слова: *кластеризация, нечеткое множество, мера близости, функция принадлежности, нечеткое отношение*

Authors propose a mathematical model and a method of clustering which takes into account the intuitive idea of grouping the data without imposing a priori assumptions about data structures. Clustering method FuzzyCluster was developed on the base of fuzzy description, capable of functioning under conditions of uncertainty about data structures, as well as that which takes into account the intuitive idea of data grouping. When comparing results of the five standard datasets clustering made by the algorithm FuzzyCluster with those made by the algorithms k-means and c-means we can see the advantages of the FuzzyCluster.

Keywords: *clustering, fuzzy set, similarity measure, accessory function, fuzzy relation*

Рекомендовано до публікації докт. техн. наук Б.С. Бусигіним. Дата надходження рукопису 28.02.11

УДК 528.854

Б.С. Бусыгин, д-р техн. наук, проф.,
Е.Л. Сергеева

Государственное высшее учебное заведение „Национальный горный университет“, г. Днепропетровск, Украина,
e-mail: sergieieva@i.ua

МОНИТОРИНГ СОСТОЯНИЯ ТЕРРИКОНОВ ДОНБАССА ПО ДАННЫМ МУЛЬТИСПЕКТРАЛЬНЫХ КОСМИЧЕСКИХ СЪЕМОК

B.S. Busygin, Dr. Sc. (Tech.), Professor,
Ye.L. Sergeyeva

State Higher Educational Institution “National Mining University”, Dnipropetrovsk, Ukraine, e-mail: sergieieva@i.ua

MONITORING OF A WASTE BANKS STATE AT THE TERRITORY OF DONETS COAL BASIN USING MULTISPECTRAL SATELLITE IMAGERY

Представлены практические аспекты применения данных мультиспектральных космических съемок к решению задач мониторинга терриконов горнопромышленных регионов, в частности, составлению карты расположения негорящих, тлеющих и горящих терриконов участка Донецкого угольного бассейна. Разработан подход к мониторингу состояния терриконов по набору разновременных космоснимков. Выполнена проверка предложенного подхода на данных космических съемок Landsat-TM участка Донбасса.

Ключевые слова: *мониторинг, космическая съемка, террикон, приповерхностная температура, классификация с обучением, Landsat-TM*

Введение. Донецкий каменноугольный бассейн, открытый в 1720 г., в течение почти 300 лет является крупнейшим индустриальным и промышленным центром Украины. Общая площадь бассейна составляет около 60 тыс. км^2 и охватывает территории Днепропетровской, Донецкой и Луганской областей [1]. В Донбассе разведано свыше 800 месторождений более 50 видов минерального сырья, общая стоимость которых – свыше 3 триллионов долларов США. Из них запасы угля до глубины 1800 м составляют 140,8 миллиардов