УДК 330.4

К. Л. Юрченко,
аспірант кафедри економічної кібернетики,
Київський національний університет імені Тараса Шевченка

ВИКОРИСТАННЯ АНСАМБЛЮ АЛГОРИТМІВ ДЛЯ КЛАСТЕРИЗАЦІЇ КАТЕГОРІАЛЬНИХ ДАНИХ

K. Yurchenko,
post-graduate student, Department of Economic Cybernetics, Taras Shevchenko Kyiv National University

USING OF ENSEMBLE OF ALGORITHMS FOR CLUSTERIZATION OF CATEGORICAL DATA

У статті розглядається процес кластеризації категоріальних даних за допомогою ансамблю алгоритмів кластеризації. Запропоновано модифікацію двоетапного алгоритму кластеризації. Ефективність запропонованого алгоритму протестовано на прикладі даних фізичних осіб.

The paper studies process of categorical data clusterization using the ensemble of algorithms of clusterization. The modification of two-stage clustering algorithm is proposed. The research proves the effectiveness of given algorithm using individuals' data.

Ключові слова: Ансамбль кластерів, мультиноміальна логістична регресія, двоетапний алгоритм кластеризації, неперервні змінні, категоріальні змінні.
Key words: ensemble of clusters, multinomial logistic regression, two-stage clustering algorithm, continuous variables, categorical variables.

ПОСТАНОВКА ПРОБЛЕМИ

Кластерний аналіз — задача розбиття заданої вибірки спостережень на підмножини, що називаються кластерами, так, що кожний кластер складається зі схожих об'єктів, а об'єкти різних кластерів суттєво відрізняються. Задача кластеризації відноситься до статистичної обробки, а також до широкого класу задач навчання без учителя.

Навчання без учителя — один зі способів машинного навчання, при розв'язку яких досліджувана система спонтанно навчається виконувати поставлену задачу, без втручання з боку експериментатора. Навчання без учителя часто протиставляється навчанню з учителем, коли для кожного об'єкту наперед задається правильна відповідь і вимагається знайти залежність між стимулами та реакціями системи.

Використана термінологія:

Об'єкт спостереження — елементарна множина даних, з якою оперує алгоритм кластеризації.

Кожному об'єкту відповідає вектор характеристик, атрибутів:

$x = (x_1, \ldots, x_k)$

де k — кількість окремих характеристик об'єкту, або розмірність простору характеристик.

Множина всіх об'єктів $(X_1, \ldots X_n)$, де $X_i = (x_{i1}, \ldots, x_{id})$.

Кластер — множина близьких одне до одного об'єктів.

$d(x_i; x_j)$ — "відстань" між двома об'єктами, результат застосування метрики, або квазі-метрики у просторі характеристик.

Переваги та недоліки існуючих алгоритмів:

K-середнє — найбільш популярний метод кластеризації, був винайдений в 1950-х роках математиком Гуго Штейнгаузом [5] та майже одночасно Стюартом Ллойдом [3]. Особливу популярність набрав після робіт Маккуіна [4]. Є інтуїтивно зрозумілим, більш швидким у порівнянні з ієрархічними алгоритмами. Варто відзначити, що метод є більш орієнтованим на неперервні дані та евклідову відстань. Більшість даних, що описують фінансовий стан підприємств добре описуються за допомогою n — мірного нормального розподілу. В принципі, випадковий вибір центрів кластерів, що використовується за замовчуванням при використанні методу K-середнє, прийнятний у даному випадку. Варто очікувати, що центри кластерів, незалежно від обраної їх кількості, будуть розташовуватись еліптично. Кількість спостережень буде вищою у центральних кластерах. Вищенаведені висновки були зроблені, виходячи з попереднього уявлення про структуру даних. Зазвичай кластерний аналіз і проводиться для складання такого

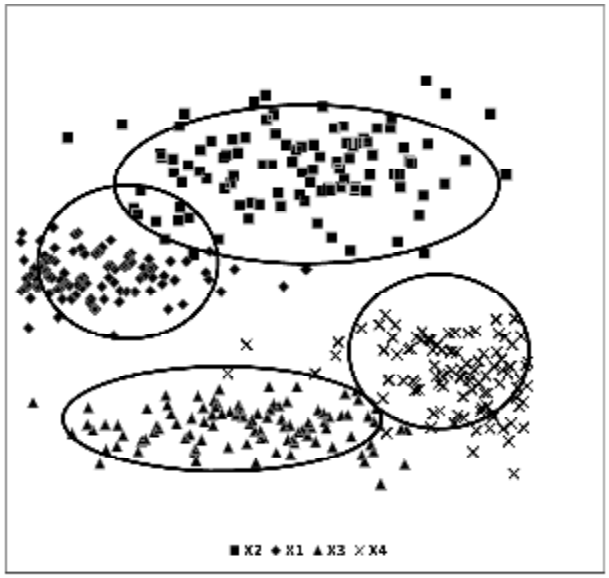

Рис. 1. Результат роботи K-середнє за умови, що справжня кількість кластерів — чотири, а обрана кількість для кластеризації — також чотири

попереднього, досить грубого уявлення про структуру даних. За умови, що дослідник має початкову уяву про структуру даних, то проведення кластеризації не надасть йому суттєвої інформації. В той же час, початкове уявлення про структуру даних може надати інформацію про кількість кластерів та їх стартові, орієнтовні центри для використання алгоритму K-середнє. В той же час, варто очікувати суттєвої різниці в результатах кластеризації для підприємств різноманітних галузей, регіонів. Отже, за умови наявності таких параметрів результати алгоритмів кластеризації будуть більш інформативними.

Використання алгоритму K-середнє для категоріальних даних є більш проблематичним.

Недоліки алгоритму [4]:

— Не гарантується досягнення глобального мінімуму сумарного квадратичного відхилення, а тільки одного з локальних мінімумів.

— Результат залежить від вибору початкових центрів кластерів, їх оптимальний вибір невідомий.

— Кількість кластерів треба знати заздалегідь. Початковий вибір кількості кластерів суттєво впливає на результат.

Наприклад, за умови, що справжня кількість кластерів — чотири (див. рис. 1) було обрано кількість кластерів рівну двом для виконання алгоритму. В результаті наступні спостереження будуть об'єднані у наступні кластери (див. рис. 2).

Майже всі ієрархічні алгоритми кластеризації позбавлені недоліків K-середнє, проте вимагають набагато більшого часу для виконання. Більшість статистичних пакетів пропонують можливість ієрархічної кластеризації тільки для незначних об'ємів даних — до 1000 спостережень.

Зазвичай можливо використати попередню агрегацію, що призведе до суттєвого зменшення кількості аналізованих даних без втрати точності. Наприклад, при відмінностях між значенням фінансових показників менше ніж на 1%, 0,1% ми вважаємо їх однаковими. Межа для кожного показника може бути встановлена безпосередньо дослідником, виходячи з практичних міркувань. Такий спосіб агрегації дозволяє швидко об'єднати об'єкти, характеристики яких близькі до середніх значень. У випадку категоріальних змінних можна об'єднувати без будь-якого попереднього аналізу об'єкти, що відрізняються тільки однією характеристикою [5].

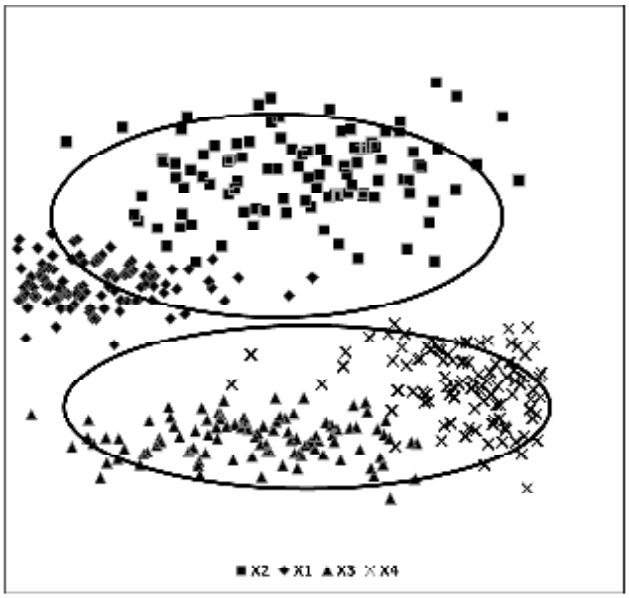

Рис. 2. Результат роботи K-середнє за умови, що справжня кількість кластерів — чотири, а обрана кількість для кластеризації — два

		X2	
		0	1
X1	0	25%	25%
	1	25%	25%

Рис. 3. Рівномірний розподіл спостережень між кластерами

		X2	
		0	1
X1	0	49%	1%
	1	1%	49%

Рис. 4. Розподіл спостережень з сильнокорельованими параметрами

		$\hat{X}_2$	
		0	1
$\hat{X}_1$	0	50%	0%
	1	0%	50%

Рис. 5. Розподіл прогнозів параметрів спостережень, отриманих на основі мультиноміальних регресій

На основі низки подібних міркувань побудований двоетапний алгоритм кластеризації, що реалізований у пакеті SPSS. Важливою перевагою даного алгоритму є можливість одночасного опрацювання як категоріальних так і неперервних даних.

Двоетапний алгоритм кластеризації:

Двоетапний алгоритм кластеризації — метод, розроблений для обробки дуже великих вибірок даних. Він здатний оброблювати неперервні та категоріальні змінні. Опис взятий з документації пакету SPSS. Алгоритм був розроблений на основі низки доповідей на конференціях [6, 7]. Він складається з наступних кроків:

— Попередня кластеризація.

— Обробки викидів (опціональний пункт. Може ігноруватись дослідником у разі впевненості в чистоті даних).

— Безпосередньо кластеризація.

Попередня кластеризація:

На етапі попередньої кластеризації використовується послідовний підхід до групування об'єктів. Спостереження оброблюються одне за одним та додаються до існуючих кластерів чи утворюють нові в залежності від відстані між кожним об'єктом та попередньоутвореними кластерами (відстань між об'єктами не завжди є Евклідовою відстанню — звичним уявленням про відстань, а зазвичай залежить від специфіки досліджуваних даних).

В основі процедури лежить побудова модифікованого дерева характеристик. Таке дерево складається з рівнів вузлів, а кожен вузол складається з окремих спостережень.

Вузли останнього рівня представляють кластери — результат процедури попередньої кластеризації.

Вузли неостанніх рівнів використовуються для швидкого віднесення кожного спостереження до певного кластеру.

Кожен вузол характеризується характеристиками, на основі яких він сформований, кількістю спостережень, що його утворюють, середнім значенням та середньоквадратичним відхиленням кожної характеристики, якими описуються спостереження.

Для кожного наступного спостереження, починаючи з кореневого вузла починається пошук найближчого за кожною характеристикою. Відстань має бути не більша, ніж порогове значення. У разі знаходження найближчого вузла на останньому рівні його значення (кількість спостережень у кожному вузлі, середнє значення та середньоквадратичне відхилення за кожною характеристикою) перераховуються.

У випадку незнаходження найближчого вузла проводиться перебудова дерева з виділенням нових вузлів.

Якщо у результаті кількість кінцевих кластерів отримується завелика для подальшої обробки ієрархічним алгоритмом, то значення порогового значення має бути збільшеним. Взагалі кажучи, найбільш оптимальною є така максимальна кількість кластерів отримана після першого етапу, що може бути оброблена ієрархічним алгоритмом кластеризації.

Безпосередньо кластеризація:

Після отримання кластерів за результатами першого етапу, вони оброблюються ієрархічним алгоритмом кластеризації. Кількість під кластерів, кластерів отриманих на першому етапі, суттєво менша ніж кількість спостережень на початку, а отже ієрархічна кластеризація може бути виконана за прийнятний час.

Ієрархічна кластеризація — низка алгоритмів з наступною стартовою спільною ідеєю кластеризації:

Спочатку кожен об'єкт спостереження відноситься до окремого кластеру.

Процедура кластеризації продовжується до тих пір, доки всі спостереження не будуть об'єднані в один, або будь-яке інше, наперед задане число кластерів.

Алгоритми ієрархічної кластеризації відрізняються способом, за яким на кожному етапі знаходиться найближчий кластер для певного спостереження — означенням відстані між кластерами. Три найбільш популярні методи визначення відстані наведені нижче [8]:

$d(C_i, C_j) = \min_{x \in C_i, C_j} d(x, x')$.

У даному випадку кластери мають тенденцію бути довгими та тонкими. У разі утворення одного кластеру він буде еквівалентний мінімальному остовому дереву.

$d(C_i, C_j) = \max_{x \in C_i, C_j} d(x, x')$.

У даному випадку кластери мають тенденцію бути компактними, приблизно рівними за діаметром.

$$d(C_i, C_j) = \frac{\sum x \in C_i, x' \in C_j d(x, x')}{|C_i||C_j|}.$$

В даному випадку отримуються кластери, що є проміжними варіантами між першим та другим варіантами.

Запропонована модифікація:

1. Усі дані мають бути агреговані до категоріальних даних. Неперервні величини такі, як час, відстань, вага, вартість мають бути приведені до певних груп.

2. Для кожної характеристики проводиться оцінка мультиноміальної регресії.

Наприклад:

Початкова вибірка:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{nk} \end{pmatrix} = (t_1, \dots, t_k)$$

t_i — вектор i-их характеристик кожного спостереження.

Отже, $\forall\, i = \overline{1,k}$ *оцінюється мультиноміальна регресія виду* $f(t_1, \dots \dots t_{i-1}, t_{i+1}, \dots t_k) = t_i + \varepsilon$. Тут опущено індекс спостережень.

Припускаємо, що $\forall\, i, t_i$ *приймає одне з* m_i *значень.* $m_i \in N$..

Це є результатом пункту 1 та необхідною умовою для використання мультиноміальної регресії.

Отримується прогноз кожної характеристик на основі інших характеристик, що описують кожне спостереження.

$\hat{t_i} = f(t_1, \dots \dots t_{i-1}, t_{i+1}, \dots t_k)$.

Model Summary	
Algorithm	TwoStep
Inputs	7
Clusters	3
Measure of cohesion and separation	0.4
Size of Smallest cluster	27779 (26.5%)
Size of Largest cluster	39547 (37.8%)
Ratio of Sizes	1.42

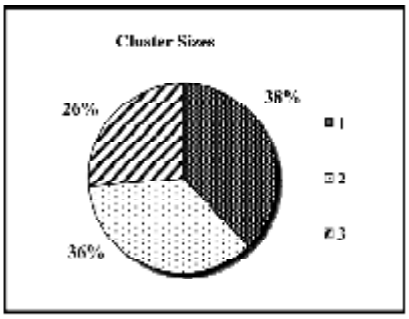

Рис. 6. Результат роботи ансамблю алгоритмів без накладання обмежень

Model Summary	
Algorithm	TwoStep
Inputs	7
Clusters	2
Measure of cohesion and separation	0.25
Size of Smallest cluster	42075 (40.2%)
Size of Largest cluster	62606 (59.8%)
Ratio of Sizes	1.42

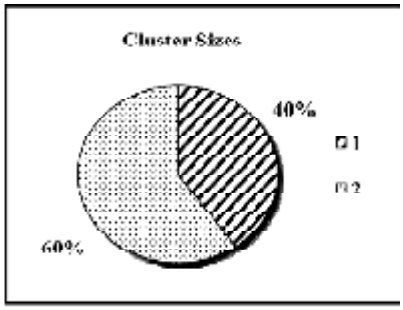

Рис. 7. Результат роботи двоетапного алгоритму без накладання обмежень

Model Summary	
Algorithm	TwoStep
Inputs	7
Clusters	3
Measure of cohesion and separation	0.3
Size of Smallest cluster	28602 (27.3%)
Size of Largest cluster	41432 (39.6%)
Ratio of Sizes	1.45

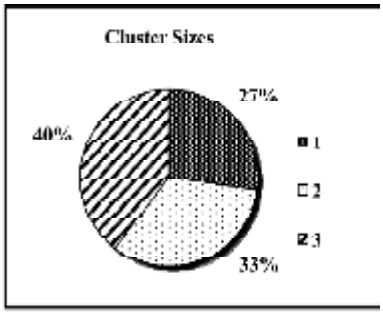

Рис. 8. Результат роботи двоетапного алгоритму з заданим числом кластерів, рівним 3

Спостереження, що були отримані як результати прогнозів на основі мультиноміальної регресії в подальшому кластеризуються за допомогою двоетапного алгоритму кластеризації.

Інтуїтивне пояснення доцільності запропонованої модифікації:

Наведемо інтуїтивне пояснення доцільності такого методу. Уявімо, що в нас наявна змінна, що відображає приналежність кожного спостереження до відповідного кластеру. В такому випадку ми могли б побудувати модель, що оцінює ймовірність приналежності кожного спостереження до певного кластеру. Наприклад, це можна було б зробити використовуючи мультиноміальну логістичну регресію.

Мультиноміальна логістична регресія фактично являє собою К звичайних логістичних регресій. Кожна регресія оцінює гіпотезу про те, чи спостереження належить до певного кластеру чи ні. Найбільш ймовірним кластером є той, що отримав найбільшу ймовірність в одній з оцінених регресій. Прогноз найбільш ймовірного кластеру для спостереження з заданими параметрами є логлінійною функцією від цих параметрів.

Недолік методу: метод показує погану результативність у разі некорельованості характеристик об'єктів. Наприклад, нехай є об'єкти, що характеризуються двома параметрами X1 та X2, кожен з яких приймає два значення 0 або 1 та вони є рівномірно розподіленими (рис. 3).

У такому випадку кожен з чотирьох варіантів буде обраний як центр кластеру. В той же час, для даного прикладу жоден з алгоритмів кластеризації не зможе більш коректно агрегувати дані чотири кластери. Це варто віднести до спільного недоліку алгоритмів без учителя, що є менш точні, ніж алгоритми з учителем, та слугують здебільшого для розвідувального аналізу.

Можна відзначити, що даний підхід неявно враховує кореляцію між змінними. Наприклад, нехай спостереження описуються двома параметрами та мають розподіл зображений на рисунку 4.

У результаті використання прогнозів на основі мультиноміальних регресій ми отримаємо наступний розподіл спостережень рис. 5.

У нашому випадку кількість значень кожного параметру рівна 2. А отже, кожна мультиноміальна регресія є звичайною логістичною регресією. Отримані прогнози на основі регресій добре відображають те, як мали б бути кластеризовані дані. Це добре зрозуміло у випадку двох змінних та двох значень у кожній з них. У випадку суттєвої кількості параметрів та значень кожного з них такі висновки були б важкодосяжні.

Можна зробити висновок, що використання запропонованої модифікації є певним аналогом використання відстані Махаланобіса як відстані між спостереженнями. Використання цієї відстані в алгоритмах кластеризації є особливо доречним при аналізі даних з високими рівнями кореляції між характеристиками [5].

Опис даних:

Практичне застосування вищеописаного алгоритму та порівняння з існуючими алгоритмами було зроблено на основі вибірки з 104681 спостереження, що являють собою аплікаційні дані фізичних осіб та інформацію про їх кредитну історію.

Інформація про кожного позичальника містить наступні характеристики:

— співвідношення щомісячних виплат по кредитам та щомісячного доходу;

— вік;

— місячний дохід;

— кількість відкритих кредитних ліній та кредитів;

— кількість іпотечних кредитів;

— кількість фактів просрочки, що перевищують 90 днів;

— кількість фактів просрочки від 60 до 90 днів;

— кількість фактів просрочки від 30 до 60 днів;

— кількість утриманців позичальника;

Дані було отримано з платформи Kaggle.

Kaggle — іноваційний варіант розв'язку статистичних, аналітичних задач. Це лідуюча платформа для змагань з прогнозування та моделювання. Компанії, уряди та дослідники презентують вибірки даних та проблеми. В свою чергу найкращі в світі дослідники та їх команди пропонують свої розв'язки [9].

Результати:

Для оцінки якості даного алгоритму буде проведено порівняння результатів кластеризацій на основі комбінації алгоритмів (мультиноміальні регресії + двоетапний алгоритм кластеризації) та безпосередньо двоетапного алгоритму кластеризації.

У результаті виконання ансамблю алгоритмів та просто двоетапного алгоритму кластеризації було отримано наступні результати:

— Ансамбль алгоритмів зупинився на утворенні трьох кластерів (рис. 6).

Просто двоетапний алгоритм кластеризації зупинився на утворенні двох кластерів (рис. 7).

Якість кластеризації виявилась кращою у випадку ансамблю алгоритмів. Це відображає горизонтальний індикатор на лівій частині зображення.

Для більш коректного порівняння було вирішено зупинити двоетапний алгоритм на етапі утворення 3 кластерів — за крок до утворення двох кластерів. Результати роботи двоетапного алгоритму кластеризації з фінальним кроком, на якому утворилися 3 кластери, представлені на рисунку 8.

Якість кластеризації виявилась кращою у випадку ансамблю кластерів, ніж у просто двоетапного алгоритму кластеризації. Можна зробити висновок, що запропонована модифікація відчутно покращує якість кластеризації.

ВИСНОВКИ ТА ЗАУВАЖЕННЯ

У даній роботі було запропоновано метод кластеризації, що є ансамблем методів кластеризації. Алгоритм складається з оцінок, уточнень значень кожної характеристики на основі низки мультиноміальних регресій та подальшої кластеризації за допомогою двоетапного алгоритму кластеризації.

Такий ансамбль моделей показав більш високу стійкість утворених кластерів, а отже його використання є доцільним. Використання прогнозів на основі мультиноміальних регресій покращує якість вхідних даних для двоетапного алгоритму кластеризації. Фактично відбувається прибирання шумів, проводиться попередня, найбільш груба агрегація даних.

Зауважимо, що запропонований метод можна використовувати тільки для категоріальних даних. Для його використання усі неперервні величини мають бути агреговані до категоріальних, інтервальних величин.

Перевагою даного методу можна відзначити врахування кореляції між змінними. Ця особливість є особливо актуальною при роботі з даними, що відображають фінансовий стан суб'єктів господарювання: підприємств нефінансового сектору, банків, страхових компаній, фізичних осіб.

У подальших роботах планується модифікувати метод для можливого використання неперервних величин без попереднього групування.

Література:

1. Пономаренко В.С. Моделирование социально-экономических систем: теория и практика: монография / Под. ред. В.С. Пономаренко, Т.С. Клебановой, Н.А. Кизима. — Х.: ФЛП Александрова К.М.; ИД "ИНЖЭК", 2012. — С. 375—386.

2. Chiu T. A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. / T. Chiu, D. Fang, J. Chen, Y. Wang, and C. Jeris // In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, CA: ACM. — 2001.

3. Lloyd S. Least square quantization in PCM's. Bell Telephone Laboratories Paper. — 1957.

4. MacQueen J. B. Some methods for classification and analysis of multivariate observations. In Proc. 5th Berkeley Symp. on Math. Statistics and Probability. University of California Press. — 1967. Pages 281—297.

5. Steinhaus H. Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci. — 1956. C1. III vol IV: 801-804.

6. Zhang T. BIRCH: An efficient data clustering method for very large databases./ T. Zhang, R. Ramakrishnon, and M. Livny // In: Proceedings of the ACM SIGMOD Conference on Management of Data. Montreal, Canada: ACM. — 1996.

7. Матеріали конференції CS769 Spring 2010 Advanced Natural Language Processing [Електронний ресурс]. — Режим доступу: http:// pages.cs.wisc.edu

8. Метод k-середніх: [Електронний ресурс]. — Режим доступу: http://en.wikipedia.org/wiki/K-means_-clustering.

9. Платформа для збору і обробки даних: [Електронний ресурс]. — Режим доступу: http://www.kaggle.com.

References:

1. Ponomarenko, V.S., Klebanova, T.S., Kizima, N.A. (2012), Modelirovanie sotsialno-ekonomicheskih system: teoriia i praktika [Modelling of social-economic systems: theory and practice], ID "INGEK", Kharkov, Ukraine, p. 375—386.

2. Chiu, T. (2001), "A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment" Seventh ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco.

3. Lloyd, S. (1957), "Least square quantization in PCM's", Bell Telephone Laboratories Paper.

4. MacQueen, J. B. (1967), "Some methods for classification and analysis of multivariate observations", 5th Berkeley Symp. on Math. Statistics and Probability, University of California Press, pp. 281—297.

5. Steinhaus, H. (1956), "Sur la division des corps materiels en parties", Bull. Acad. Polon. Sci, C1. III, vol. IV, pp. 801—804.

6. Zhang, T., Ramakrishnon, R., Livny, M. (1996), "BIRCH: An efficient data clustering method for very large databases", ACM SIGMOD Conference on Management of Data, Montreal, Canada: ACM.

7. CS769 Spring 2010 Advanced Natural Language Processing (2010), available at: http://pages.cs.wisc.edu (Accessed 6 August 2013).

8. Wikipedia (2013), "Method of k-avaraged", available at: http://en.wikipedia.org/wiki/K-means_clustering (Accessed 3 September 2013).

9. Platform for data collection and processing (2013), available at: http://www.kaggle.com (Accessed 3 September).

Стаття надійшла до редакції 10.09.2013 р.