Савчук Т. Удосконалений метод кластеризації станів комп'ютерної техніки K-MEANS / Т. Савчук, С. Петришин // Вісник ТНТУ — Тернопіль : ТНТУ, 2015. — Том 78. — № 2. — С. 198-206. — (Приладобудування та інформаційно-вимірювальні технології).

УДК 004.89

Т. Савчук, к.т.н.; С. Петришин

Вінницький національний технічний університет

УДОСКОНАЛЕНИЙ МЕТОД КЛАСТЕРИЗАЦІЇ СТАНІВ КОМП'ЮТЕРНОЇ ТЕХНІКИ K-MEANS

Робота присвячена удосконаленню методу кластеризації станів комп'ютерної техніки K-MEANS з метою підвищення якості розбиття множини таких станів. При дослідженні відомих модифікацій означеного методу щодо можливості їх застосування при аналізі станів комп'ютерної техніки було виявлено недостатню точність віднесення такого стану до певного кластеру через випадковий вибір їх початкових центрів.

Виявлений недолік було усунуто шляхом визначення початкових центрів кластерів на основі значень потенціалів, а також, виділення в окремий таксон станів, які могли бути помилково віднесені до кластера за рахунок допустимих відхилень значень параметрів та характеристик таких станів, що дозволило підвищити якість розбиття множини станів комп'ютерної техніки в середньому на 7%.

***Ключові слова**: метод K-MEANS, кластеризація, стан комп'ютерної техніки.*

T. Savchuk; S. Petrishyn

IMPROVED METHOD OF CLUSTERING STATES OF COMPUTER EQUIPMENT K-MEANS

The work is dedicated to improving the method of clustering classes of computer equipment K-MEANS to improve the quality partition of such states. The object of research is the process of clustering state of computer equipment. Subject of research – methods of cluster analysis states of computer equipment.

Relevance of these studies is conditioned by the rapid scientific and technical progress, in which significantly increased the number of computer equipment, which is used in various fields, and therefore increases the likelihood of situations specific to this equipment, given the diversity of functions that it performs. Thus, depending on the state of computer equipment taken various administrative decisions regarding its further functioning. So important is the development or improvement of methods of cluster analysis state of computer equipment that will determine the decision on its further functioning.

In the analysis of decomposition of objects that can be used for solving the problem of clustering state of computer equipment, it was determined that such methods must be clear, non-hierarchical and scalable, expressed by characteristics inherent in the clustering method K-MEANS. When tested method known modifications appointed on the possibility of their use in the analysis of the state of computer equipment was found insufficient accuracy of classification of this state to a cluster through a random selection of initial centers.

Identified deficiencies have been corrected by determining the initial cluster centers based on the values of potentials, as well as the allocation of a separate cluster of conditions that could be mistakenly attributed to the cluster due to tolerances values of parameters and characteristics of these states, thus improving the quality of the partition of the states of computer equipment an average of 7%.

***Key words**: method K-MEANS, clustering, state of computer equipment.*

Вступ. Аналіз станів комп'ютерної техніки (КТ) потребує швидкого опрацювання великих об'ємів інформації про такі стани, з такими завданнями можуть справитись технології Data Mining. Основними із задач Data Mining є: класифікація, пошук асоціативних правил і кластеризація [1].

Розглянемо основні стадії та етапи життєвого циклу комп'ютерної техніки. На кожному з них доцільно використовувати різні методи Data Mining для аналізу інформації про її стани з метою надання рекомендацій щодо подальших дій, враховуючи той чи інший стан. Необхідно відзначити, що кластеризацію доцільно застосовувати на етапах розробки документацій з обслуговування такої техніки, оскільки на даному етапі не відомо на які саме класи або кластери буде розбито їх множину та які характеристики буде враховано [2].

Постановка задачі. Нехай X – матриця, в якій кожен рядок $\{x_{i1},\ldots,x_{ij},\ldots,x_{im}\}$ описує певний стан КТ.

$$X=\{X_1,X_2,...,X_n\}=\begin{Bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{Bmatrix},$$

де X_i – стан комп'ютерної техніки;

x_{ij} – значення j-го параметра або характеристики i-го стану КТ;

m – кількість параметрів та характеристик станів, що збережені в базі даних (БД);

n – кількість станів КТ, що збережені в БД.

Тоді, на основі зазначеного, задача кластеризації станів КТ зводиться до розбиття їх вибірки на $k\,(k \le n)$ непересічних підмножин, що називаються кластерами, так, щоб:

- кожен кластер складався зі станів, близьких за метрикою ω (при яких формуються подібні рекомендації щодо подальших дій при їх виникненні);
- стани, які знаходяться в різних кластерах, значно відрізнялися (при яких формуються різні рекомендації щодо подальших дій при їх виникненні) [1, 3].

При цьому кожному стану $X_i \in X$ приписується номер його кластера. І в залежності від номера кластера користувач отримує необхідну інформацію про формалізовану групу, до якої було віднесено стан.

Аналіз методів кластеризації станів комп'ютерної техніки. Розглянемо основні методи кластеризації, що можуть бути використані при аналізі станів КТ [1, 4–8].

1) За способом обробки даних розрізняють такі методи кластеризації:

- неієрархічні методи, що застосовуються у випадках, коли наперед відомо кількість кластерів об'єктів та характеризуються високою стійкістю до шумів і викидів, некоректного вибору метрики, віднесення незначущих змінних в набір, який бере участь в кластеризації, що особливо важливо при проведенні кластерного аналізу станів КТ в реальному часі. Крім того, означені методи дозволяють опрацьовувати потужні БД параметрів та характеристик станів КТ;
- ієрархічні методи, які застосовують у випадках, коли наперед невідомо число кластерів об'єктів, для яких характерним є висока чутливість до викидів у кластерах та некоректного вибору метрики. Особливістю означених методів є те, що вони орієнтовані на роботу з не потужними БД, що не є характерним для опису станів КТ.

2) За способом аналізу даних:

- чіткі методи кластеризації, що характеризуються розбиттям множини об'єктів на кластери, що не перетинаються, що є важливим при аналізі станів КТ, оскільки для кожного з станів має бути вироблено управлінські рішення характерні для кожного із кластерів;

– нечіткі методи, для яких є допустимим віднесення одного і того ж об'єкта одночасно до кількох або навіть всіх кластерів з різною ймовірністю, що при проведенні кластерного аналізу станів КТ може призвести до неоднозначності при прийнятті рішень щодо подальшого її функціонування.

3) За можливістю розширення обсягу даних, що обробляються:

– масштабовані, що передбачають можливість проведення кластерного аналізу над множиною, кількість об'єктів якої може змінюватись, що є характерною рисою для КТ з урахуванням різних умов її експлуатації;

– немасштабовані, характеризуються проведенням декомпозиції незмінної множини об'єктів, що не є характерним для предметної області, яка аналізується.

Таким чином, за результатами проведеного аналізу методів, що можуть бути використані для кластеризації станів КТ, було з'ясовано, що є найбільш прийнятними методами кластерного аналізу для розв'язання поставленої задачі є неієрархічні чіткі масштабовані методи, що дозволяють працювати з потужними обсягами даних, а результатом їх роботи є непересічні кластери станів комп'ютерної техніки.

Розробка удосконаленого методу кластеризації станів комп'ютерної техніки K-MEANS. Одним із описаних методів є метод K-MEANS. Використання відомого методу K-MEANS для кластеризації станів КТ при здатності працювати з потужними обсягами даних і при наперед визначеній кількості кластерів k, має такі недоліки [1, 9–13]:

– вибір випадковим чином станів КТ, які і будуть вважатись початковими центрами кластерів, негативно впливає на якість кластеризації, тобто на точність віднесення такого стану до коректного кластера;

– чутливість до викидів, які можуть бути в множині вхідних даних, негативно впливає та якість кластеризації;

– існуючі модифікації методу не враховують особливостей КТ, що негативно впливає на якість кластеризації станів.

Означені недоліки існуючого методу K-MEANS можна усунути його удосконаленням, що дозволяє підвищити якість розбиття множини станів КТ за рахунок визначення початкових центрів кластерів на основі значень потенціалів, а також виділення в окремий таксон станів, які могли бути помилково віднесені до кластера за рахунок допустимих відхилень значень параметрів та характеристик такої техніки, що враховує особливості предметної області. Для цього обчислюється для кожного стану значення потенціалу, який показує можливість формування кластера в його околі

$$P_1(X_i) = \sum_{j=1, j\neq i}^{n} \exp\left(-\varepsilon \cdot a_{зE}(X_i, X_j)\right), i = 1, \ldots, n, \qquad (1)$$

де $P_1(X_i)$ – значення першого потенціалу станів КТ;

n – кількість станів КТ, що підлягають кластеризації;

$a_{зE}(X_i, X_j)$ – відстань між станами X_i та X_j;

ε – додатна константа, яка характеризує масштаб відстаней між станами КТ

$$\varepsilon = \frac{1}{a_{зE}(X_i, X_j)}, \qquad (2)$$

де $\overline{a_{зE}(X_i, X_j)}$ – середнє значення відстаней між станами КТ.

Чим щільніше розміщені стани в околі потенційного центра кластера, тим вище значення його потенціалу. Центром першого кластера μ_{y_1} обирається стан з найбільшим потенціалом. Оскільки декілька станів КТ з найбільшими значеннями потенціалів, як правило, розміщені поруч, то для знаходження інших центрів кластерів необхідно усунути вплив знайденого. Для цього значення потенціалів перераховується таким чином: від поточного значення віднімається значення потенціалу знайденого центра

$$P_j(X_i) = P_{j-1}(X_i) - P_{j-1}(\mu_{y_{(j-1)}}) \cdot \exp\left(-\beta \cdot a_{зE}(X_i, \mu_{y_{(j-1)}})\right), i = 1, \ldots, n, \quad (3)$$

де $P_j(X_i)$ – значення *j*-го потенціалу станів КТ;

n – кількість станів КТ, що підлягають кластеризації;

$a_{зE}(X_i, \mu_{y_{(j-1)}})$ – відстань між станом X_i та центром кластера $\mu_{y_{(j-1)}}$;

β – додатна константа, яка характеризує масштаб розміру одного кластера ($\beta \approx \varepsilon$).

Центрами кластерів μ_{y_j} на кожному кроці обираються ситуації з найбільшим значенням потенціалу. Таку операцію необхідно проводити до тих пір, поки не буде знайдено k попередніх центрів кластерів.

Після формування множини початкових центрів кластерів виконується розбиття вибірки станів КТ з використанням відомого підходу, що передбачає віднесення всіх станів до одного з k кластерів – того, відстань до центра якого є мінімальною. Далі виконується уточнення місця розташування центрів кожного кластера, якими стають стани КТ, їхні ознаки розраховуються як середнє арифметичне ознак станів, що входять до цього кластера

$$\mu_{y_j} = \frac{1}{|y_j|} \sum_{X_i \in y_j} X_i, y \in Y, j = 1, \ldots, k. \quad (4)$$

де y_j – *j*-тий кластер станів КТ.

Виконується така кількість ітерацій розбиття, поки центри кластерів стануть стійкими (тобто при кожній ітерації центрами кластерів будуть одні й ті самі стани). Тоді дисперсія всередині кластера буде мінімізована, а між кластерами – максимізована.

З метою підвищення якості розбиття множини станів КТ відбувається ініціалізація додаткового кластера та віднесення до нього станів, які можуть бути зараховані до кластерів з урахуванням допустимих відхилень значень параметрів та характеристик таких станів.

В формалізованому вигляді алгоритм K-MEANS (рисунок 1), в основу якого покладено запропонований удосконалений метод, містить такі основні кроки.

4) Задати кількість кластерів k, на які буде розбито множину станів КТ.
5) Знайти значення потенціалів всіх станів КТ з використанням формули (1).
6) Визначити стан КТ, який може бути центром першого кластера за формулою

$$\mu_{y_1} = \underset{x_i \in X}{\operatorname{argmax}} P_1(X_i),\ i = 1, \ldots, n. \tag{5}$$

7) ($j = 1$) поки $j \le k$ виконувати:

4.1) Порядковий номер *j* поточного потенціалу збільшити на 1.

4.2) Обчислити значення поточного потенціалу для всіх станів, за винятком тих, які були обрані як початкові центри кластерів за формулою (3).

4.3) Визначити точку, яка може бути центром *j*-го кластера

$$\mu_{y_j} = \underset{x_i \in X}{\operatorname{argmax}} P_j(X_i),\ i = 1, \ldots, n. \tag{6}$$

8) Віднести кожен із станів до найближчого центра кластера

$$y_j = \underset{y \in Y}{\operatorname{argmin}} a_{3E}(X_i, \mu_y),\ i = 1, \ldots, n, \tag{7}$$

де y_j – кластер станів КТ.

9) Обчислити нове положення центрів з використанням виразу (4).

10) Виконувати пункти 5–6, поки центри кластерів y_i не перестануть змінюватись.

11) Ініціалізувати додатковий кластер станів КТ $y_0 = \varnothing$.

12) Обчислити значення допустимого відхилення відстані Δ, яка визначається з використанням зваженої евклідової відстані.

13) Для всіх кластерів $y_i (i = 1, \ldots, n)$ виконувати:

10.1) Поки $|y_0| \ne const$ для всіх центрів кластерів $\mu_{y_j}\ (j \ne i)$ виконувати:

10.1.1) Знайти

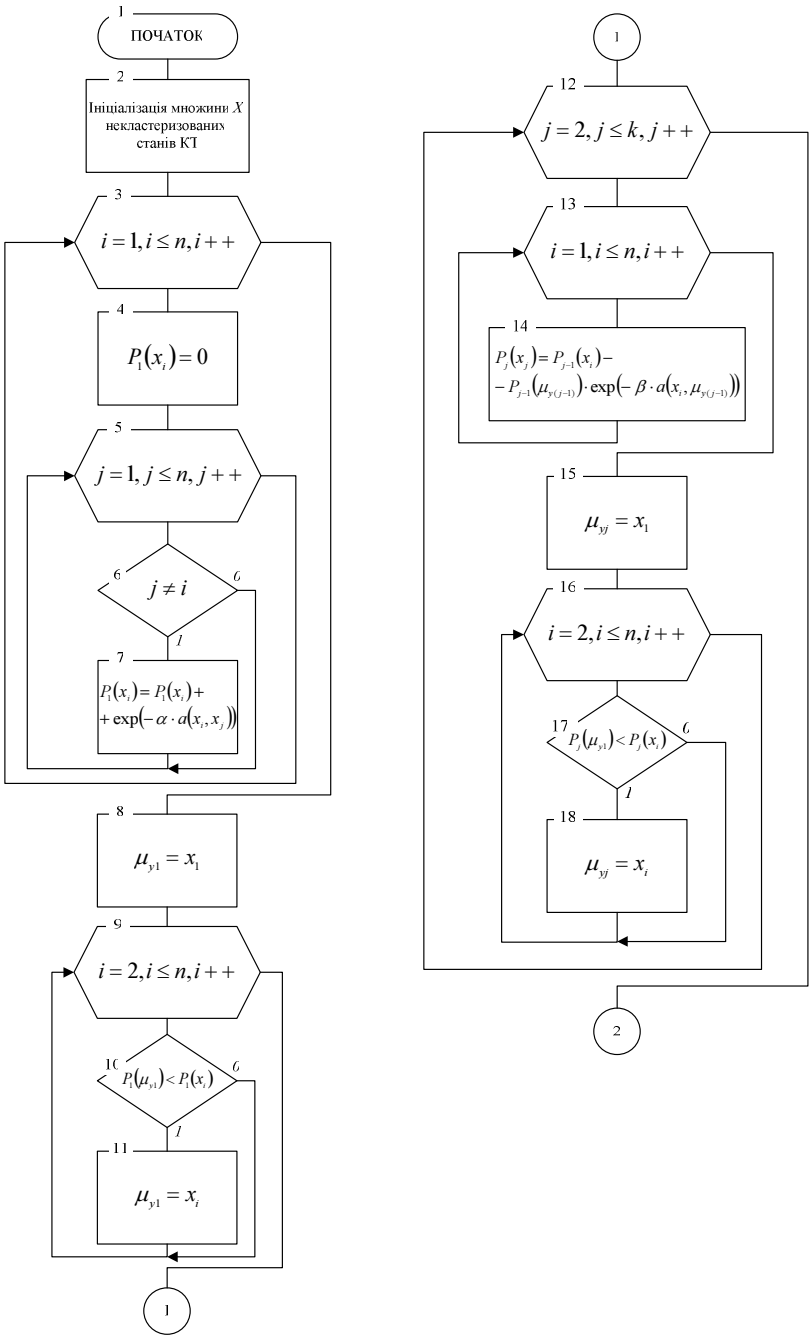

Рисунок 1. Схема удосконаленого алгоритму кластеризації станів КТ K-MEANS

Figure 1. The scheme ща improved algorithm of clustering states of computer equipment K-MEANS

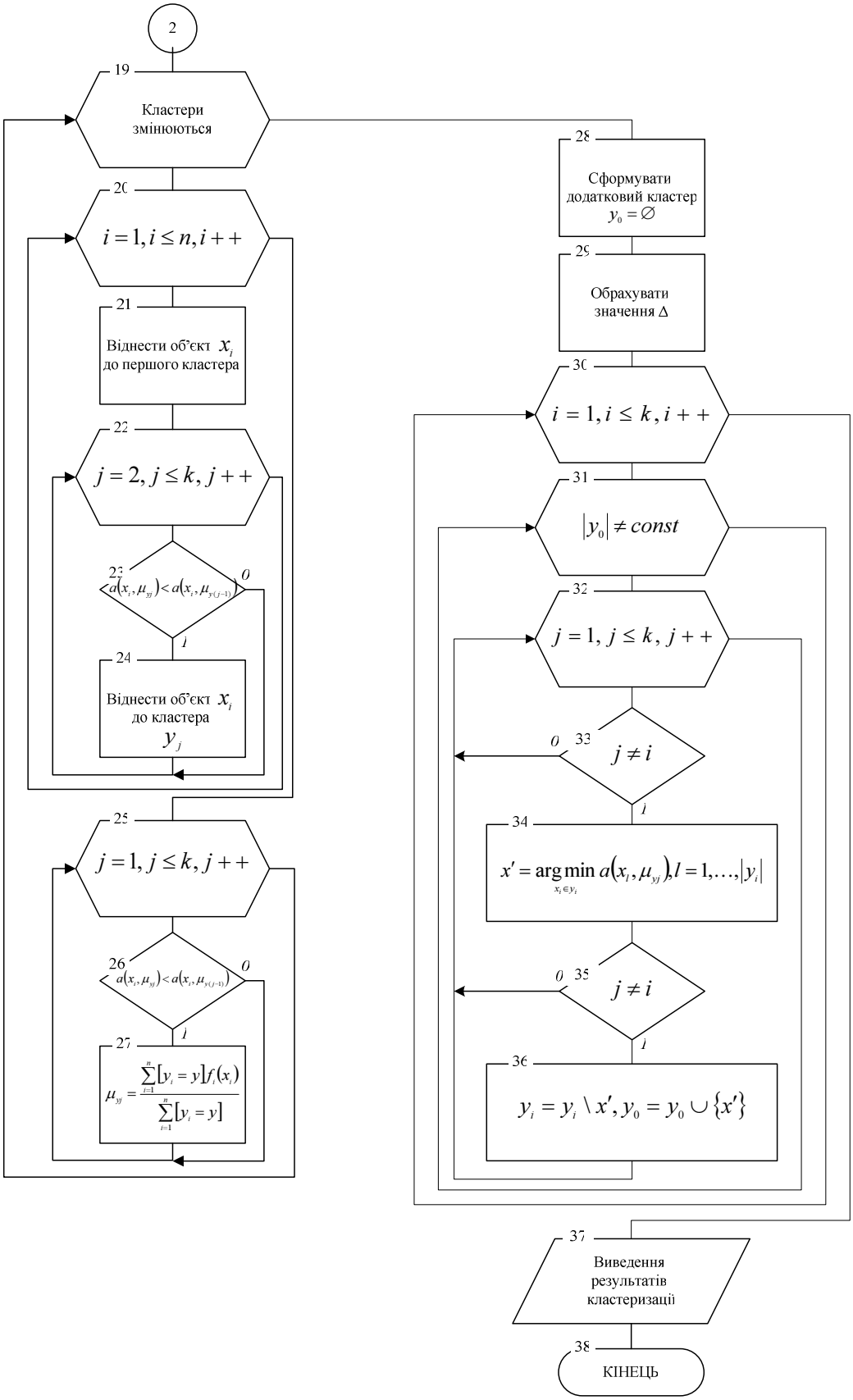

Рисунок 1. Продовження

Figure 1. Continuation

$$y_i = y_i \setminus X', y_0 = y_0 \cup \{X'\}. \quad (9)$$

$$X' = \arg\min_{X_l \in y_i} a_{зE}\left(X_l, \mu_{y_j}\right), l = 1, \ldots, |y_i|, \quad (8)$$

де X' – найближчий стан до центра кластера $y_j (j \neq i)$.

10.1.2. Якщо $a_{зE}(X', \mu_{y_j}) - \Delta < a_{зE}(X', \mu_{y_i}) - \Delta$ то

Отже, запропоновано удосконалення існуючого методу кластеризації K-MEANS визначення початкових центрів кластерів на основі потенціалів, а також відповідний алгоритм K-MEANS, що дозволить підвищити якість розбиття множини станів КТ на таксони.

Висновки. Таким чином, удосконалено метод кластеризації станів комп'ютерної техніки K-MEANS, що передбачає визначення початкових центрів кластерів на основі потенціалів, а також враховує особливості станів такої техніки та виділяє в окремий таксон стани, які могли бути помилково віднесені до кластера за рахунок допустимих відхилень значень параметрів та характеристик. Це дозволило підвищити якість кластеризації станів комп'ютерної техніки на 7%.

Conslusions. Thus, improved clustering method K-MEANS for computer equipment states, which involves determining the initial cluster centers based on potential and takes into account the states of this equipment and provides a separate taxon conditions that may be erroneously attributed to the cluster due to tolerances parameter values and characteristics. This enhanced the quality of clustering states of computer equipment by 7%.

Список використаної літератури:

1. Барсегян А. А. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – [2-е изд., перераб. и доп.]. – Спб. : БХВ, Петербург, 2007. – 384с. : ил. – ISBN 5-94157-991-8.
2. Партыка Т. Л. Вычислительная техника / Т. Л. Партыка, И. И. Попов. – [2-е изд., перераб. и доп.]. – М. : ФОРУМ: ИНФА-М, 2007. – 608с. : ил. – ISBN 5-91134-050-X.
3. Савчук Т. О. Використання ієрархічних методів кластеризації для аналізу надзвичайних ситуацій на залізничному транспорті / Т. О. Савчук, С. І. Петришин // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2009. – № 1. – С. 193-198. – ISSN 2219-9365.
4. David G. Kleinbaum Applied Regression Analysis and Other Multivariable Methods / David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam – [5 edition]. – Cengage Learning, 2013. – 1072 p. – ISBN 978-1285051086.
5. Data Mining – Cluster Analysis [Electronic resource]. – Access mode : http://www. tutorialspoint. com/data_mining/dm_cluster_analysis. htm.
6. Гитис П. Х. Статистическая классификация и кластерний анализ / П. Х. Гитис – М. : Московский государственный горный университет, 2003. – 157 с.
7. Файсал М. Е. Сардієх. Методи і алгоритми неієрархічної кластеризації для задач інтелектуального аналізу даних: автореф. дис. : 05. 13. 06 / Файсал М. Е. Сардієх ; Нац. ун-т «Львів. політехніка». — Л., 2011. — 20 с.
8. Jain A. Data clustering: A review / Jain A., Murty M., Flynn P. – ACM Computing Surveys. 1999. – P. 264–323.
9. Савчук Т. О. Розробка модифікованого алгоритму K-MEANS для аналізу надзвичайних ситуацій на залізничному транспорті / Т. О. Савчук, С. І. Петришин // Обчислювальний інтелект (результати, проблеми, перспективи) : матеріали 1-ї Міжнародної науково-технічної конференції (10–13 травня 2011 р.). – Черкаси, 2011. – С. 236-237.
10. Савчук Т. О. Порівняльний аналіз використання методів кластеризації для ідентифікації надзвичайних ситуацій на залізничному транспорті / Т. О. Савчук, С. І. Петришин // Системний аналіз та інформаційні технології : матеріали 12-ї Міжнародної науково-технічної конференції SAIT 2010, Київ, 25–29 травня 2010 р. / ННК «ІПСА» НТУУ «КПІ». – К. : ННК «ІПСА» НТУУ «КПІ», 2010. – С. 485.

11. Tapas Kanungo An Efficient k-Means Clustering Algorithm: Analysis and Implementation / Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu // IEEE Transactions on Pattern Analysis and Machine Intelligence - Los Alamitos, USA, 2002 - №24(7), С. 881–892.
12. An Example Inference Task: Clustering [Electronic resource]. – Access mode : http://www. inference. phy. cam. ac. uk/mackay/itprnn/ps/284. 292. pdf.
13. Andrea Vattani K-means Requires Exponentially Many Iterations Even in the Plane [Electronic resource] / Andrea Vattani – Access mode : http://cseweb. ucsd. edu/~avattani/papers/kmeans-journal. pdf.

Отримано 06.04.2015