УДК 004.852; 004.93'14; 004.94

П.О. Кравець
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

СТОХАСТИЧНА ІГРОВА МОДЕЛЬ КЛАСТЕРИЗАЦІЇ ДАНИХ

Запропонована стохастична ігрова модель кластеризації даних в умовах дії завад. Розроблено адаптивний рекурентний метод та алгоритм розв’язування стохастичної гри. Виконано комп’ютерне моделювання ігрової кластеризації зашумлених даних. Вивчено вплив параметрів на збіжність стохастичного ігрового методу кластеризації даних. Проаналізовано отримані результати.

Ключові слова: кластеризація даних, стохастична ігрова модель, адаптивний ігровий метод.

The stochastic game model of the data clustering under the influence of noise is offered. Adaptive rerrurent method and algotithm of stochastic game solving are developed. Computer modelling of game of noisy data clustering is executed. The parameter influences on convergence of stochastic game method of the data clustering are studied. The analysis of received results is realized.

Key words: data clustering, stochastic game model, adaptive game method.

Вступ

Розв’язування задач інтелектуального аналізу та візуалізації даних, групування та розпізнавання образів, видобування знань та пошуку інформації, класифікації об’єктів можуть бути виконані методами кластерного аналізу. Метою кластерного аналізу є знаходження груп подібних об’єктів у заданій множині або вибірці. На відміну від дискримінантного аналізу, де класи є наперед заданими, кластерним аналізом визначають склад кластерів [1, 2].

Кластерний аналіз використовують у хімії, біології, медицині, соціології, педагогіці, психології, філології, маркетингу, опрацюванні сигналів, розпізнаванні образів, документознавстві, інформатиці, науковій роботі та інших галузях людської діяльності для організації даних у вигляді класів з метою їх систематизації та групового аналізу [3 – 10].

Кластеризація – це поділ множини об’єктів на підмножини залежно від їх подібності. Виділені підмножини називаються кластерами. Елементи одного кластера мають спільні властивості. Елементи різних кластерів значно відрізняються між собою.

Загальна схема кластеризації є такою:

1) виділення характеристик об’єктів;
2) визначення метрики об’єктів;
3) розділення множини об’єктів на кластери;
4) інтерпретація результатів кластеризації.

У постановках задач кластеризації кількість кластерів може задаватися або бути невідомою апріорі [11].

Нехай кожен об’єкт $x \in X$ із множини об’єктів $X = (x_1, x_2, ..., x_L)$ описується вектором властивостей $x = (x[1], x[2], ..., x[m])$, які можуть бути кількісними або якісними характеристиками об’єкта.

Подібність двох об’єктів x_i та x_j визначається метрикою їх близькості $D(x_i, x_j)$ у просторі характеристик. Як метрики використовують евклідову відстань, манхеттенівську відстань, відстань Чебишова, відсоток невідповідності, коефіцієнт кореляції Пірсона тощо [3].

Для поділу множини об’єктів на кластери найчастіше використовують такі методи [3, 4]:

1) ієрархічна деревоподібна кластеризація;
2) метод *k*-середніх;
3) метод найближчого або найвіддаленішого сусіда;
4) метод незваженого або зваженого попарного середнього;
5) методи нечіткої кластеризації;
6) застосування нейронних мереж;
7) генетичні алгоритми;
8) метод загартування.

У загальному випадку кластеризацію об’єктів можна розглядати як задачу оптимального поділу об’єктів на групи. Критерієм оптимізації може бути мінімізація середньоквадратичної похибки виділення кластерів:

$$\delta = \sum_{j=1}^{N} \sum_{i=1}^{C_j} \left\| x_i^{(j)} - \overline{x}_j \right\|^2 \to \min ,$$

де $\overline{x}_j$ – центр мас j-го кластера, точка у просторі характеристичних векторів із середнім для цього кластера значенням характеристик; C_j – кількість елементів j-го кластера.

Завершальним етапом кластерного аналізу є змістова інтерпретація сформованих кластерів, під час якої виявляють фактори або причину групування об’єктів у кластери. Тут необхідно зважати на те, що різні методи кластеризації можуть породжувати різні кластерні рішення. Крім того, метод кластеризації може виявляти привнесені структури даних, яких насправді немає в аналізованих даних. Тому необхідно обирати такі методи кластеризації, які дають найбільш усвідомлені рішення у досліджуваній предметній області. До оцінювання якості кластеризації залучають експертів відповідних предметних областей.

На основі результатів кластерного аналізу можна класифікувати об’єкти, виявляти концептуальні схеми групування об’єктів, перевіряти та формувати гіпотези щодо моделей організації даних, стиснення даних заміною кластера його типовим елементом, виявлення новизни даних, які не увійшли до жодного із кластерів.

Для розв’язування задач кластеризації даних розроблено засоби, до яких належать популярні програмні пакети (Matlab, Fuzzy Clustering and Data Analysis Toolbox, Cluster Validity Analysis Platform) та комерційні розробки (SPSS Statistics, STATISTICA). Програмні засоби кластеризації даних надають такі можливості: 1) методи кластеризації даних (K-means, K-medoid, PAM, Hierarchical, SOM, FCMclust, GKclust, GGclust та ін.); 2) функції аналізу, призначені для оцінювання фіксованих розбиттів на кластери методами, основаними на індексах (Dunn, Alternative Dunn, Xie and Beni’s, Partition index та ін.); 3) функції візуалізації, які відображають дані у простір меншої розмірності (Sammon); 4) демонстраційні приклади, які реалізують алгоритми для реальних промислових даних.

У практичних застосуваннях призначені для кластеризації дані, як правило, містять елементи невизначеності. Це можуть бути нечітко задані характеристики об’єктів, пропущені атрибути об’єктів у базах даних, зашумлені сигнали тощо. В умовах невизначеності застосовують методи нечіткої, адаптивної кластеризації, генетичні алгоритми, нейронні мережі з навчанням без учителя [12 – 16].

Кластеризація даних формулюється як конкурентна або кооперативна задача віднесення об’єкта до того або іншого кластера. Проблеми конкуренції та кооперації об’єктів вивчає теорія ігор, а в умовах невизначеності – теорія стохастичних ігор [17]. Тому актуальним з наукового, пізнавального та практичного кутів зору є застосування методів стохастичних ігор для кластеризації даних в умовах невизначеності.

Метою цієї роботи є побудова ігрової моделі кластеризації даних з елементами невизначеності. Для досягнення мети необхідно розв’язати такі задачі: сформулювати задачу ігрової кластеризації даних, розробити адаптивний ігровий метод та алгоритм розв’язування задачі, розробити комп’ютерну програмну модель, проаналізувати та інтерпретувати отримані результати.

Постановка ігрової задачі

Нехай множину $X=\{x_1,x_2,...,x_L\}$ задано координатами точок $x\in R^m$ в m-вимірному параметричному просторі. Координати точок визначають нормалізований характеристичний вектор призначених для кластеризації об'єктів. У цій множині необхідно виділити N кластерів $\left\{Y_n, n=1..N \middle| \bigcup_{n=1..N} Y_n = X,\ Y_i \bigcap_{i\neq j} Y_j = \varnothing\ \forall (i,j)\in\{1..N\}\right\}$ за критеріями

$$\frac{1}{C_n}\sum_{x\in Y_n}\left\|x_l - x_k\right\| \to \min,\ n=1..N, \tag{1}$$

де $C_n = |Y_n|$ – кількість елементів, що входять до кластера Y_n; $\|*\|\in R^1$ – евклідова норма вектора.

Розподіляємо множину X на кластери Y_n ($n=1..N$)методом стохастичної гри, яка задається кортежем $(I, A^i, \Xi^i \mid i\in I)$, де I – множина гравців; $L=|I|$ – кількість гравців; $A^i=\{a^i(1),...,a^i(N)\}$ – множина чистих стратегій i-го гравця, які визначають вибір одного із кластерів; N – кількість стратегій i-го гравця; $\Xi^i: A\to R^1$ – функція програшів i-го гравця; $A=\underset{i\in I}{\times} A^i$ – множина комбінованих стратегій гравців.

Суть гри полягає у випадковому переміщенні гравців з одного кластера до іншого. Для цього у моменти часу $t=1,2,...$ кожен гравець на основі генератора випадкових подій незалежно від інших вибирає чисту стратегію $a^i\in A^i$, яка визначає його входження у відповідний кластер. Із урахуванням (1), після реалізації комбінованого варіанта $a\in A$ гравці отримують випадкові програші $\xi^i(a)$ з апріорі невідомими стохастичними характеристиками:

$$\xi_t^i = \frac{1}{C_t^i}\sum_{j\in I}\chi\left(a_t^i = a_t^j\right)\left\|x^i - x^j\right\| + \mu \quad \forall i\in I, \tag{2}$$

де $C_t^i = \sum_{j\in I}\chi\left(a_t^i = a_t^j\right)$ – поточна кількість елементів кластера, до якого входить i-й гравець; $\chi(*)\in\{0,1\}$ – індикаторна функція події; $\mu \sim Normal(0,d)$ – нормально розподілена випадкова величина, яка моделює невизначеність системи; d – дисперсія розподілу.

Ефективність ходу гри визначається функціями середніх програшів:

$$\Xi_t^i = \frac{1}{t}\sum_{\tau=1}^{t}\xi_\tau^i \quad \forall i\in I. \tag{3}$$

Мета гри полягає у мінімізації системи функцій середніх програшів (3) у часі:

$$\overline{\lim_{t\to\infty}}\,\Xi_t^i \to \min \quad \forall i\in I. \tag{4}$$

Отже, на основі спостереження поточних програшів $\{\xi_n^i\}$ кожен гравець $i\in I$ повинен навчитися вибирати чисті стратегії $\{a_t^i\}$ так, щоб з часом $t=1,2,...$ забезпечити виконання системи критеріїв (4).

Розв'язки ігрової задачі задовольнятимуть одну з умов колективної рівноваги, наприклад, Неша або Парето залежно від методу формування послідовностей стратегій $\{a_t^i\}\forall i\in I$.

Метод розв'язування задачі

Розв'язування стохастичної гри (1) виконаємо за допомогою адаптивного рекурентного перетвореннях векторів змішаних стратегій.

Побудову методу розв'язування стохастичної гри виконаємо на основі стохастичної апроксимації умови доповняльної нежорсткості детермінованої гри, справедливої для змішаних стратегій у точці рівноваги за Нешем [18].

Для цього визначимо полілінійну функцію середніх програшів детермінованої гри:

$$V^i(p) = \sum_{a\in A} v^i(a)\prod_{j\in I; a^j\in a} p^j(a^j),$$

де $v(a)=M\{\xi_t^i(a)\}$.

Тоді векторна умова доповняльної нежорсткості (CS, Complementary Slackness) матиме вигляд:

$$\vec{\mathrm{CS}}=\nabla_{p^i}V^i(p)-e^{N_i}V^i(p)=0 \quad \forall i\in D,$$

де $\nabla_{p^i}V^i(p)$ – градієнт функції середніх програшів; $e^N=(1_j \mid j=1..N)$ – вектор, всі компоненти якого дорівнюють 1; $p\in S^M$ – комбіновані змішані стратегії гравців, задані на опуклому одиничному симплексі S^M ($M=N^L$).

Для врахування розв'язків у вершинах одиничного симплексу виконаємо зважування умови доповняльної нежорсткості елементами векторів змішаних стратегій:

$$diag(p^i)(\vec{CS})=0 \quad \forall i\in D, \tag{5}$$

де $diag(p^i)$ – квадратна діагональна матриця порядку N, побудована з елементів вектора p^i.

Враховуючи, що $diag(p^i)[\nabla_{p^i}V^i-e^{N_i}V^i]=E\{\xi_t^i[e(a_t^i)-p_t^i]\mid p_t^i=p^i\}$, з (5) на основі методу стохастичної апроксимації отримаємо рекурентну залежність:

$$p_{t+1}^i=\pi_{\varepsilon_{t+1}}^N\left\{p_t^i-\gamma_t\xi_t^i(e(a_t^i)-p_t^i)\right\} \quad \forall i\in I, \tag{6}$$

де $\pi_{\varepsilon_{t+1}}^N$ – проектор на одиничний N-вимірний симплекс S^N [19]; $\gamma_t>0$, $\varepsilon_t>0$ – монотонно спадні послідовності додатних величин; $e(a_t^i)$ – одиничний вектор, який вказує на вибір чистої стратегії $a_t^i=a^i\in A^i$.

Параметри γ_t та ε_t визначають умови збіжності стохастичної гри і можуть бути задані так:

$$\gamma_t=\gamma t^{-\alpha},\ \varepsilon_t=\varepsilon t^{-\beta}, \tag{7}$$

де $\gamma>0$; $\alpha>0$; $\varepsilon>0$; $\beta>0$.

Збіжність стратегій (6) до оптимальних значень з імовірністю 1 та у середньоквадратичному визначається співвідношеннями параметрів γ_t та ε_t, які повинні задовольняти базові умови стохастичної апроксимації [20].

Проектування на розширюваний ε_t-симплекс $S_{\varepsilon_{t+1}}^N$ забезпечує виконання умови $p_t^i[j]\ge\varepsilon_t$, $j=1..N$, необхідної для повноти статистичної інформації про вибрані чисті стратегії, а параметр $\varepsilon_t\to 0$, $t=1,2,...$ використовується як додатковий елемент керування збіжністю рекурентного методу.

Вибір чистих стратегій a_t^i здійснюється гравцями на основі динамічних випадкових розподілів (6):

$$a_t^i=\left\{A^i(k)\middle| k=\arg\left(\min_k\sum_{j=1}^k p_t^i(a_t^i(j))>\omega\right),\ k=1..N\right\} \quad \forall i\in I, \tag{8}$$

де $\omega\in[0,1]$ – дійсне випадкове число з рівномірним розподілом.

Стохастична гра розпочинається з ненавчених векторів змішаних стратегій зі значеннями елементів $p_0^i(j)=1/N$, де $j=1..N$. У наступні моменти часу динаміка векторів змішаних стратегій визначається марківським рекурентним методом (6) – (8).

У момент часу t кожен гравець на основі змішаної стратегії p_t^i вибирає чисту стратегію a_n^i, за що до моменту часу $t+1$ отримує поточний програш ξ_t^i, після чого обчислює змішану стратегію p_{t+1}^i згідно з (6).

Завдяки динамічній перебудові змішаних стратегій на основі опрацювання поточних програшів метод (6) – (8) забезпечує адаптивний вибір чистих стратегій у часі.

Якість ігрової кластеризації даних оцінюється:

1) функцією середніх втрат:

$$\Xi_t = \frac{1}{L}\sum_{i=1}^{L}\Xi_t^i, \tag{9}$$

де $L = |I|$ – потужність множини гравців;

2) середньою нормою змішаних стратегій гравців:

$$\Delta_t = \frac{1}{tL}\sum_{\tau=1}^{t}\sum_{i=1}^{L}\left\|p_\tau^i\right\|. \tag{10}$$

Алгоритм розв'язування стохастичної гри

1. Задати початкові значення параметрів:

$t = 0$ – початковий момент часу;

$L = |I|$ – кількість гравців;

$X = \{x_1, x_2, ..., x_L\}$ – множина призначених для кластеризації параметрів;

m – кількість вимірів параметрів $x \in R^m$;

N – кількість чистих стратегій гравців (кількість кластерів Y_n, $n = 1..N$);

$A^i = \{a^i(1), a^i(2), ..., a^i(N)\}$, $a^i(j) = j$, $i = 1..L$, $j = 1..N$ – вектори чистих стратегій гравців;

$p_0^i = (1/N, ..., 1/N)$, $i = 1..L$ – початкові змішані стратегії гравців;

$\gamma > 0$ – параметр кроку навчання;

$\alpha \in (0,1]$ – порядок кроку навчання;

ε – параметр ε-симплекса;

$\beta > 0$ – порядок швидкості розширення ε-симплекса;

$d > 0$ – дисперсія завад;

$t_{\max}$ – максимальна кількість кроків методу.

2. Вибрати варіанти дій $a_t^i \in A^i$, $i = 1..L$ згідно з (8).

3. Отримати значення поточних програшів ξ_t^i, $i = 1..L$ згідно з (2). Поточні значення гауссівського білого шуму обчислюються так:

$$\mu_t = \sqrt{d}\left(\sum_{j=1}^{12}\omega_{j,t} - 6\right),$$

де $\omega \in [0,1]$ – дійсне випадкове число з рівномірним законом розподілу.

4. Обчислити значення параметрів γ_t, ε_t згідно з (7).

5. Обчислити елементи векторів змішаних стратегій p_t^i, $i = 1..L$ згідно з (6).

6. Обчислити характеристики якості кластеризації Ξ_t (9), Δ_t (10).

7. Задати наступний момент часу $t := t + 1$.

8. Якщо $t < t_{\max}$, то перейти на крок 2, інакше – кінець алгоритму.

Результати комп'ютерного моделювання

Розв'язування стохастичної гри кластеризації даних виконаємо за допомогою ігрового методу (6) – (8) з параметрами: $m = 2$, $N = 2$, $A^i = (1,2)$, $\gamma = 1$, $\varepsilon = 0.999/N$, $\alpha = 0.01$, $\beta = 2$, $t_{\max} = 10^5$.

Нехай у межах базової множини $X = \{Y_1, Y_2\}$ візуалізуються дві непорожні підмножини $Y_1 \cup Y_2 = X$. Розглянемо такі три варіанти організації множини точок, призначених для кластеризації.

Варіант 1. Підмножини не перетинаються: $Y_1 \cap Y_2 = \varnothing$. Відстань між підмножинами перевищує діаметри підмножин: $S(Y_1, Y_2) > D(Y_1)$, $S(Y_1, Y_2) > D(Y_2)$, де $S(Y_1, Y_2) = \min_{y_1 \in Y_1, y_2 \in Y_2}\left\|y_1 - y_2\right\|$, $D(Z) = \max_{z_1, z_2 \in Z}\left\|z_1 - z_2\right\|$.

Ці умови задовольняє множина $X = \{\{(1,3),(3,1),(3,3)\},\{(7,7),(7,9),(9,7)\}\}$. Підмножини $Y_1 = \{(1,3),(3,1),(3,3)\}$ та $Y_2 = \{(7,7),(7,9),(9,7)\}$ не перетинаються. Застосування методу (6) – (8) забезпечує отримання розв'язків стохастичної гри у чистих стратегіях. Для цього варіанта даних розв'язок гри є таким: $Y_1 = \{(1,3),(3,1),(3,3)\}$, $Y_2 = \{(7,7),(7,9),(9,7)\}$.

На рис. 1 у логарифмічному масштабі зображено графіки функцій середніх програшів гравців Ξ_t та середньої норми змішаних стратегій Δ_t, які характеризують збіжність стохастичної гри кластеризації даних.

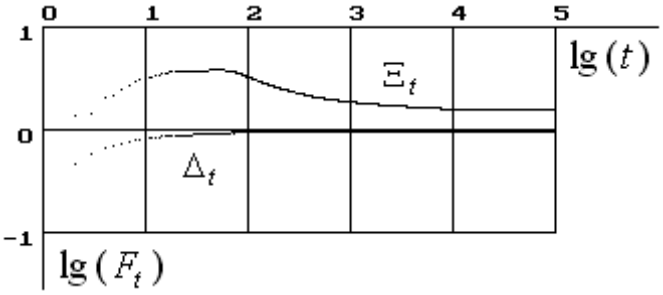

Рис. 1. Характеристики збіжності стохастичної гри

Ігровий метод (6) – (8) забезпечує мінімізацію функції середніх програшів у часі. Функція середньої норми змішаних стратегій прямує до логарифмічного нуля, що ілюструє отримання розв'язків гри у чистих стратегіях.

Порядок швидкості збіжності ігрового методу визначається співвідношенням параметрів α та β. Значення цих параметрів повинні задовольняти базові умови стохастичної апроксимації [20].

Залежність середньої кількості кроків $\overline{t}$ навчання гри від параметра α наведено на рис. 2. Значення $\overline{t}$ усереднено за $k_{\exp} = 100$ реалізаціями випадкових процесів. Момент зупинки гри визначається умовою $\Delta_t \geq 0.99$ наближення середньої норми змішаних стратегій до 1 та правильним віднесенням елементів множини X до одного із кластерів Y_1 або Y_2 (так, як візуалізуються ці кластери у множині X). Результати отримано для значення дисперсії завад $d = 0$.

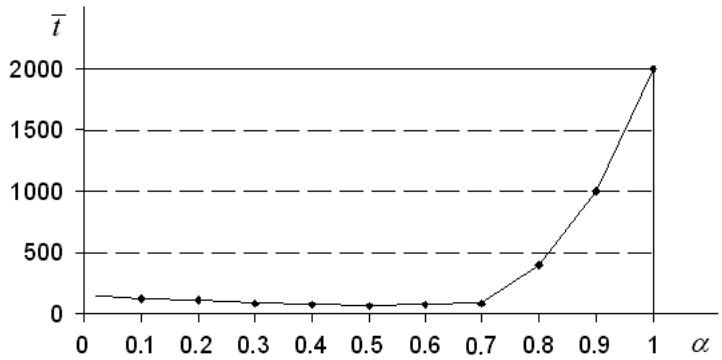

Рис. 2. Вплив параметра α на збіжність гри

Для розв'язуваної задачі зростання значення параметра α від 0 до 0.7 значно не погіршує збіжності стохастичної гри. Значно зростає середня кількість кроків гри при $\alpha > 0.7$.

Досліджено стійкість стохастичної гри при дії завад у вигляді білого шуму. Вплив дисперсії завад d на значення середньої кількості кроків $\overline{t}$ гри кластеризації даних зображено на рис. 3. Результати отримано для значень параметрів $\alpha = 0.3$ та $\beta = 2$.

Значення дисперсії $d \in [0;50]$ значно не впливає на розв'язування задачі кластеризації даних за допомогою ігрового методу (6) – (8). Зростання інтенсивності завад ($d > 50$) призводить до

значного зростання середньої кількості кроків гри, необхідних для правильного віднесення елементів множини X до одного із кластерів Y_1, Y_2 на рівні навчання гри $\Delta_t \geq 0.99$. Встановлені межі зміни дисперсії залежать від абсолютних значень поточних програшів гравців.

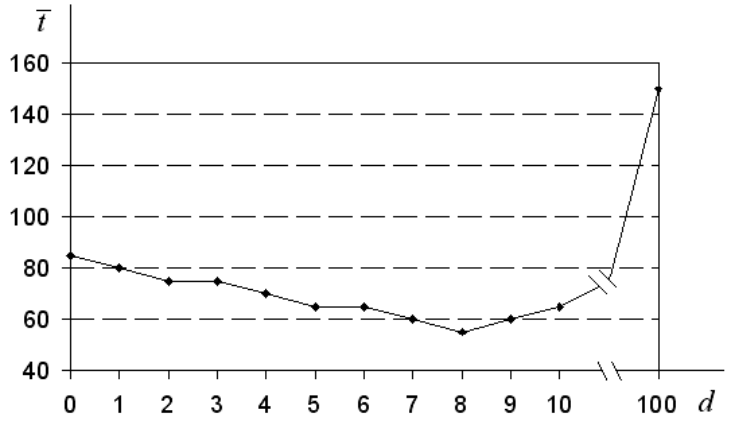

Рис. 3. Вплив дисперсії d на збіжність гри

При зменшенні відстані $S(Y_1, Y_2)$ між підмножинами Y_1 та Y_2 (коли порушуються умови варіанта 1) їхні граничні елементи можуть бути віднесені як до підмножини Y_1, так і до підмножини Y_2.

Варіант 2. Підмножини перетинаються: $Y = Y_1 \cap Y_2 \neq \varnothing$. У спільній підмножині існують точки $y \in Y$, розміщені на однаковій відстані від підмножин $Y_1 - Y$ та $Y_2 - Y$: $|s(y, Y_1 - Y) - s(y, Y_2 - Y)| < \varepsilon$, де $s(y, Z) = \min_{z \in Z} \|y - z\|$.

Ці умови задовольняє множина $X = \{\{(1,3),(3,1),(5,5)\}, \{(5,5),(7,9),(9,7)\}\}$. Точка $(5,5) \in Y$ знаходиться на однаковій відстані від підмножин $Y_1 - Y = \{(1,3),(3,1)\}$ та $Y_2 - Y = \{(7,9),(9,7)\}$ і з однаковою ймовірністю може бути віднесена як до кластера Y_1, так і до кластера Y_2. Для заданих вхідних даних метод (6) – (8) забезпечує розв'язування стохастичної гри у чистих стратегіях. Можливими розв'язками є такі:

1) $Y_1 = \{(1,3),(3,1),(5,5)\}$, $Y_2 = \{(7,9),(9,7)\}$;

2) $Y_1 = \{(1,3),(3,1)\}$, $Y_2 = \{(5,5),(7,9),(9,7)\}$.

Варіант 3. У множині X не візуалізуються підмножини Y_1 та Y_2: $X = Y_1 = Y_2$.

Нехай $X = \{(4,6),(5,5),(6,4)\}$. Цей варіант є частковим випадком варіанта 2. Множина X може бути розділена на кластери згідно із критеріями (4). При $N = 2$ можливими розв'язками є такі:

1) $Y_1 = \{(4,6),(5,5)\}$, $Y_2 = \{(6,4)\}$;

2) $Y_1 = \{(4,6)\}$, $Y_2 = \{(5,5),(6,4)\}$.

У граничних випадках, наприклад, коли $0 < |X| \leq 2$, метод (6) – (8) забезпечує розв'язування ігрової задачі кластеризації даних у змішаних стратегіях. На рис. 4 зображено характеристики збіжності стохастичної гри розділення базової множини $X = \{(4,6),(6,4)\}$ на $N = 2$ кластери. Параметри гри є такими: $\alpha = 0.3$, $\beta = 2$, $d = 0$.

Як видно на рис. 4, функція середньої норми змішаних стратегій Δ_t не досягає значення логарифмічного нуля, що свідчить про отримання розв'язку у змішаних стратегіях. Можливими розв'язками є такі:

1) $Y_1 = \{(4,6),(6,4)\}$, $Y_2 = \varnothing$;

2) $Y_1 = \varnothing$, $Y_2 = \{(4,6),(6,4)\}$;

3) $Y_1 = \{(4,6)\}$, $Y_2 = \{(6,4)\}$;

4) $Y_1 = \{(6,4)\}$, $Y_2 = \{(4,6)\}$.

Рис. 4. Характеристики збіжності стохастичної гри 2×2

Зростання потужності множини X та відповідне зростання кількості гравців призводить до зменшення швидкості збіжності стохастичної гри, яке проявляється зростанням кількості кроків, необхідних для кластеризації даних.

На рис. 5 зображено графік середньої кількості кроків навчання стохастичної гри від кількості вхідних даних. Результати отримано для таких значень параметрів ігрового методу: $\alpha = 0.3$, $\beta = 2$, $d = 0$, $N = 2$. Призначені для кластеризації дані отримано випадково за допомогою нормального закону розподілу координат точок на площині. Згенеровано два скупчення точок з параметрами нормального розподілу $Normal(E\{(5,5)\}, d(9))$ та $Normal(E\{(10,10)\}, d(9))$. Момент $\bar{t}$ завершення гри визначається з умови $\Delta_t \geq 0.99$. Отримані результати усереднено за $k_{\exp} = 100$ експериментами.

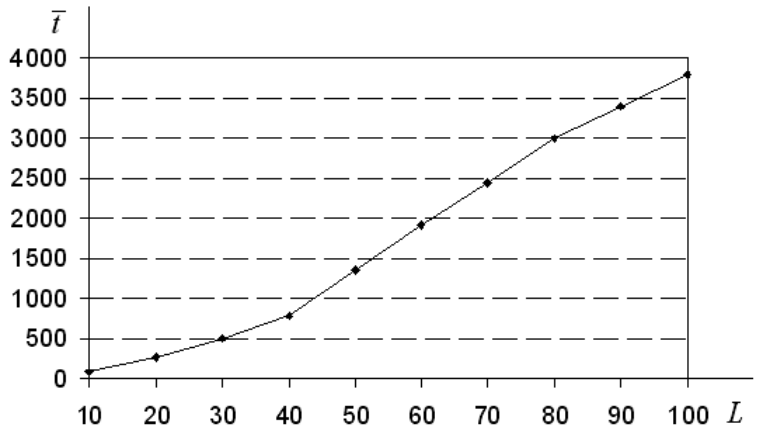

Рис. 5. Залежність середньої кількості кроків гри від кількості точок кластеризації

За результатами експериментів видно, що із збільшенням кількості призначених для кластеризації точок зростає кількість кроків, необхідних для навчання стохастичної гри розділяти дані на кластери.

Досягнення прийнятних для практичних застосувань характеристик збіжності стохастичної гри визначається тонким настроюванням параметрів ігрового методу у рамках базових співвідношень, які надає теорія стохастичної апроксимації [20].

Висновки

У цій статті запропоновано новий метод кластеризації даних, оснований на результатах теорії стохастичних ігор. Розроблений та досліджений ігровий метод (6) – (8) забезпечує розв'язування задачі кластеризації зашумлених даних. Для цього кожна точка множини даних розглядається як окремий гравець з можливістю навчання та адаптації до невизначеностей системи. Чистими стратегіями гравців є вибір одного із фіксованої кількості кластерів. Після завершення вибору кластерів усіма гравцями обчислюють відповідні програші за критеріями мінімізації сумарної відстані між точками кластерів, утворюваних вільним вибором стратегій гравців. Отримані програші гравці використовують для перебудови динамічних векторів змішаних стратегій,

покладених в основу випадкового механізму генерування чистих стратегій гравців. Побудований на основі стохастичної апроксимації метод перебудови змішаних стратегій (6) забезпечує мінімізацію функцій середніх програшів на одиничних симплексах.

Задачі кластеризації даних розв'язуються під час розв'язання стохастичної гри у реальному масштабі часу під час збирання поточної інформації та її адаптивного опрацювання.

Розроблена програмна модель підтверджує збіжність адаптивного ігрового методу (6) – (8) під час розв'язування задачі кластеризації даних. Ефективність ігрового методу оцінено за допомогою характеристичних функцій середніх програшів та середньої норми змішаних стратегій. Збіжність ігрового методу залежить від розмірності стохастичної гри, величини завад та співвідношень параметрів ігрового методу. При зростанні кількості гравців та інтенсивності завад швидкість та ефективність ігрової кластеризації даних зменшуються. Достовірність отриманих результатів підтверджується повторюваністю значень розрахованих характеристик стохастичної гри для різних послідовностей випадкових величин.

Запропонований ігровий метод (6) – (8) кластеризації даних належить до класу методів, що ґрунтуються на опрацюванні реакцій середовища на дії агентів, і має відносно невисоку степеневу швидкість збіжності, що пов'язано з апріорною невизначеністю системи. Збирання інформації здійснюється у процесі навчання шляхом адаптивної перебудови векторів змішаних стратегій пропорційно до значень поточних програшів. Цей недолік долається високою швидкодією сучасних засобів обчислювальної техніки та можливістю розпаралелювання ігрової задачі.

Невирішеним у цій роботі залишилося питання автономного визначення кількості кластерів під час розв'язування стохастичної гри кластеризації зашумлених даних.

1. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с. 2. Mirkin B.G. Clustering for Data Mining. A Data recovery Approach / B.G. Mirkin. – Taylor & francis Group, 2005. – 278 p. 3. Бериков В. С. Современные тенденции в кластерном анализе / В. С. Бериков, Г. С. Лбов // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению "Информационно-телекоммуникационные системы". – 2008. – 26 с. 4. Jain A.K. Data Clustering: A Review / A.K. Jain, M.N. Murty, P.J. Flynn // ACM Computing Surveys. – 1999. – Vol. 31, № 3. – P. 254–323. 5. Хайдуков Д. С. Применение кластерного анализа в государственном управлении / Д. С. Хайдуков // Философия математики: актуальные проблемы. – М.: МАКС Пресс, 2009. – 287 с. 6. Наследов А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках / А.Д. Наследов. – СПб.: Питер, 2004. – 416 с. 7. Клічук О. Особливості інтелектуальних методів кластеризації у реляційних базах даних / О. Клічук // Штучний інтелект. – 2010. – № 1. – С. 25 – 31. 8. Ткаченко О. М. Методи кластеризації даних у цифровій обробці мовленнєвих сигналів / О. М. Ткаченко, О. Д. Феферман // Інформаційні технології та комп'ютерна техніка. Наукові праці ВНТУ. – 2010. – № 1. – С. 1 – 8. 9. Мельник Р. А. Пошук зображень у базах даних за структурними коефіцієнтами на основі алгоритму триступеневої кластеризації / Р. А. Мельник, Р. Б. Тушницький // Вісник Нац. ун-ту "Львівська політехніка": Комп'ютерні системи проектування. Теорія і практика. – 2009. – № 651. – С. 190–196. 10. Кораблев Н.М. Кластеризация данных методом k-means с использованием искусственных имунных систем / Н.М. Кораблев, А.А. Фомичев // Бионика интеллекта. – 2011. – № 3 (77). – С. 102–106. 11. Linde Y. An Algorithm for Vector Quantizer Design / Linde Y., Buzo A., Gray R. // IEEE Transactions on Communications. – 1980. – № 28. – PP. 84 – 94. 12. Говорухін С.О. Кластеризація об'єктів із нечітко заданими значеннями характеристик / С.О. Говорухін // Штучний інтелект. – 2008. – № 4. – С. 567–576. 13. Назаренко М.В. Алгоритм кластеризації на основі нечітких множин / М.В. Назаренко, Л.В. Саричева // Науковий вісник НГУ. – 2011. – № 2. – С. 36 – 39. 14. Алгулиев Р.М. Быстрый генетический алгоритм решения задачи кластеризации текстовых документов / Р.М. Алгулиев, Р.М. Алыгулиев // Искусственный интеллект. – 2005. – № 3. – 698 – 707. 15. Хайкин С. Нейронные сети: полный курс / С. Хайкин. – М.: Вильямс, 2006. – 1104 с. 16. Шафроненко А. Ю. Адаптивна кластеризація даних з пропущеними значеннями / А. Ю. Шафроненко, В. В. Волкова,

Є. В. Бодянський // Радіоелектроніка, інформатика, управління. – 2011. – № 2. – С. 115 – 119. 17. Доманский В.К. Стохастические игры / В.К. Доманский // Математические вопросы кибернетики. – 1988. – № 1. – С. 26 – 49. 18. Мулен Э. Теория игр с примерами из математической экономики / Э. Мулен. – М.: Мир, 1985. – 200 с. 19. Назин А.В. Адаптивный выбор вариантов: Рекуррентные алгоритмы / А.В. Назин, А.С. Позняк. – М.: Наука, 1986. – 288 с. 20. Граничин О.Н. Введение в методы стохастической аппроксимации и оценивания: Учеб. пособие / О.Н. Граничин. – СПб.: Издательство С.-Петербургского университета, 2003. – 131 с.

УДК 004.8

Ю.М. Романишин[1,2], С.Р. Петрицька[1]
[1]Національний університет "Львівська політехніка", кафедра ЕЗІКТ,
[2]University of Warmia and Mazury in Olsztyn, Poland

ПОБУДОВА ЗАДАНОЇ ПОСЛІДОВНОСТІ ІМПУЛЬСІВ НА ОСНОВІ БАГАТОВХОДОВОГО СПАЙК-НЕЙРОНА

Розглянуто задачу побудови ідеалізованої спайк-послідовності імпульсів як зваженої суми вхідних послідовностей, що має значення для процедури навчання спайк-нейронної мережі на основі апарату лінійної алгебри з використанням ідеалізованих імпульсів з нульовою тривалістю і одиничною амплітудою та поняття простору спайк-послідовностей. Для визначення вагових коефіцієнтів використано метод найменших квадратів. Результуюча спайк-послідовність формується з використанням нечітких чисел. Наведено два приклади наближення заданої спайк-послідовності.

Ключові слова: спайк-нейронна мережа, спайк-послідовність, метод лінійної алгебри, спайк-навчання.

The problem of construction of idealizing sequence of spikes as the weighted sum of input sequences of spikes, which is important for learning procedure of spike neural network on the basis of methods of linear algebra with the use of idealizing impulses with a zero duration and single amplitude and conception of space of sequences of spikes is considered. For determination of weight coefficients a least squares method is used. Resultant spike sequence is formed with the use of fuzzy numbers. Two examples of approximation of the sequence of spikes are demonstrated.

Key words: spike neural network, spike sequence, method of linear algebra, spike learning.

Вступ

Спайк-нейронні мережі належать до нейронних мереж третього покоління, в яких інформація кодується часовими проміжками між імпульсами ("спайками") [1, 2]. Внаслідок цього процедури навчання таких мереж суттєво відрізняються від процедур навчання нейронних мереж попередніх поколінь з пороговими та неперервними функціями активації. Серед процедур навчання спайк-нейронних мереж можна назвати [3]: 1) градієнтні методи (SpikeProp), близькі до аналогічних для нейронних мереж з неперервними функціями активації; 2) статистичні методи; 3) методи на основі апарату лінійної алгебри; 4) методи на основі еволюційних стратегій; 5) метод ReSuMe та інші. Однак, незважаючи на значну кількість публікацій за цією тематикою, окремі питання, пов'язані з процедурами навчання спайк-нейронних мереж, досліджено недостатньо.