IEEE Second International Conference on Data Stream Mining & Processing
August 21-25, 2018, Lviv, Ukraine

Game Model for Data Stream Clustering

Petro Kravets
Computer Science Department
Lviv Polytechnic National University
Lviv, Ukraine
Petro.O.Kravets@lpnu.ua

Abstract—**In this article, the stochastic game model for data stream clustering is offered. Players represent numerical values of the clustering data. The essence of the game is that players perform a self-learning random move from one cluster to another in order to minimize the differences between the data of the same cluster. To solve the game, an adaptive recursive method has been developed. Computer modeling confirms the convergence of the game method with certain limitations of its parameters.**

Keywords—data stream clustering; stochastic game model; adaptive game method.

I. Introduction

The clustering is a partition of the set of objects into subsets depending on their similarity. The separated subsets are called as clusters. Elements of one cluster have the general properties. Elements of different clusters considerably differ among themselves.

The clustering is used for problem-solving of the intellectual analysis and visualization of the data, grouping, and recognition of images, extraction of new knowledge and for information search. The clustering purpose consists in the finding of groups of similar objects in the set [1].

The clustering of objects also is used in chemistry, biology, medicine, sociology, pedagogics, psychology, philology, marketing, signal processing, pattern recognition, scientific discipline of documentation, computer science, scientific work and other areas of human activity for data structure in a cluster form for the purpose of their ordering and the group analysis.

The general clustering scheme is such: extraction of characteristics of objects; definition of the metric affinity of objects; partition of a set of objects on clusters; interpretation of clustering results.

Let each object $x \in X$ from a set of objects $X = (x_1, x_2, ..., x_L)$ is described by a vector of properties $x = (x[1], x[2], ..., x[k])$, which can be quantitative or qualitative characteristics of the object.

In problems of data stream clustering vectors of properties change in time: $x = (x_t \,|\, t = 1, 2, ...)$ [2, 3]. As a rule, the law of change of object properties is unknown a priori.

The similarity of two objects x_i also x_j is defined by the metrics of their affinity $\delta(x_i, x_j)$ in space of characteristics. As the metrics, the Euclidean distance, the Tchebyshev distance, the Manhatten distance, the percentage of inconsistency, the Pierce correlation factor and others are used.

Generally, the clustering of objects it is possible to consider as a problem of optimum distribution of objects on groups. The minimization of a root-mean-square error of clusters setting can be the criterion of optimization:

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

where $x_i^{(j)}$ is the point belonging to j -th cluster; $\overline{x}_j$ is the center of the j -th cluster; C_j is a number of elements of the j -th cluster.

Substantial interpretation of the generated clusters for a finding of factors or the reasons of a grouping of objects in clusters is the final stage of clustering. For estimation of the quality of clustering, involve experts from corresponding subject domains.

The data intended for clustering, as a rule, contains uncertainty elements in practical applications. It can be indistinctly specified characteristics of the objects, missed attributes of objects in databases, noisy signals etc. In the uncertainty conditions apply fuzzy clustering, adaptive clustering, genetic algorithms, neural networks without the teacher learning.

The data clustering is formulated as a competitive or cooperative problem of assigning an object to one or another cluster. Problems of a competition and cooperation of objects are studied by the theory of games [4], and in uncertainty conditions, they are studied by the theory of stochastic games [5]. Therefore actual from the scientific, informative and practical points of view there are applications of methods of stochastic games for the data clustering in the conditions of uncertainty.

Construction of a game model of the data clustering with uncertainty elements is the goal of this paper. For purpose achievement it is necessary to solve such problems: to carry out a formulation of a problem game of the data clustering, to develop an adaptive game method and algorithm for solving the problem, to develop computer program model, to analyze and interpretation of the received results.

II. Game Problem Statement

Let $X = \{x_1, x_2, ..., x_L\}$ is the set by coordinates of points $x \in R^k$ in k -dimensional parametrical space. Coordinates of points define the normalized characteristic vector intended

978-1-5386-2874-4/18/$31.00 ©2018 IEEE

for objects clustering. In this set, it is necessary to separate N clusters

$$\left\{ Y_n, n=1..N \left| \begin{array}{l} \bigcup\limits_{n=1..N} Y_n = X, \\ Y_i \bigcap\limits_{i\neq j} Y_j = \varnothing \ \forall (i,j)\in\{1..N\} \end{array} \right. \right\}$$

by criteria

$$\frac{1}{C_n}\sum_{x\in Y_n}\left\|x_i - x_j\right\| \to \min\ ,\quad n=1..N \qquad (1)$$

where $C_n = |Y_n|$ is a quantity of elements which enter into a cluster Y_n; $\|*\|\in R^1$ is the Euclidean distance of a vector.

Let parameters of objects are vector random variables with stationary normal distribution:

$$x \sim Normal(m_x, d_x) \in R^k\ ,$$

where m_x is an expectation value; d_x is a dispersion.

The separation of clusters Y_n ($n=1..N$) in the set X will be done using the stochastic game method described by the tuple $(I, A^i, \Xi^i \,|\, i\in I)$, where I is a set of players; $L = |I|$ is a quantity of players; $A^i = \{a^i[1],...,a^i[N]\}$ is a set of pure strategies of the i-th player which define a choice of one of clusters; N is a quantity of strategies of the i-th player ($N < L$); $\Xi^i : A \to R^1$ is a lose function of the i-th player; $A = \underset{i\in I}{\times} A^i$ is a set of the combined strategies.

The game essence consists in the random moving of players from one cluster to another. For this purpose during time moments $t = 1,2,...$, each player on the basis of the generator of random events independently of others chooses a pure strategy $a^i \in A^i$ which defines its accessory to the corresponding cluster. According to (1), after the realization of the combined variant $a\in A$, players receive random losses $\xi^i(a)$ with a priori unknown stochastic characteristics:

$$\xi_t^i = \frac{1}{C_t^i}\sum_{j\in I}\chi\left(a_t^i = a_t^j\right)\left\|x_i - x_j\right\| \quad \forall i\in I\ , \qquad (2)$$

where $C_t^i = \sum_{j\in I}\chi\left(a_t^i = a_t^j\right)$ is a current quantity of elements of a cluster which contains the i-th player; $\chi(*)\in\{0,1\}$ is an indicator of the event.

The efficiency of a game course is defined by functions of average losses:

$$\Xi_t^i = \frac{1}{t}\sum_{\tau=1}^{t}\xi_\tau^i \qquad \forall i\in I\ . \qquad (3)$$

The game purpose consists in minimization of the system of functions of average losses (3) in time:

$$\overline{\lim_{t\to\infty}}\,\Xi_t^i \to \min \qquad \forall i\in I\ . \qquad (4)$$

So, on the basis of a supervision of current losses $\{\xi_n^i\}$ each player $i\in I$ should learn to choose pure strategy $\{a_t^i\}$ so that with time course $t = 1,2,...$ to provide the performance of criteria system (4).

The game problem solutions will satisfy one of the conditions of collective balance, for example, on Nash or Pareto, depending on a method of formation of a sequence of strategies $\{a_t^i\}\forall i\in I$.

III. Method of Problem Solving

Stochastic game solving we will execute by means of adaptive recurrent transformation of vectors $p_t^i\ \forall i\in I$ of the mixed strategies.

Construction of a method of stochastic game solving we will carry out on the basis of stochastic approximation of a complementary slackness condition of a determined game, correct for the mixed strategies in a balance point on Nash [5].

For this purpose, we will define a polylinear function of average losses for the determined game:

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

where $v(a) = M\{\xi_t^i(a)\}$.

Then the vector of a complementary slackness condition (CS) will be of the form:

$$\overset{\rightarrow}{\mathrm{CS}}{}^i = \nabla_{p^i} V^i(p) - e^N V^i(p) = 0 \qquad \forall i\in I\ ,$$

where $\nabla_{p^i} V^i(p)$ is a gradient of the polylinear function of average losses; $e^N = (1_j \,|\, j=1..N)$ is a vector whose all components are equal to 1; $p\in S^M$ is the combined mixed strategy of players set on a convex unit simplex S^M ($M = N^L$).

To take account of the solutions in vertices of the unit simplex we will execute weighing of a CS^i-vector by elements of a vector p^i of the mixed strategies:

$$diag(p^i)(\overset{\rightarrow}{CS}{}^i) = 0 \quad \forall i\in I\ , \qquad (5)$$

where $diag(p^i)$ it is the square diagonal matrix of an order N constructed of elements of a vector p^i.

Considering that

$$diag(p^i)[\nabla_{p^i} V^i - e^N V^i] = \\ = E\{\xi_t^i [e(a_t^{\ i}) - p_t^i] \mid p_t^i = p^i\},$$

on the basis of a method of stochastic approximation [6] we will receive recurrent expression:

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

where E is an expectation symbol; $\pi_{\varepsilon_{t+1}}^N$ is a projector on N -dimensional ε_t -simplex $S_{\varepsilon_{t+1}}^N$ [5]; $\gamma_t > 0$, and $\varepsilon_t > 0$ are monotonously descending sequences of positive values; $e(a_t^{\ i})$ is the unit vector specifying in a choice of pure strategy $a_t^{\ i} = a^i \in A^i$.

Parameters γ_t and ε_t can be calculated as follows:

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

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

Convergence of strategies (6) to optimum values with probability 1 and in the root-mean-square is defined by the ratio of parameters γ_t and ε_t which should satisfy fundamental conditions of stochastic approximation [6].

Projection on expanded an ε_t -simplex $S_{\varepsilon_{t+1}}^N$ provides the performance of the condition $p_t^i[j] \ge \varepsilon_t$, $j = 1..N$ necessary for completeness of the statistical information on chosen pure strategies, and the parameter $\varepsilon_t \to 0$ is used as an additional element for controlling the convergence of the recurrent method.

The choice of pure strategy $a_t^i[k]$ $\forall i \in I$ is carried out by players on the basis of dynamic random distributions (6):

$$k = \arg\left(\min_{k=1..N} \sum_{j=1}^{k} p_t^i(a_t^{\ i}(j)) > \omega \right) \in \{1..N\}, \qquad (8)$$

where $\omega \in [0, 1]$ it is the real random number with the uniform distribution law.

The stochastic game begins from not learned vectors of the mixed strategies with a value of elements $p_0^i(j) = 1/N$, where $j = 1..N$. During following moments of time the dynamics of vectors of the mixed strategies are defined by a Markovian recurrent method (6) – (8).

So, at the moment of time t each player on the basis of the mixed strategy p_t^i chooses a pure strategy a_n^i and until the moment of time $t+1$ receives current loss ξ_t^i then calculates the mixed strategy p_{t+1}^i according to (6).

Thanks to the dynamic reorganization of the mixed strategies based on the processing of current losses, the method (6) – (8) provide an adaptive choice of pure strategies in time.

Quality of game of the data clustering is estimated by:

1) the average loss function:

$$\Xi_t = \frac{1}{L} \sum_{i=1}^{L} \Xi_t^i, \qquad (9)$$

where $L = |I|$ is a cardinality of a set of players;

2) the function of the average norm of mixed player strategies:

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

The algorithm of Stochastic Game Solving

1. To set initial values of parameters: $t = 0$ is an initial moment of time; N is a quantity of pure strategies of players (otherwise it is a number of clusters Y_n, $n = 1..N$); $L = |I|$ is a quantity of players; $X = \{x_1, x_2, ..., x_L\}$ is a set of objects intended for clustering; k is a quantity of characteristic factors of objects $x \in R^k$; $m_x = (m_x[1], m_x[2], ..., m_x[k])$ is an expectation value of parameters of object $x \in X$; $d_x = (d_x[1], d_x[2], ..., d_x[k])$ is a dispersion of parameters of object $x \in X$; $A^i = \{a^i[1], a^i[2], \ \ ..., a^i[N]\}$, $a^i(j) = j$, $i = 1..L$, $j = 1..N$ is a vectors of pure strategies of players; $p_0^i = (1/N, ..., 1/N)$, $i = 1..L$ is an initial mixed strategies of players; $\gamma > 0$ is a parameter of a step of learning; $\alpha \in (0,1]$ is an order of a step of learning; ε is an ε -simplex parameter; $\beta > 0$ is an order of an ε -simplex expansion rate; $t_{\max}$ is a maximum quantity of steps of a method.

2. To choice variants of actions $a_t^{\ i} \in A^i$ of players $i = 1..L$ according to (8).

3. Get current property values of objects as random variables with the normal distribution law:

$$x_t = m_x + \sqrt{d_x} \left(\sum_{j=1}^{12} \omega_{j,t} - 6 \right),$$

where $\omega_{j,t} \in [0, 1]$ it is the real random number with the uniform distribution law.

4. To calculate the value of current losses ξ_t^i, $i = 1..L$ according to (2).

5. To calculate the value of parameters γ_t and ε_t, according to (7).

6. To calculate elements of vectors of the mixed strategies p_t^i, $i = 1..L$ according to (6).

7. To calculate quality characteristics Ξ_t (9) and Δ_t (10) of the data clustering.

8. To set the following moment of time $t := t + 1$.

9. If $t < t_{\max}$ then go to a step 2, else to stop.

IV. Results of Computer Modelling

We will solve a stochastic game by means of a recurrent method (6) – (8) for test parameters: $k = 2$, $N = 2$, $A^i = \{1, 2\}$, $\gamma = 1$, $\varepsilon = 0.999 / N$, $\alpha = 0.3$, $\beta = 2$, $t_{\max} = 10^5$, $d = 0.01$.

Let in the base set $X = \{Y_1, Y_2\}$ two non-empty subsets $Y_1 \cap Y_2 = \varnothing$ are visualized such that intracluster distances are less than intercluster distances. Elements of these subsets are received as the random points generated on a plane on the normal distribution law for different mathematical expectations.

On Fig. 1 graphs of functions Ξ_t of average losses of players and average norm Δ_t of the mixed strategies which characterize the convergence of stochastic game of data clustering are represented in logarithmic scale.

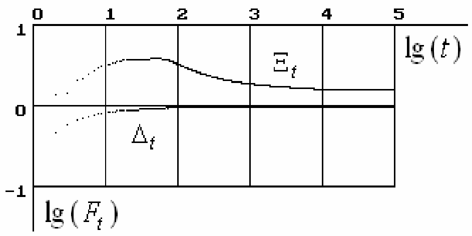

Fig. 1. Characteristics of solving the stochastic game in pure strategies

The game method (6) – (8) provides minimization of the function Δ_t of average losses in time. The function of the average norm of mixed strategies reaches the logarithmic zero, which illustrates the obtaining of the game's solving in pure strategies.

Dependence of average quantity of game learning steps $\overline{t}$ on the parameter α is shown on Fig. 2. Value $\overline{t}$ is averaged on realizations of random processes.

The moment of a game stop is defined by a condition of the approach of the average norm of mixed strategies to 1 ($\Delta_t \geq 0.99$) and correct assignment of elements of the set X to one of the clusters Y_1 or Y_2 (how these clusters are visualized in the set X).

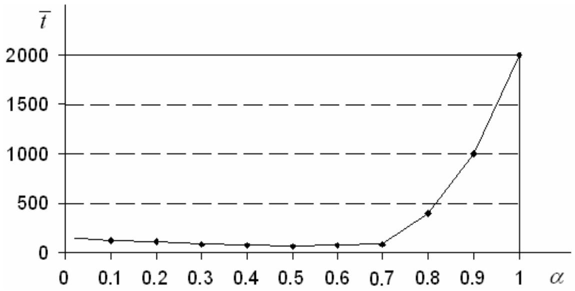

Fig. 2. Influence of the parameter α on the game convergence

For a solved problem, the growth of parameter α from 0 to 0.7 does not lead to considerable deterioration of stochastic game convergence. Considerable growth of the average quantity of game steps occurs at $\alpha > 0.7$.

The order of convergence rate of a game method is defined by a parity of parameters α and β. For the convergence of the offered method, it is necessary that these parameters satisfy the conditions of stochastic approximation [6]. Dependence of average quantity of steps $\overline{t}$ of clustering game from a dispersion d_x of parameters of objects $x \in X$ it is representing by the diagram on Fig. 3.

Value of a dispersion $d_x \in [0; 50]$ does not a material effect on the quantity of the steps necessary for the data clustering by means of a game method (6) – (8). For values $d_x > 50$ of a dispersion, considerable growth of the average quantity of game steps necessary for correct adding of elements of the set X to one of the clusters Y_1 or Y_2 at the level $\Delta_t \geq 0.99$ of game learning is observed.

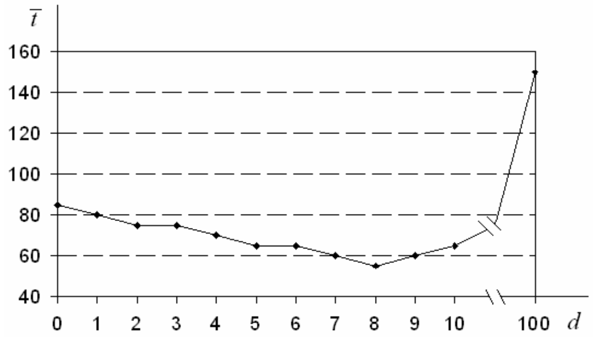

Fig. 3. Influence of the dispersion on the game convergence

The boundary elements of the subsets can be assigned to both cluster Y_1 and cluster Y_2, that is, clusters can intersect.

Let in the general set X there are the points $y \in Y$ placed on equally spaced from subsets $Y_1 - Y$ and, $Y_2 - Y$ that is, $| s(y, Y_1 - Y) - s(y, Y_2 - Y) | < \varepsilon$, where $s(y, Z) = \min\limits_{z \in Z} \| y - z \|$. Then the method (6) – (8) provides solving the game in mixed strategies as shown on Fig. 4.

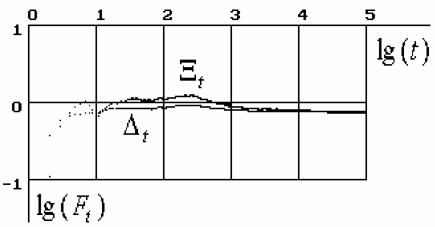

Fig. 4. Characteristics of solving the stochastic game in mixed strategies

On Fig. 4 it is visible that the function Δ_t of the average norm does not reach the logarithmic zero indicating that the game is solved in mixed strategies.

The growth of cardinality of a set X and corresponding growth of the number of players leads to a reduction of convergence rate of the stochastic game, which appears in the growth of the quantity of the steps necessary for the data clustering.

On Fig. 5 the graph of the dependence of average quantity of steps of stochastic game learning on the number of clustering objects is represented. The data intended for clustering is received randomly by means of the normal distribution law of coordinates of points on a plane. It is generated two concentrations of points with parameters of the normal distribution. The moment $\overline{t}$ of the game termination is defined by a condition $\Delta_t \geq 0.99$. The obtained results are averaged on $k_{\exp} = 100$ experiments.

By results of experiments, it is visible that with an increase in the quantity of clustering objects the quantity of the steps necessary for stochastic game learning increases.

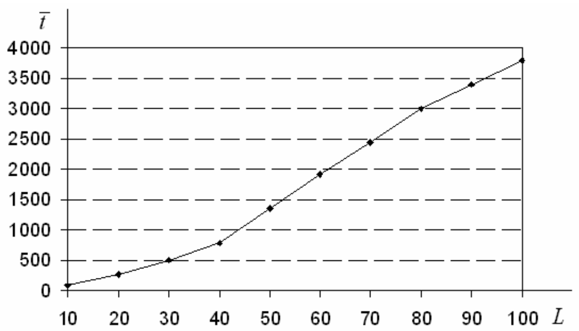

Fig. 5. Dependence of the average quantity of game steps on the number of clustering points

Achievement of the characteristics of the stochastic game convergence, which is acceptable in practice, is determined by fine-tuning of the parameters of the game method within the framework of the basic relations given by the theory of stochastic approximation.

V. Conclusions

In this paper, the new game model for data stream clustering is proposed. An adaptive recursive method was constructed to solve the game. Random moving of points on a plane simulates data streams.

Convergence of a game method depends on the dimension of the stochastic game, the intensity of the noise and the parity of parameters of the game method.

The efficiency of the game of data clustering decreases at an increase of the number of players and noise intensity.

Simulation veracity proves repeatability of values of average characteristics of the stochastic game obtained for various realizations of random variables.

The offered game method of the data clustering belongs to a class of methods, which are based on the processing of reactions of the environment. This method has a relatively small (power-law) order of convergence rate due to the a priori uncertainty of the system.

This limitation can be overcome by the high performance of modern computer and possibility a game problem parallelization.

References

[1] A. Jain, M. Murty, and P. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, vol 31, no. 3, pp. 264-323, September 1999.

[2] D. Barbara, “Requirements for clustering data streams”, ACM SIGKDD Explorations Newsletter, vol. 3, №. 2, pp. 23-27, 2003.

[3] J. Chandrika, and K.R. Ananda Kumar, “Dynamic Clustering Of High-Speed Data Streams”, International Journal of Computer Science Issues, vol. 9, iss. 2, №. 1, pp. 224-228, 2012.

[4] T. Roughgarden, E. Tardos and V. V. Vazirani. Algorithmic Game Theory, edited by Noam Nisan, Cambridge University Press, 2007.

[5] A. Nazin, and A. Poznyak, Adaptive Choice of Variants, Moscow, Nauka, 1986 (in Russian).

[6] H. J. Kushner, G. George Yin, Stochastic Approximation and Recursive Algorithms and Applications. New York: Springer Verlag, 2003.