მონაცემთა კლასტერიზაცია ნაწილაკთა გროვის მეთოდის გამოყენებით

პეტრე პეტაშვილი
საქართველოს ტექნიკური უნივერსიტეტი

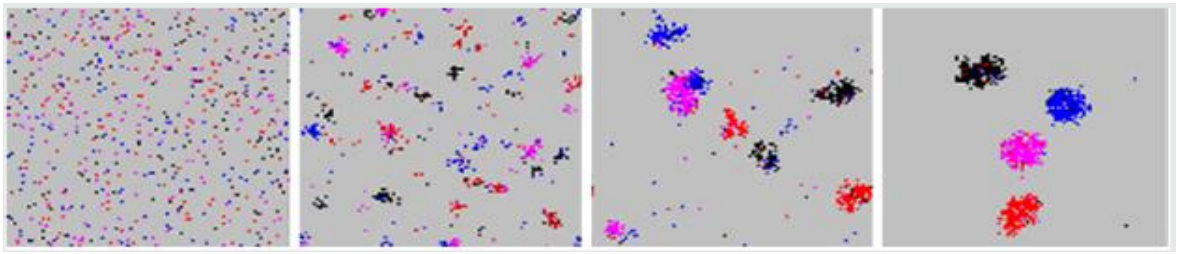
რეზიუმე

მონაცემთა კლასტერიზაცია მნიშვნელოვან როლს თამაშობს მონაცემთა მოპოვებისა და დამუშავების, დიდი მოცულობის მონაცემთა ინტელექტუალური ანალიზის, აგრეთვე მრავალ-აგენტური მოდელირებისა და ოპტიმიზაციის ამოცანების გადაწყვეტის დროს. დღეისათვის შემუშავებულია მონაცემთა კლასტერიზაციის ალგორითმების მთელი კლასი. თუმცა ბოლო პერიოდში, კლასტერიზაციის თვალსაზრისით, პერსპექტიულ და საინტერესო მიმართულებად მიიჩნევა ე.წ. ბიო-ინსპირირებული ოჯახის ალგორითმები, ცნობილი როგორც „გროვის ინტელექტი" (Swarm Intelligence). სტატიაში განხილულია მონაცემების კლასტერიზაციისადმი ახლებური მიდგომა, რომელიც ეფუძნება ნაწილაკთა გროვის ინტელექტის მეთოდებს, რომლის გამოყენება ხდება მრავალკრიტერიუმიანი ოპტიმიზაციის ამოცანების გადაჭრის უწყვეტ პროცესში. კვლევის შედეგებისა და ალგორითმის ეფექტურობის თვალსაჩინოებისათვის გამოყენებულ იქნა კომპიუტერული სიმულაცია.

საკვანძო სიტყვები: მონაცემთა კლასტერიზაცია. მონაცემთა ინტელექტუალური ანალიზი. კოლექტიური ინტელექტი. ნაწილაკების გროვის ოპტიმიზაცია.

I. შესავალი

კლასტერიზაცია არის მონაცემების დაჯგუფება ობიექტების მსგავსობის მიხედვით, რომლის ევოლუციური პროცესი წარმოდგენილია 1-ელ ნახაზზე. თითოეული ჯგუფი, ანუ კლასტერი, შეიცავს მსგავს ობიექტებს ჯგუფის შიგნით და განსხვავებულ ობიექტებს სხვა ჯგუფებისგან.



ნახ.1

ბოლო ათწლეულების განმავლობაში კლასტერული ანალიზის როლის მნიშვნელობა იზრდება სხვადასხვა სფეროში, ინჟინერია (მანქანათა შემეცნება, ხელოვნური ინტელექტი, გამოსახულების ამოცნობა), კომპიუტერული მეცნიერებები (ინფორმაციის ძიება ინტერნეტში, ტექსტური და გრაფიკული მონაცემების მოძიება-ფრაგმენტაცია), მედიცინა, საბუნებისმეტყველო და სოციალურ მეცნიერებებში. კლასტერიზაციის ამოცანები ასევე აქტუალურია სტატისტიკაში, გრაფთა თეორიაში, ხელოვნურ ნეირონულ ქსელებში, ევოლუციურ გამოთვლებსა და სხვა ოპტიმიზაციის ამოცანებში [1].

ინფორმაციის მოძიება, იგივე მონაცემთა ინტელექტუალური ანალიზი, არის ახალი მძლავრი ტექნოლოგია, რომელიც მიმართულია დიდი მონაცემთა ბაზებიდან მნიშვნელოვანი ინფორმაციის მოპოვებისკენ. ასეთი ინსტრუმენტები გამოიყენება მომავალის ტენდენციის და ქცევის პროგნოზირებისთვის და სწორი გადაწყვეტილების მიღების ხელშესაწყობად. მნიშვნელოვანი ინფორმაციის მოპოვება დიდი ბაზებიდან, საჭიროებს მრავალფეროვანი მონაცემების სწრაფ და

ავტომატიზებულ კლასტერიზაციის მექანიზმს. კლასიკური კლასტერიზაციის მეთოდებით კი ამ დონის ოპტიმიზაცია შეუძლებელია. ბიო-ინსპირირებული ოჯახის ალგორითმებმა, კერძოდ კოლექტიურ ინტელექტზე დაფუძნებულმა მეთოდებმა უკეთესი შედეგი აჩვენეს ბევრ კლასიკურ მეთოდებთან შედარებით.

2. კოლექტიური ინტელექტის მეთოდების მოკლე მიმოხილვა

კოლექტიური ინტელექტის ალგორითმები დაფუძნებულია ბიოლოგიურ არსებათა ჯგუფურ ქმედებაზე. ისინი, მიუხედავად ინდივიდების შეზღუდული შესაძლებლობებისა, ახერხებენ კოლექტიური ქმედებით გადაჭრან ბევრი კომპლექსური ამოცანა. ამის ნათელი მაგალითია ჭიანჭველების კოლონის ოპტიმიზაციის მეთოდი, რომელიც გამოიყენება NP-კლასის დისკრეტული ოპტიმიზაციის ამოცანებში და ნაწილაკების გროვის ოპტიმიზაციის მეთოდი, საძიებო არის გლობალური ოპტიმუმების დასადგენად [2,3].

თვალსაჩინოების მიზნით, განვიხილოთ ნაწილაკების გროვის ოპტიმიზაციის მეთოდი (PSO). გროვის ინტელექტის ოპტიმიზაციის მეთოდი დაფუძნებულია კოლექტივის სოციალურ ქცევაზე. იგი კონცეპტუალურად ძალზე მარტივია, რადგან იყენებს მხოლოდ მარტივ არითმეტიკული ოპერაციებს საძიებო არეში ოპტიმუმების დასადგენად [4,5].

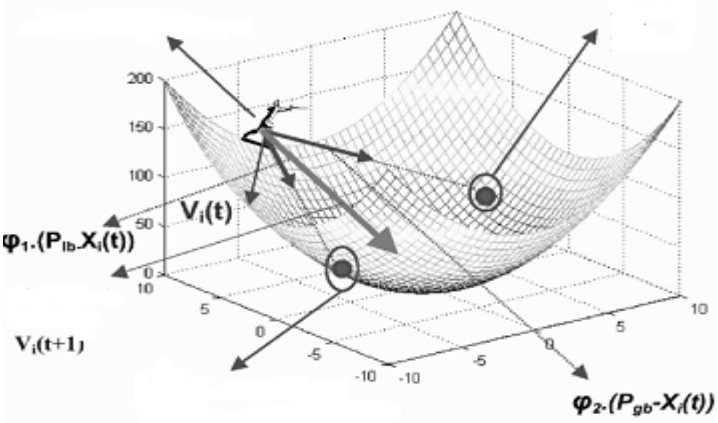
პოპულაციის ინიციალიზაცია PSO-ში ხდება თითოეული ინდივიდის (ნაწილაკის) შემთხვევითი X_i პოზოცოოს და V_i სიჩქარის არჩევით. ხოლო f ფუნქცია გამოითვლება აღნიშნული პარამეტრების საფუძველზე n-განზომილებიან საძიებო არეში $X_i = (x_{i1}, x_{i2}, x_{i3}, \dots x_{in})$ და $V_i = (v_{i1}, v_{i2}, v_{i3}, \dots v_{in})$.

ყოველ ბიჯზე ნაწილაკების პოზიციები და სიჩქარეები კორექტირდება და ხელახლა გამოითვლება ფიტნეს ფუნქცია. მარტივ განახლების განტოლებას, i-ური ნაწილაკის d-ური განზომილებისთვის აქვს შემდეგი სახე:

$$V_{id}(t+1) = \omega . V_{id}(t) + C_1 . \varphi_1 . \left(P_{lid} - X_{id}(t)\right) + C_2 . \varphi_2 . \left(P_{gd} - X_{id}(t)\right) \quad (1)$$

$$X_{id}(t+1) = X_{id}(t) + V_{id}(t+1) \quad (2)$$

სადაც φ_1 და φ_2 არის შემთხვევითი დადებითი რიცხვები, C_1 და C_2 აჩქარების კონსტანტებია, ხოლო ω არის ინერცია. P_{li} - არის i-ური ნაწილაკის ლოკალური საუკეთესო მნიშვნელობა, ხოლო P_g - გროვის გლობალური საუკეთესო მნიშვნელობა. სიჩქარის განახლება იტერაციებს შორის ილუსტრირებულია მე-2 ნახაზზე. ალგორითმის მიხედვით, იტერაციების დამთავრების შემდეგ, ნაწილაკების უმეტესობა მოექცევა საძიებო არის გლობალური ოპტიმუმის ახლო რადიუსში.

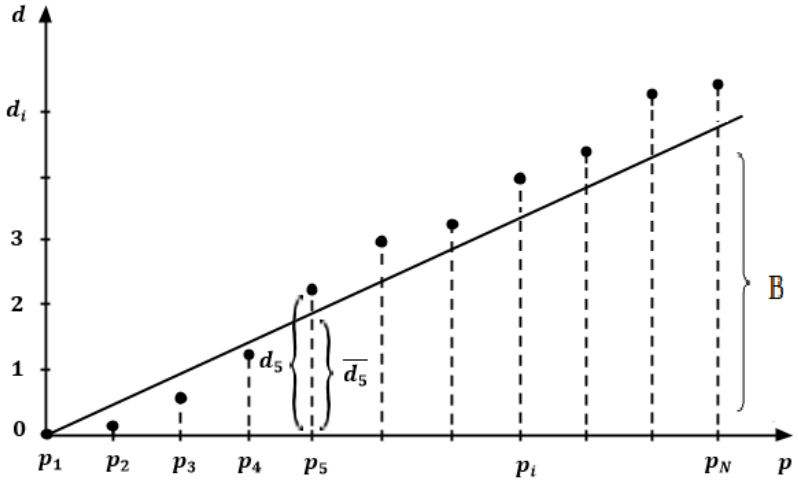


ნახ.2

3. კლასტერიზაციის ახლებური მიდგომა

მრავალკრიტერიუმიანი ანუ ლოკალური ოპტიმიზაციის ამოცანების გადაწყვეტის პროცესში აუცილებელი ხდება ნაწილაკების (აგენტების) კლასტერიზაცია, რისთვისაც ხდება თითოეული ნაწილაკისთვის საწყის კოორდინატებში ფიტნეს ფუნქციის გამოთვლა. ფიტნეს ფუნქციის საფუძველზე, ანუ საუკეთესო პოზიციებზე ირჩევა M რაოდენობის ლიდერი ნაწილაკები, დანარჩენი ნაწილაკები კი ავტომატურად ხდება აუთსაიდერები (ნახ.3).

$$l_j = \{p_k^i\}, \quad j=1,2, \ldots M. \tag{3}$$



ნახ.3. ლიდერების არჩევა

ლიდერების არჩევა შემდეგნაირად ხდება: იგება გრაფიკი, სადაც აბსცისთა ღერძზე განლაგდება ნაწილაკები (p) დალაგებული ფიტნეს-ფუნქცის დონის მიხედვით კლებადობით, ხოლო ორდინატთა ღერძზე გადაიზომება ფიტნეს-ფუნქცის დონეებს შორის დისტანციები (d). ვიმახსოვრებთ თითოეული კოორდინატს (p, d), რომელიც შეესაბამება p ნაწილაკს d ფიტნეს-ფუნქცის დონის სხვაობით. რა თქმა უნდა, კოორდინატთა სათავეში მოხვდება p_1 ნაწილაკი ფიტნეს-ფუნქცის მაქსიმალური დონით.

$$r_1 = r_{max}; \quad r_i = f(p_i), \quad i=1,2, \ldots N. \tag{3.2}$$

$$d_i = r_{max} - r_i \tag{3.3}$$

$$\overline{d} = \frac{\sum_{i=1}^{N} d_i}{N}; \quad \overline{d_i} = \frac{\sum d_i}{i} \tag{3.4}$$

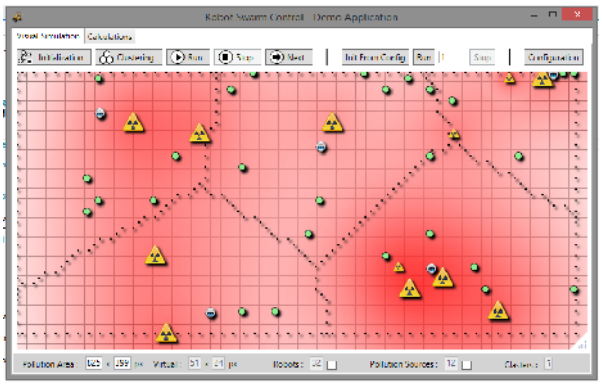
ამის შემდეგ კოორდინატთა სათავიდან აიგება წრფე შემდეგი ფორმულით:

$$\tan \alpha = \frac{\sum_{i=1}^{N} r_i}{N(r_{max} - r_N)} \tag{3.5}$$

წერტილები, რომლებიც მოხვდება აღნიშნული წრფის ქვემოთ გახდება ლიდერები, ხოლო ზემოთ აუთსაიდერები:

$$p_i \equiv \begin{cases} d_i \le \overline{d_i} & - \ leader \\ otherwise & - outsider \end{cases} \tag{3.6}$$

ლიდერების მიხედვით K-Means ალგორითმის გამოყენებით ხდება მოცემული არის კლასტერიზაცია (ნახ.4). თითოეულ კლასტერში ნაწილაკებს შორის ურთიერთობის ფორმა „star“ - ტოპოლოგიით არის განსაზღვრული. ანუ თითოეულ აუთსაიდერ ნაწილაკს კავშირი აქვს მხოლოდ თავის შიდაკლასტერულ ლიდერ ნაწილაკთან.



ნახ.4

K-Means საშუალებას იძლევა გადავანაწილოთ N აუთსაიდერი M ლიდერების სიმრავლეზე $L = \{l_r\}$, r=1,2, …M, იგი ცდილობს მინიმუმამდე დაიყვანოს კლასტერის წერტილების საერთო კვადრატული გადახრა კლასტერის ცენტრიდან, რომელიც ჩვენს შემთხვევაში ლიდერ ნაწილაკს შეესაბამება:

$$\underset{L}{\operatorname{argmin}} = \sum_{l_c=1}^{M} \sum_{p_k^i \in l_r} \left\| p_k^i - p_k^j \right\|^2 \tag{3.7}$$

4. დასკვნა

ამოცანა მდგომარეობს იმაში, რომ ჩვენ შევძლოთ გარემოს დაბინძურების, განსაკუთრებით ატომური ენერგეტიკით, მონიტორინგი, რაც მიიღწევა მობილური, უკაბელო სენსორული ქსელიდან ინფორმაციის უწყვეტად მიღებით და ამ ინფორმაციის შეგროვება-დამუშავებით. ამ საქმეში ჩვენი მთავარი ინსტრუმენტი არის მულტი-რობოტული სისტემა. ჩვენ განვსაზღვრეთ ამ სისტემის პარამეტრები და მისი მართვის სტრატეგიები. ამისთვის კი ამოსავალ წერტილად ავიღეთ ნაწილაკების გროვის ოპტიმიზაციის მეთოდები. ამრიგად, დასმული ამოცანის გადასაჭრელად ჩვენ ვმუშაობთ PSO-ზე დაფუძნებულ ადაპტურ ალგორითმზე, რომელიც გარემო პირობების შეცვლის შემდეგაც კი შეძლებს სწრაფად მოძებნოს ოპტიმალური შედეგი.

ლიტერატურა:

1. Evangelou I.E., Hadjimitsis D.G., Lazakidou A.A., Clayton C. (2001). Data Mining and Knowledge Discovery in Complex Image Data using Artificial Neural Networks, Workshop on Complex Reasoning an Geographical Data, Cyprus.

2. Lumer E., Faieta B. (1994). Diversity and Adaptation in Populations of Clustering Ants. In Proceedings Third Intern. Conf. on Simulation of Adaptive Behavior: from animals to animates 3, Cambridge, Massachusetts MIT press, pp. 499-508.

3. Eberhart R.C., Shi Y. (2001). Particle swarm optimization: Developments, applications and resources, In Proceedings of IEEE Intern. Conf. on Evolutionary Computation, vol.1, pp. 81-86.

4. Ahmed M.N., Yaman S.M., Mohamed N., Farag A.A., Moriarty T.A. (2002). Modifed fuzzy c-means algorithm for bias estimation and segmentation of MRI data. IEEE Trans Med Imaging, 21, pp. 193-199.

5. Kennedy J., Eberhart R.C. (1997). A discrete binary version of the particle swarm algorithm, Proceedings of the Conf. on Systems, Man and Cybernetics, IEEE Service Center, Piscataway, NJ, pp. 4104-4109.

DATA CLUSTERING USING PARTICLE SWARM METHOD

Petre Petashvili
Georgian Technical University

Summary

Clustering aims at representing large datasets by a fewer number of prototypes or clusters. It brings simplicity in modeling data and thus plays a central role in the process of knowledge discovery and data mining. Data mining tasks, in these days, require fast and accurate partitioning of huge datasets, which may come with a variety of attributes or features. This, in turn, imposes severe computational requirements on the relevant clustering techniques. A family of bio-inspired algorithms, well-known as Swarm Intelligence (SI) has recently emerged that meets these requirements and has successfully been applied to a number of real world clustering problems. This paper explores the role of SI in clustering different kinds of datasets. It finally describes a new SI technique for partitioning any dataset into an optimal number of groups through one run of optimization. Computer simulations undertaken in this research have also been provided to demonstrate the effectiveness of the proposed algorithm.

КЛАСТЕРИЗАЦИЯ ДАННЫХ С ПРИМЕНЕНИЕМ МЕТОДА РОЯ ЧАСТИЦ

Петре Петашвили
Грузинский Технический Университет

Резюме

Кластеризация данных играет большую роль при решении задач добычи и обработки данных, интеллектуального анализа большого объема данных, а также мультиагентного моделирования и оптимизации. На сегодняшный день разработан целый класс алгоритмов кластеризации данных. Хотя в последнее время с точки зрения кластеризации перспективным и весьма интересным направлением считается т.н. группа био-инспирированных алгоритмов, известная как интеллект роя (Swarm Intelligence). В статье рассматривается новый подход к кластеризации данных, основанный на методах роя частиц, которые успешно применяются для решения задач многокритериальной оптимизации. Для наглядности результатов решения и эффективности разработанного алгоритма применена компьютерная симуляция.