

UDC 008

SCOPUS CODE 1702

<https://doi.org/10.36073/1512-0996-2025-2-133-141>

ევოლუციური მანქანური სწავლება მონაცემთა ტბების სერვისებში

- ბადრი მეფარიშვილი** საინფორმაციო ტექნოლოგიების დეპარტამენტი, საქართველოს ტექნიკური უნივერსიტეტი, საქართველო, 0160, თბილისი, მ. კოსტავას 77
E-mail: meparishvilibadri08@gtu.ge
- გულნარა ჯანელიძე** საინფორმაციო ტექნოლოგიების დეპარტამენტი, საქართველოს ტექნიკური უნივერსიტეტი, საქართველო, 0160, თბილისი, მ. კოსტავას 77
E-mail: janelidzegulnara08@gtu.ge
- გიორგი მურადოვი** საინფორმაციო ტექნოლოგიების დეპარტამენტი, საქართველოს ტექნიკური უნივერსიტეტი, საქართველო, 0160, თბილისი, მ. კოსტავას 77
E-mail: muradovi.giorgi22@gtu.ge

რეცენზენტები:

მ. ჩხაიძე, სტუ-ის ინფორმატიკისა და მართვის სისტემების ფაკულტეტის პროფესორი

E-mail: m.chkhaidze@gtu.ge

ა. კობიაშვილი, სტუ-ის ინფორმატიკისა და მართვის სისტემების ფაკულტეტის პროფესორი

E-mail: a.kobiashvili@gtu.ge

ანოტაცია. დიდ მონაცემთა ანალიტიკაში აქტუალურია დიდი მოცულობის სხვადასხვა ტიპის, მათ შორის არასტრუქტურირებული მონაცემების შენახვა, დამუშავება და ანალიზი, მათი ბუნებრივი ფორმატით. ორგანიზაცია სულ უფრო მეტად იყენებს მონაცემთა ტბებს, რათა სხვადასხვა სახის გარე მონაცემთა წყაროებიდან შეინახოს და მართოს დიდი მოცულობის არასტრუქტურირებული და სტრუქტურირებული მონაცემები. ტრადიციულ მონაცემთა საცავებისგან განსხვავებით, რომლებიც მონაცემთა წინასწარ დამუშავებას და ორგანიზებას საჭიროებს შენახვამდე, მონაცემთა ტბები საშუალებას იძლევა

შევინახოთ დიდი მონაცემები თავის ფორმატში, რაც უზრუნველყოფს უპრეცედენტო მოქნილობას და მასშტაბურობას. სხვადასხვა ტიპის მონაცემთა მხარდაჭერის ეს უნარი მონაცემთა ტბებს მნიშვნელოვან კომპონენტად აქცევს დიდ მონაცემთა ანალიტიკის, მანქანური სწავლებისა და მონაცემთა დამუშავების სხვა მოწინავე აპლიკაციებისთვის.

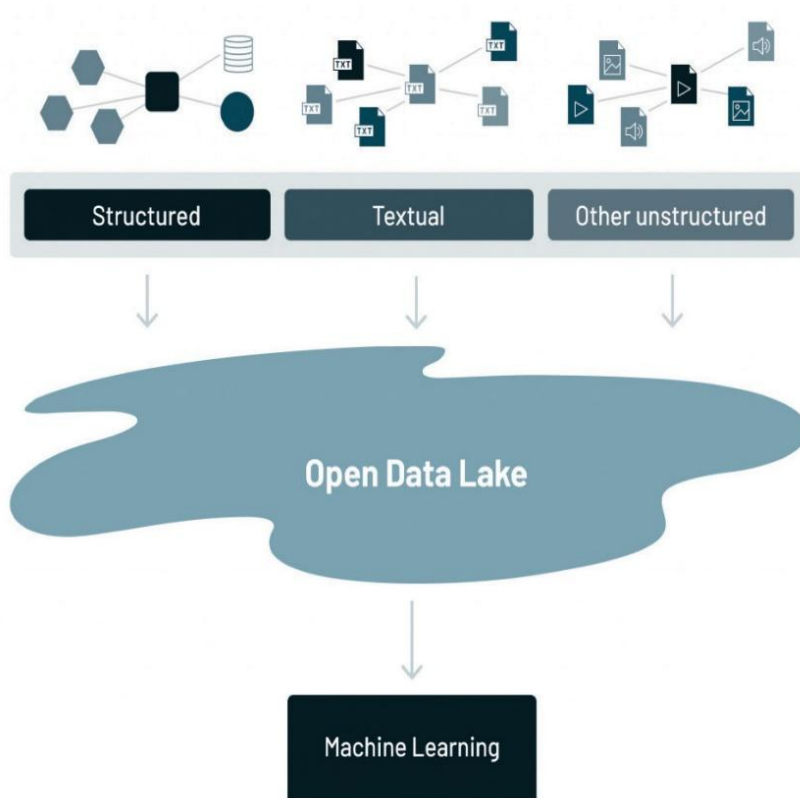
განხილულია მონაცემთა ტბების მნიშვნელობა დიდი მონაცემების ანალიტიკისთვის, მათი ფუნქციები, შესაძლებლობები და სერვისები. აღწერილია ევოლუციური მანქანური სწავლების როლი ტექსტური მონაცემების დამუშავებაში და ამ პროცესებისადმი ახალი მიდგომები. მონაცემთა ტბებში მო-

თხოვნების ოპტიმიზაცია, კერძოდ ევოლუციური ოპტიმიზაცია, დიდ მონაცემთა მართვის ერთ-ერთი მნიშვნელოვანი ასპექტია, რომელიც ადაპტირებულ მიდგომებს იყენებს მოთხოვნების დასამუშავებლად. ამ თვალსაზრისით, ნაშრომში დამუშავებულია მანქანური სწავლების გრაფული მოდელირების ინოვაციური მიდგომები და შემუშავებულია მონაცემთა ტოპოლოგიური მოდელის ოპტიმიზაციის ალგორითმი.

საკვანძო სიტყვები: ევოლუციური მანქანური სწავლება; მონაცემთა ტბები; მონაცემთა ტბის სერვისები; სემანტიკური მოთხოვნების ოპტიმიზაცია.

შესავალი

ნახევრად სტრუქტურირებული და არასტრუქტურირებული დიდი მონაცემების ბუნებრივ ფორმატში პირდაპირი შენახვის აუცილებლობამ განაპირობა ე.წ. „მონაცემთა ტბების“ შექმნა, რომელიც მათი შენახვისთვის Hadoop HDFS-ს იყენებს. მონაცემთა ყოველი ელემენტის შენახვა შესაძლებელია მათთან ასოცირებულ მეტამონაცემებთან (მონაცემთა წყარო, ფორმატი და ა.შ.) ერთად. მონაცემთა ტბების ცნება გაჩნდა მონაცემთა საცავებისა და მონაცემთა ვიტრინების საპირწონედ [1]. მონაცემთა ტბები იყენებს ღია ფორმატს. მონაცემთა ტბის არქიტექტურა წარმოდგენილია სურათზე.



მონაცემთა ტბის არქიტექტურა

მონაცემთა ტბებში არასტრუქტურირებული მონაცემების შენახვისა და ანალიზის შესაძლებლობა საკვანძო უპირატესობაა იმ ორგანიზაციებისთვის, რომლებიც ცდილობენ განავითარონ მოწინავე მანქანური სწავლების მოდელები [2].

მონაცემთა ტბის აგება და ექსპლუატაცია ჯერჯერობით რთულად ითვლება როგორც მეთოდოლოგიურად, ისე ტექნიკური მიზეზებით. თუმცა, მონაცემთა ტბების განვითარებასთან ერთად, მონაცემთა მომსახურების გასაუმჯობესებლად, შემუშავდა მონაცემთა ტბის სერვისები, რომლებიც ფუნქციებისა და შესაძლებლობების ფართო სპექტრს გთავაზობს, რათა გაადვილდეს მონაცემების მიღება, შენახვა, დამუშავება და ანალიზი. მონაცემთა ტბის სერვისები გამოიყენება სხვადასხვა ინდუსტრიასა და დომენში, რაც ორგანიზაციას საშუალებას აძლევს ამოიღოს ქმედითი ინფორმაცია და წარმართოს ბიზნესის ინოვაციები.

მონაცემთა ტბის სერვისები. მონაცემთა ტბის ყველაზე პოპულარულ სერვისებად მიჩნეულია [3]:

Amazon Redshift, რომელიც ახორციელებს მოთხოვნათა სწრაფ შესრულებას, განსაკუთრებით რთული ანალიტიკური მოთხოვნებისთვის, სვეტოვანი შენახვისა და პარალელური დამუშავების შესაძლებლობების გამო.

ძირითადი ფუნქციები მოიცავს მონაცემთა მრავალი ტიპისა და ფორმატის მხარდაჭერას, მონაცემთა დამუშავების პლატფორმებთან უწყვეტ ინტეგრაციას (მაგ., Apache Hadoop, Apache Spark), გაფართოებულ უსაფრთხოებასა და წვდომის კონტროლს და ინტეგრაციას მანქანური სწავლებისა და ხელოვნური ინტელექტის ინსტრუმენტებთან გაფართოებული ანალიტიკისთვის.

Snowflake, რომლის არქიტექტურა საცავისგან გამოყოფს გამოთვლით რესურსებს, რაც მომხმარებლებს საშუალებას აძლევს დამოუკიდებლად გააფართოონ გამოთვლა სამუშაო დატვირთვის მოთხოვნების საფუძველზე. მიუხედავად იმისა, რომ Snowflake ინტეგრირდება პოპულარულ BI და ETL ინსტრუმენტებთან, მისი ეკოსისტემა შეიძლება არ იყოს ისეთი ვრცელი, როგორც ზოგიერთი სხვა პლატფორმა.

Azure Synapse Analytics (ყოფილ SQL მონაცემთა საწყობი), რომელიც შეუფერხებლად ინტეგრირდება Azure-ის სხვა სერვისებთან, როგორიცაა Azure Data Lake Storage, Azure Databricks და Power BI, რაც ხელს უწყობს მონაცემთა ანალიტიკის სამუშაო ნაკადებს თავიდან ბოლომდე. Synapse Analytics გთავაზობს მოთხოვნების სწრაფ შესრულებას, განაწილებული მოთხოვნების დამუშავებისა და შენახვის ოპტიმიზებული ფორმატების გამოყენებას.

Google Cloud Dataflow-ს შეუძლია ადვილად დაამუშაოს როგორც სერიული, ისე ნაკადის მონაცემები, რაც შესაფერის ხდის რეალურ დროში ანალიტიკის გამოყენების შემთხვევებს. მონაცემთა ნაკადი შეუფერხებლად ინტეგრირდება Google Cloud სერვისებთან, როგორიცაა BigQuery, Cloud Storage და Pub/Sub, რაც ხელს უწყობს მონაცემთა დამუშავებისა და ანალიზის სამუშაო ნაკადებს.

Google BigQuery არის სერვერის გარეშე მონაცემთა საცავი, რაც ნიშნავს, რომ მომხმარებლებს არ სჭირდებათ რაიმე ინფრასტრუქტურის მართვა. იგი შექმნილია უპრობლემოდ მასშტაბირებისთვის და მომხმარებლებს საშუალებას აძლევს პეტაბაიტი მონაცემები გააანალიზონ მაღალი წარმადობით. BigQuery შეუფერხებლად ინტეგრირდება Google

Cloud Platform-ის (GCP) სხვა სერვისებთან, როგორცაა Google Cloud Storage, Google Data Studio და Google Cloud Dataprep, უზრუნველყოფს მონაცემთა უწყვეტი ტრანსფორმაციისა და ვიზუალიზაციის სამუშაო ნაკადებს.

ზემოგანხილულ ზოგიერთ სერვისში (AzureML, Spark MLlib, Azure Synapse Analytics, BigQuery) ერთ-ერთი ცენტრალური ადგილი უჭირავს მანქანური სწავლების კომპონენტებს, რაც განპირობებულია მათი არქიტექტურით.

მანქანური სწავლების როლი მონაცემთა ტბებში.

მანქანური სწავლება (ML) ტრანსფორმაციულ როლს ასრულებს მონაცემთა ტბების მუშაობაში. ერთ-ერთი მთავარი სფეროა, სადაც ML-ს შეუძლია მნიშვნელოვანი გავლენა მოახდინოს, მანქანური სწავლების ალგორითმებს შეუძლია გააანალიზოს მონაცემთა გამოყენების შაბლონები და ავტომატურად დაარეგულიროს მონაცემთა ტბის კონფიგურაციები [4].

სტატიაში აღნიშნულია ევოლუციური მანქანური სწავლების (Evolutionary Machine Learning - EML) როლი, განსაკუთრებით არასტრუქტურირებული, კერძოდ ტექსტური მონაცემების დამუშავებაში. მონაცემთა ტბებში უნიკალური გამოწვევებია მონაცემთა ჰეტეროგენურობის, სქემის ცვალებადობისა და ფართომასშტაბიანი ოპერაციების პირობებში მოთხოვნის დამუშავება. ცნობილია, რომ ამ პრობლემების გადასაჭრელად გამოიყენება მოთხოვნის ოპტიმიზაციის რამდენიმე ტრადიციული მეთოდი, როგორიცაა: დინამიკური დაპროგრამება (DP), გენეტიკური ალგორითმი (GA), ევოლუციური დაპროგრამება (EP), გენეტიკური დაპროგრამება (GP), ხელოვნური ნეირონული ქსელები (ANN), ჭიანჭველების კოლონიის ალგორითმი (ACA) და ნაწილა-

კების გროვის ოპტიმიზაცია (PSO) და ა.შ. ისე, რომ სრულად არ შეეხოთ მონაცემთა ტბების უნიკალურ მახასიათებლებს, მათ შორის სქემის მოქნილობას და მონაცემთა ჰეტეროგენულობას.

მიუხედავად იმისა, რომ არსებობს რამდენიმე მოხილვა ევოლუციური გამოთვლის (EC) გამოყენების შესახებ მანქანური სწავლების (EML) პრობლემებისთვის, ისინი ხშირად ფოკუსირებულია კონკრეტულ ასპექტებზე ან ამოცანებზე, როგორცაა მახასიათებლების შერჩევა ან კლასიფიკაცია გენეტიკური დაპროგრამების გამოყენებით. თუმცა, ჯერ კიდევ არ არსებობს ყოვლისმომცველი კვლევა EML ტექნიკის შესახებ სხვადასხვა პრობლემისთვის, როგორცაა ცოდნის მოპოვება დიდი მონაცემებიდან.

ნაშრომი მიზნად ისახავს ახალი მიდგომა წარმოადგინოს EC ტექნიკის გამოყენებით ძირითადი ML პრობლემების გადასაჭრელად, ამ სფეროში სწრაფი პროგრესის გათვალისწინებით, რაც ML აპლიკაციებს უფრო მეტ შესაძლებლობას აძლევს. ევოლუციური ოპტიმიზაციისა და მანქანური სწავლების მეთოდების სფეროში მნიშვნელოვანი მუშაობის მიუხედავად, ჩვენი აზრით, ზოგიერთი კვლევის პრობლემა გადაუჭრელი რჩება.

ძირითადი ნაწილი

მანქანური სწავლების გრაფული მოდელის აღწერა. წარმოგიდგენთ მოდელირებისა და ოპტიმიზაციის ჩვენს ინოვაციურ მიდგომას. მათემატიკური ფორმალიზმის იარაღად გამოიყენება მონაცემთა ტოპოლოგიური მოდელი, ე.ი. *აქსონ-დენდრიტების*, მათ შორის სინაპსური კავშირებით. ამ ნაშრომის ორიგინალობა მდგომარეობს სისტემის, როგორც ნეირონული გრაფის აღწერაში სინერგიუ-

ლი წიბოებით, წარმოდგენილი ენტროპიის/სინერგიის არაცხადი სუპერპოზიციით, სადაც თითოეული კომპონენტი კვანძია. ფართო გაგებით, ნებისმიერი ფართომასშტაბიანი სისტემა და მისი კომპონენტები პირობითად შეიძლება ჩაითვალოს გრაფად (ან ჰიპერგრაფად). ზოგადად, ეს მოდელები შეიქმნა ნებისმიერი მასშტაბის რთული განაწილებული სისტემების ასაღწერად, რომლებიც მოქმედებს, როგორც სამყაროს სრულყოფილი ფრაქტალი [5].

აღვნიშნოთ პირობითად მიუმართავი გრაფი $G = (V, E, C)$ თითოეული აგენტისთვის, როგორც ნეირონული (აქსონ-დენდრიტული) მოდელი მიმდებარე მატრიცით, სადაც V კვანძების სიმრავლეა და E – წიბოების სიმრავლე. კვანძოვანი ინდექსები მიეკუთვნება ნეირონებს აქსონ-დენდრიტების ნაკრებით. C კავშირია ნეირონებს შორის, რომლებიც რეალიზებულია სინაპსების საშუალებით. ნაკრები V შედგება N დინამიკური აგენტებისგან, რომლებიც შეიძლება წარმოდგენილი იყოს რობოტებით ან ნებისმიერი სოციალური სისტემის წევრებით, რომლებიც განაწილებულია სიბრტყეზე ან სივრცეში.

შემდეგ განვიხილავთ სემანტიკური ქრომოსომის მოდელის ახალ კონცეფციას. ამ მოდელში ქრომოსომა, და არა უბრალო სტრიქონი, წარმოდგენილია სინერგიული გრაფით (ან ჰიპერგრაფით), სადაც ნეირონები გენების როლს ასრულებს. სინერგიული გრაფის სტაბილურობა (ჰომეოსტაზი) ემსახურება მოცემული ქრომოსომის ვარგისობის ფუნქციას. ამრიგად, ქრომოსომა ან მისი კომპონენტები (ე.წ. გენები) შეიძლება ჩაითვალოს სინერგიულ ან ნეირონულ (აქსონ-დენდრიტულ) მოდელად.

როგორც წესი, თითოეული აქსონი ან დენდრიტი აღწერილია, როგორც ტერმინალი [6]:

$$t_{ik} = \{s_{ik}, d_{ik}, \omega_{ik}\}. \quad (1)$$

სადაც $s_{ik} \in \{-1, +1\}$ არის ტერმინალის ნიშანი; $d_{ik} \in D$ – ტერმინალის ტიპი; $\omega_{ik} \in [0, 1]$ – ტერმინალის წონის კოეფიციენტი.

ნეირონებს შორის კავშირი რეალიზდება სინაპსებით

$$C_{ij} = \{t_{ik} \circ t_{kj}\}, \quad (2)$$

სადაც $\circ$ არის სინაპსი ანუ კავშირი.

ალგორითმში განიხილება თავსებადობის ორი განსხვავებული ტიპი: ფუნქციური და პარამეტრული.

მეზობელი კვანძების სიმრავლე წარმოდგენილია სინაპსით განსხვავებულ კვანძებს შორის. თითოეული სინაპსი კონკრეტული პირობების საფუძველზე იქმნება:

$$C_{ij} = \left\{ (s_{ik} = -s_{kj}) \wedge (d_{ik} = d_{jk}) \wedge (|\omega_{ik} - \omega_{jk}| = \min_k) \right\}, \quad (3)$$

სადაც $s_{ik} = -s_{kj}$ არის ტერმინალების საპირისპირო პოლარობა, რაც თავსებადობის ფუნქციურობას ასახავს;

$$d_{ik} = d_{jk} - \text{ტიპის იდენტურობა};$$

$$(\omega_{ik} - \omega_{jk}) = \min - \text{წონის კოეფიციენტებში მინი-}$$

მალური სხვაობა, რომელიც რეალურად პარამეტრული შეუთავსებლობის ხარისხს განსაზღვრავს.

განვიხილოთ გარემო, როგორც სისტემის ვირტუალური ელემენტი. წონის კოეფიციენტები მისი ყველა ტერმინალისთვის იქნება: $\omega_{or} = 0$,

$$\text{სადაც } r = \overline{1, F}.$$

$$R - \text{სინაპსების რაოდენობა: } R = \text{Card}\{C_{ij}\};$$

F – თავისუფალი ტერმინალების რაოდენობა:

$$F = Q - 2R;$$

Q - ტერმინალების საერთო რაოდენობა.

$$Q = \sum_{i=1}^N \sum_{k=1}^L t_{ik} \quad (4)$$

μ – შეუთავსებლობის ხარისხი

$$\mu_{r_{(ik)}} = |\omega_{ik} - \omega_{jk}|. \quad (5)$$

გრაფი ითვლება დაკავშირებულად, თუ სინაპსი გრაფის ორ განსხვავებულ კვანძს შორისაა. გრაფის ენტროპია შეიძლება განისაზღვროს მოცემული მდგომარეობის მიღწევის გზების რაოდენობით და გამოითვლება შემდეგი ფუნქციით:

$$H = - \sum_{r_{(ik)}=1}^F \mu_{r_{(ik)}} \log \mu_{r_{(ik)}} - \sum_{r_{(ik)}=1}^R P_{r_{(ik)}} (\mu_{r_{(ik)}} \log \mu_{r_{(ik)}} + (1 - \mu_{r_{(ik)}}) \log(1 - \mu_{r_{(ik)}})). \quad (6)$$

განაწილებული სისტემის ქცევას განსაზღვრავს როგორც გარე, ისე შინაგანი თავისუფლება. სინაპსების თავსებადობა აუცილებელი პირობაა ნეირონის გრაფის ერთიანობისთვის.

ზოგადად, სისტემის მოდელი არის მრავალგანზომილებიანი გრაფი, სადაც განზომილება განისაზღვრება ტერმინალების ტიპების რაოდენობით. კლასტერი შეიძლება ჩაითვალოს, როგორც ქვეგრავი ან გრაფის პროექცია ნებისმიერი ტიპის ტერმინალზე, რაც გრაფს კლასტერების კრებულად აქცევს.

სინერგია, თავის მხრივ, არის ფუნქცია:

$$S = \log \sum_{i=1}^n \mu_i - \sum_{i=1}^h p_i \log p_i, \quad (7)$$

სადაც h იზომორფული ჯგუფების ორბიტების რაოდენობაა;

p – იზომორფული ჯგუფების ორბიტების ალბათობა.

თითოეული სინაპსის არამკაფიოება ანუ არაცხადი მდგომარეობა განისაზღვრება ენტროპია-სინერგიის სუპერპოზიციის ალბათობით, რაც ძირითადად გავლენას ახდენს მთლიან გრაფზე. მთლიანი სისტემის სტაბილურობა ანუ ჰომეოსტაზი მოცემულ მომენტში სინერგიისა და ენტროპიის სხვაობით განისაზღვრება:

$$M_h = S - H. \quad (8)$$

როგორც წესი, ენტროპია ასოცირდება უწყვეტობის, არეულობის და/ან ქაოსის დონესთან. შემოგთავაზებთ ენტროპიის ახალ კონცეფციას, როგორც შინაგან ქცევით შეუთავსებლობას (რეზისტენტობა) ანუ დეზორიენტირებული კომპონენტების ქცევითი ვექტორებს შორის შეუსაბამობას. მოდელის აგება და ოპტიმიზაცია შესაძლებელია ენტროპიის მინიმიზაციის გზით. სისტემის მდგომარეობას განაპირობებს კვანძების ქცევითი ვექტორები, რომლებიც განისაზღვრება თითოეული ნეირონის სინაპსური კავშირებით. ნებისმიერი სწავლების პროცესი სისტემაში შეიძლება ჩაითვალოს, როგორც გადასვლა ერთი მდგომარეობიდან მეორეზე, რაც მჭიდროდ არის დაკავშირებული ენტროპიის ცვლილებებთან. ძირითადი თვისება, რომელიც გამოიყენება ოპტიმიზაციის დასახასიათებლად, არის სისტემის მდგომარეობის ცვლილების კრიტერიუმი, რომელიც იდენტიფიცირებულია ნეირონული გრაფის ენტროპიასთან.

მონაცემთა ტოპოლოგიური მოდელის ოპტიმიზაცია. ამ მიდგომის ფუნდამენტური პრინციპი და უნიკალურობა მდგომარეობს ნეირონული გრაფის ეტაპობრივ რეკონფიგურაციაში მისი ფიტნესფუნქციის გასაძლიერებლად, ისეთი გენეტიკური ოპერაციების გამოყენების გარეშე, როგორიცაა კროსოვერი,

მუტაცია ან ინვერსია. ხელახლა კონფიგურირებადი ქრომოსომის მოდელი წააგავს თვითშეკრებას, ფართოდ გავრცელებულ ბუნებრივ პროცესს, სადაც არსებული კომპონენტების უწესრიგო ნაკრები ავტონომიურად ერწყმის უფრო რთულ და ორგანიზებულ სტრუქტურებს, რაც ჰგავს, სუპრამოლეკულურ თვითშეკრების პროცესების მსგავსად, ტვინში ნეირონული ქსელის ფორმირებას. თავდაპირველად, გრაფის ოპტიმიზაცია რეკონფიგურაციის გზით ჯერ ზუსტად არ არის განსაზღვრული. საწყისი მოდელის შემთხვევითი ინიციალიზაციის შემდეგ ნეირონული აქსონ-დენდრიტის კავშირების რეკომბინაცია გამოიყენება ახალი გადაწყვეტილების შესაქმნელად, რომელიც დაფუძნებულია კრიტერიუმებზე, როგორიცაა ქრომოსომის ენტროპიის მინიმიზაცია ან სინერგიის მაქსიმიზაცია (ფიტნეს-ფუნქცია).

ალგორითმის ბლოკ-სქემის აღწერა. მოცემულ განყოფილებაში წარმოგიდგენთ ალგორითმულ აღწერას [7]. კვლევა კონცენტრირებული იყო სწავლების ახალ მიდგომაზე, რომელიც ორიენტირებულია ნეირონული გრაფების სტრუქტურულ რეკონფიგურაციაზე. შემდგომში, ამ ალგორითმების ეფექტურობისა და შესრულების შეფასება ხდება სხვა მიდგომებთან შეადარებითი ანალიზის საფუძველზე. ალგორითმი იწყება ნეირონების ნაკრებით, მათი ტერმინალებითა და ასოცირებული ატრიბუტებით, როგორიცაა ნიშნები, ტიპები, წონის კოეფიციენტები, როგორც შეზღუდვები – წონის კოეფიციენტების ლიმიტები. ეს საწყისი ნაწილი ადგენს საფუძველს ალგორითმის ნაკადის შემდგომი ბიჯებისთვის:

ბიჯი 1. ნეირონული გრაფის საწყისი კონფიგურაციის შემთხვევითი ფორმირება (დაკავშირების მატრიცა) სინაპსების თავსებადობის თითოეული

მდგომარეობისთვის: $C_{ij} = \{(s_{ik} = -s_{kj}) \wedge (d_{ik} = d_{jk})\}$.

ბიჯი 2. თითოეული ნეირონისთვის თითოეული სინაპსის μ შეუთავსებლობის ხარისხის გამოთვლა განტოლებით: $\mu_{r_{(ik)}} = |\omega_{ik} - \omega_{jk}|$, ენტროპია განტოლებით (გამოსახულება 6) და ენტროპია საწყისი გრაფისთვის.

ბიჯი 3. არსებობს ორი გზა:

ა) თუ გრაფის ენტროპია უდრის ნულს, გადავდივართ მე-9 საფეხურზე;

ბ) წინააღმდეგ შემთხვევაში:

ბიჯი 4. ენტროპიული ნეირონების დახარისხება-სორტირება.

ბიჯი 5. არსებობს ორი გზა:

ა) თუ ენტროპიული ნეირონების რაოდენობა ნულს უდრის, გადავდივართ მე-7 ბიჯზე;

ბ) წინააღმდეგ შემთხვევაში:

ბიჯი 6. საწყისი დედუქციური მუტაცია ანუ „აუქციონი“ (გრაფის რეკონფიგურაცია).

ვბრუნდებით მე-5 ბიჯზე.

ბიჯი 7. არსებობს ორი გზა:

ა) თუ გრაფის მთელი ენტროპია ნულს უდრის, გადავდივართ მე-9 ბიჯზე;

ბ) წინააღმდეგ შემთხვევაში:

ბიჯი 8. ახალი ნეირონის გენერაცია თითოეული ენტროპიული ნეირონის მეზობლად. ვბრუნდებით მე-4 ბიჯზე.

ბიჯი 9. ნეირონული გრაფის სწავლა დასასრულს უახლოვდება.

დადებითი და უარყოფითი მხარეები. შემუშავებული ალგორითმის მთავარი ღირსება მდგომარეობს მის ორიგინალურობაში და ბუნებრივი ევოლუციის სიმულაციის უნარში, ოპტიმიზაციის პრობლემების გადასაჭრელად. შედეგები აჩვენებს,

რომ შემოთავაზებული ევოლუციური ოპტიმიზაციის მიდგომა გთავაზობთ რამდენიმე უპირატესობას, გენეტიკურ ალგორითმებთან შედარებით:

1. არ არის საჭირო ქრომოსომის პოპულაციების შემთხვევითი წარმოქმნა;
2. არ არის საჭირო კროსოვერების ან ინვერსიების შემთხვევითი შესრულება;
3. შემთხვევითი მუტაცია არ არის საჭირო, რადგან სინერგიული გრაფის ყოველი ხელახალი კონფიგურაცია მიზნობრივ მუტაციას ემსახურება.

როგორც ნაკლი, დამატებითი მონაცემები კვანძებისთვის (როგორიცაა ნეირონების ნაკრები, მათი ტერმინალები, ნიშნები, ტიპები, წონის კოეფიციენტები და შეზღუდვები, როგორიცაა წონის კოეფიციენტების ლიმიტები) შეიძლება ჩაითვალოს რთულად.

დასკვნა

ნაშრომში მონაცემთა ტბების ზოგიერთ სერვისისთვის (AzureML, Spark MLLib, Azure Synapse Analytics, BigQuery) შემოტანილია ევოლუციური მანქანური სწავლების ინოვაციური მიდგომა, ნეი-

რონული გრაფების საფუძველზე. ამ მიდგომის გამოჩენილი თვისება მოიცავს ეტაპობრივ ევოლუციურ რეკონფიგურაციას გრაფული მოდელის გაუმჯობესებისთვის, კერძოდ ქრომოსომის ფართო პოპულაციების შემთხვევით გენერაციისა და გენეტიკური ოპერაციების (კროსოვერი, მუტაცია და ინვერსია) გამოყენების გარეშე. რეკონფიგურირებადი ნეირონული გრაფი შეიძლება შევადაროთ თვითშეკრებას, გავრცელებულ ბუნებრივ პროცესს, სადაც არსებული კომპონენტების უწყესრიგო ნაკრები ავტონომიურად ერწყმის უფრო რთულ და ორგანიზებულ სტრუქტურებს, სუპრამოლეკულური თვითშეკრების პროცესების ან ტვინში ნეირონული ანსამბლების აგების მსგავსად.

დაბოლოს, შემოთავაზებული ევოლუციური სწავლების მიდგომა, გარდა მოცემული პრობლემატიკისა, წარმატებით შეიძლება დაინერგოს ტოპოლოგიურად განაწილებულ დინამიკურ სისტემებში, რაც ახალ პერსპექტივებს იძლევა სხვადასხვა ტოპოლოგიურ სფეროში.

ლიტერატურა

1. Inmon, W. H. (2016). Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump." Basking Ridge, NJ 07920 USA. *Technics Publications*. <https://www.TechnicsPub.com>
2. Furche, T., Gottlob, G., Libkin, L., Orsi, G. & Paton, N. W. (2016). Data wrangling for big data: Challenges and opportunities, in EDBT.
3. Reis, J. & Housley, M. Fundamentals of Data Engineering. Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. <http://oreilly.com>.
4. Telikani, A. Tahmassebi, A., Banzhaf, W. & H.Gandomi, A. (2021). Evolutionary Machine Learning: A Survey. CM Comput.Surv.1,1, Article1, p.35. <https://doi.org/10.1145/3467477>
5. Meparishvili, B., Gachechiladze, T. & Janelidze, G. (2007). NATO Science for Peace and Security Series "Complexity and Security", ISSN 1874-6276, pp.379-388.
6. Janelidze, G. & Meparishvili, B. (2006). Evolution Algorithm of Multiextreme Optimization. *Periodical Scientific Journal „Intelecti”*, N1(24), pp.119-121. ISSN 1512-0333.
7. Meparishvili, B. (2010). "New approach to evolutionary algorithms". ERA-5 Proceedings The Contribution Of Information Technology Science, Economy, Society an Education. T.E.I. of PIREAUS.

UDC 008

SCOPUS CODE 1702

<https://doi.org/10.36073/1512-0996-2025-2-133-141>

Evolutionary Machine Learning in Data Lake Services

Badri Meparishvili	Department of Information Technologies, Georgian Technical University, Georgia, 0160, Tbilisi, 77, M. Kostava str. E-mail: meparishvilibadri08@gtu.ge
Gulnara Janelidze	Department of Information Technologies, Georgian Technical University, Georgia, 0160, Tbilisi, 77, M. Kostava str. E-mail: janelidzegulnara08@gtu.ge
Giorgi Muradov	Department of Information Technologies, Georgian Technical University, Georgia, 0160, Tbilisi, 77, M. Kostava str. E-mail: muradovi.giorgi22@gtu.ge

Reviewers:

M. Chkhaidze, Professor, Faculty of Informatics and Management Systems, GTU

E-mail: m.chkhaidze@gtu.ge

A. Kobiashvili, Professor, Faculty of Informatics and Management Systems, GTU

E-mail: a.kobiashvili@gtu.ge

Abstract. In big data analytics, the storage, processing, and analysis of large volumes of various types of data, including unstructured data, in their natural format is of great relevance. The article discusses aspects of the use of machine learning in the context of big data lake services. Modern organizations are increasingly using data lakes to store and manage large volumes of unstructured and structured data from various types of external data sources. Unlike traditional data warehouses, which require pre-processing and organization of data before storage, data lakes allow us to store big data in its own format, which provides unprecedented flexibility and scalability. This ability to support various types of data makes data lakes an important component for big data analytics, machine learning, and other advanced data processing applications. In addition, query optimization in data lakes, namely evolutionary optimization, is one of the key aspects of big data management, which uses adaptive approaches to query processing. The article also discusses a novel approach to machine learning in the context of evolutionary optimization.

Keywords: Data lakes; Gata lakes services; Evolutionary machine learning; Semantic query optimization.

განხილვის თარიღი 11.12.2024

შემოსვლის თარიღი 13.12.2024

ხელმოწერილია დასაბეჭდად 16.06.2025