ამოცნობის საიმედოობის პროგნოზირება კლასტერირების შედეგების მიხედვით

ოთარ ვერულავა, თეა თოდუა, ლაშა ვერულავა
საქართველოს ტექნიკური უნივერსიტეტი

რეზიუმე

განიხილება ამოცნობის პროცესის პროგნოზირების პრობლემა კლასტერირების შედეგებზე დაყრდნობით. კერძოდ, შემოტანილია სახეების (კლასტერების) გავლენის ზონის ცნება, რის საშუალებითაც ხდება ამოცნობის პროცესში შესაძლო შეცდომების პროგნოზირება. მოცემულია გავლენის ზონების თანაკვეთისა და არათანაკვეთის შემთხვევები. როგორც ამ უკანასკნელის კერძო შემთხვევა, განხილულია არაკომპაქტური კლასტერებიც.

საკვანძო სიტყვები: სახეთა ამოცნობა. კლასტერირება. გავლენის ზონები. პროგნოზირება.

1. შესავალი

ამოცნობის საიმედოობის პროგნოზირებისათვის გამოიყენება კლასტერირების პროცესი რანგული კავშირებით [1,2,3], რომლის შედეგები საშუალებას გვაძლევს დავადგინოთ კლასტერების შემდეგი მახასიათებლები:

1. კლასტერების რაოდენობა;
2. თითოეულ კლასტერში შემავალი რეალიზაციების ნუსხა და მათი რაოდენობა;
3. კლასტერის აგების მახასიათებელი – აგების რანგის მნიშვნელობა;
4. კლასტერების განმხოლოების მაჩვენებელი – რანგების გამოტოვებების რაოდენობა.

კლასტერირების ოთხივე მახასიათებელი წარმოადგენს სკალარებს, რომელთა მნიშვნელობა მით უფრო სანდოა, რაც უფრო მეტია რეალიზაციათა რაოდენობა თითოეული სახის სასწავლო ნაკრებში. რეალიზაციათა სიმრავლის წარმომადგენლობითობის პირობა საკმაოდ მკაცრი მოთხოვნაა და მისი დაკმაყოფილება პრაქტიკული ამოცანებისათვის ხშირად ვერ ხერხდება, ამიტომ იყენებენ ემპირიულ კრიტერიუმს, რაც შემდეგში გამოიხატება: რაც მეტია რეცეპტორული ველის განზომილება, მით მეტი უნდა იყოს რეალიზაციების რაოდენობა თითოეული სახის სასწავლო ნაკრებებში.

წარმოვადგინოთ რამდენიმე ცნება კლასტერირების შედეგების შესაფასებლად.

განსაზღვრა 1. კლასტერი კომპაქტურია, თუ მასში გაერთიანებულია მხოლოდ ერთი სახის რეალიზაციები, წინააღმდეგ შემთხვევაში კლასტერი არაკომპაქტურია.

განსაზღვრა 2. სახე კომპაქტურია, თუ იგი წარმოდგენილია მხოლოდ კომპაქტური კლასტერებით, წინააღმდეგ შემთხვევაში სახე არაკომპაქტურია.

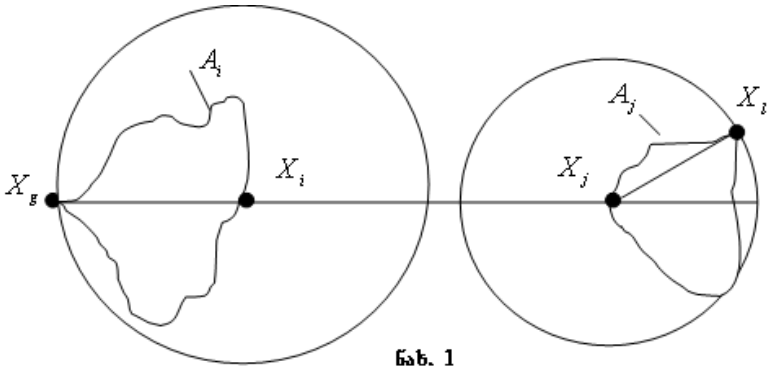
2. ამოცნობის საიმედოობის პროგნოზირება კომპაქტური სახეებისათვის

დავუშვათ, სახეთა რეალიზაციების X სიმრავლე კლასტერირებულია და მიღებულია კლასტერირების ზემოთ მოცემული მახასიათებლის მნიშვნელობები. სიმარტივისათვის დავუშვათ, რომ სახეთა სიმრავლიდან აღებულია ორი A_i და A_j ელემენტი, რომლებისთვისაც უნდა განისაზღვროს სწორად ამოცნობის ან შეცდომის დაშვების ალბათობები.

დავუშვათ, რომ A_i და A_j სახეები კომპაქტურია და მათი შესაბამისი კლასტერების განლაგება ნიშანთა სივრცეში ანალოგიურია სურ. 1-ზე ნაჩვენები შემთხვევისა.

განვსაზღვროთ სხვადასხვა კლასტერებში შემავალ რეალიზაციებს შორის მინიმალური მანძილი, რაც წარმოადგენს სიმრავლეებს შორის ჰაუსდორფის მეტრიკას.

დავუშვათ, ჰაუსდორფის მეტრიკა ხორციელდება $X_i \in A_i$ და $X_j \in A_j$ რეალიზაციებზე (სურ. 1). განვსაზღვროთ მაქსიმალური მანძილი X_i და X_j რეალიზაციებიდან თავისავე კლასტერში შემავალ რეალიზაციებამდე, რომლებიც 1-ელ ნახაზზეა წარმოდგენილი შესაბამისად X_g და X_l წერტილებით – რეალიზაციებით.



ნახ. 1

შემოვხაზოთ X_i და X_j წერტილებიდან (რეალიზაციებიდან) ჰიპერსფეროები (წრეწირები) X_iX_g და X_jX_l რადიუსებით.

განსაზღვრა 3. A_i სახის გავლენის ზონა A_j სახის მიმართ ეწოდება ნიშანთა სივრცის იმ ნაწილს, რომელიც შემოსაზღვრულია X_iX_g რადიუსით შემოწერილი ჰიპერსფეროთი.

განსაზღვრა 4. A_j სახის გავლენის ზონა A_i სახის მიმართ ეწოდება ნიშანთა სივრცის იმ ნაწილს, რომელიც შემოსაზღვრულია X_jX_l რადიუსით შემოწერილი ჰიპერსფეროთი.

მივიღოთ მხედველობაში, რომ ზემოთ მოყვანილი განსაზღვრებები შესაძლებელია გამოვიყენოთ სახეთა A სიმრავლის ელემენტების ნებისმიერი წყვილისათვის.

განვიხილოთ გავლენის ზონების განლაგების რამდენიმე შემთხვევა:

1. გავლენის ზონები არ იკვეთება.

   A_i და A_j სახეებისათვის ეს ნიშნავს, რომ X_iX_g და X_jX_l რადიუსიანი ჰიპერსფეროები არ იკვეთებიან (სურ.1), რაც აღიწერება შემდეგი გამოსახულებით:

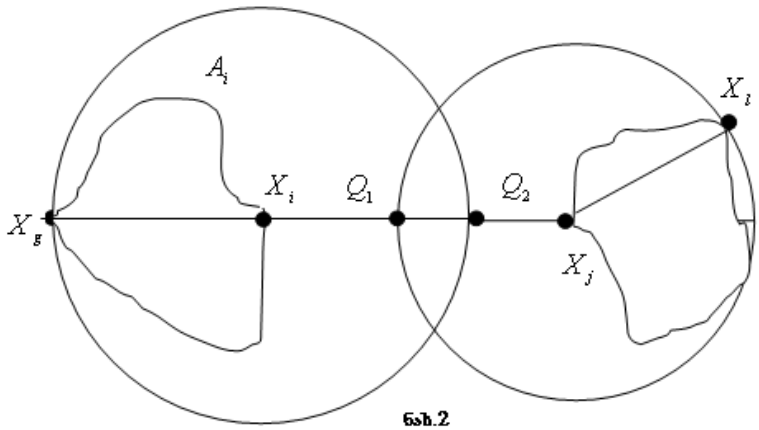
   $$X_iX_g + X_jX_l < X_iX_g \qquad (1)$$

   რანგული კავშირებით კლასტერირების შედეგების მიხედვით კი გვექნება შემდეგი უტოლობა:

   $$Rank(X_i; X_g) + Rank(X_j; X_l) < Rank(X_i; X_j)$$

2. გავლენის ზონები იკვეთება (ნახ.2).

ცხადია, რომ გავლენის ზონების გადაკვეთა არ ნიშნავს კლასტერების თანაკვეთას, რადგან ასეთ შემთხვევაში დაირღვეოდა სახეების კომპაქტურობის პირობა.



ნახ.2

მე-2 ნახაზზე გამოსახული შემთხვევა შესაძლოა წარმოავადგინოთ შემდეგი უტოლობით:

$$X_iX_g + X_jX_l > X_iX_j \qquad (2)$$

რანგული კავშირების მიხედვით გვექნება:

$$Rank(X_i;X_g)+Rank(X_j;X_l)>Rank(X_i;X_g)$$

ამოცნობის საიმედოობის პროგნოზირება ემყარება შემდეგ აქსიომებს:

აქსიომა 1. ნებისმიერი სახის რეალიზაციები სხვა სახის მიმართ შესაძლოა გამოჩნდეს მხოლოდ თავის გავლენის ზონაში ამ სახის მიმართ.

აქსიომა 2. ამოცნობის პროგნოზირებული საიმედოობა განისაზღვრება მოცემული სახის გავლენის ზონაში არსებული რეალიზაციების რაოდენობის ფარდობით კლასტერში მოთავსებულ რეალიზაციების მთელ რაოდენობასთან.

აღვნიშნოთ A_i სახის რეალიზაციების რაოდენობა M_i-ით, ხოლო M_{ij}-ით იმ რეალიზაციების რაოდენობა, რომელიც განლაგდა A_j სახის გავლენის ზონაში. ამოცნობის შეცდომის (ავღნიშნოთ P_{ij}^--ით) შესაფასებლად აქსიომა 2-ის მიხედვით, გვექნება:

$$P_{ij}^- = \frac{M_{ij}}{M_i} \qquad (3)$$

სადაც P_{ij}^- - შეცდომის ალბათობაა, $i, j = \overline{1, I}$, $i \neq j$.

სწორად ამოცნობის ალბათობებისათვის, ავღნიშნოთ P_{ij}^+, გვექნება:

$$P_{ij}^+ = 1 - P_{ij}^-$$

აქსიომა 1-დან გამომდინარე, თუ გავლენის ზონები არ იკვეთება, მაშინ ამ ზონებში სხვა სახის რეალიზაციების გამოჩენის ალბათობა ტოლია ნულის, შესაბამისად, გვექნება $M_{ij} = 0$ და გამოსახულება 3-ის მიხედვით მივიღებთ, რომ $P_{ij}^- = 0 \Rightarrow P_{ij}^+ = 1$

ამოცნობის საიმედოობის შესაფასებლად გამოვიყენოთ გავლენის ზონების თანაკვეთის მაქსიმალური სიგანის ფარდობა კლასტერებს შორის ჰაუსდორფის მეტრიკით განსაზღვრულ მანძილთან. თანაკვეთის ზონის მაქსიმალური სიდიდე მე-2 ნახაზზე აღნიშნულია Q_1Q_2 მანძილით. მის მანძილის გამოსათვლელად გამოვიყენოთ შემდეგი გამოსახულება:

$$Q_1Q_2 = X_iX_j - (X_iQ_1 + Q_2X_j) \qquad (4)$$

$X_1Q_1 = X_iX_j - X_jQ_{1;}$ $Q_2X_j = X_iX_j - X_iQ_2$. ჩავსვათ გამოსახულება (4)-ში X_1Q_1 და Q_2X_j მიღებული მნიშვნელობები; მივიღოთ მხედველობაში, რომ $X_iQ_1 = X_iX_g$ და $X_jQ_2 = X_jX_l$; მარტივი გარდაქმნებით მივიღებთ:

$$Q_1Q_2 = -X_iX_j + X_jX_l + X_iX_g \qquad (5)$$

იმის გამო, რომ Q_1Q_2 სიდიდე მანძილია და ამის გამო დადებითი სიდიდეა, ასევე საიმედოობის მნიშვნელობები მხოლოდ არაუარყოფითი სკალარებია, ამიტომ მივიღოთ, რომ

$$Q_1Q_2 = |-X_iX_j + X_jX_l + X_iX_g| \qquad (6)$$

ამოცნობის საიმედოობის შესაფასებლად გვექნება:

$$P_{ij}^- = \frac{|-X_iX_j + X_jX_l + X_iX_g|}{X_iX_j} \qquad (7)$$

იგივე შეიძლება ვიანგარიშოთ რანგული კავშირებით მიღებული გამოსახულებების მიხედვით, რისთვისაც გამოვიყენოთ (2) უტოლობა, რომელიც გადავწეროთ შემდეგნაირად:

$$Rank(X_iX_j) - (Rank(X_i;X_g) + Rank(X_j;X_l)) < 0 \qquad (8)$$

როგორც (8) გამოსახულებიდან ჩანს, მიღებული სხვაობა უარყოფითი სიდიდეა, ამიტომ

(7) გამოსახულების მსგავსად ვისარგებლოთ სხვაობის აბსოლუტური მნიშვნელობით. შესაბამისად, რანგული კავშირებით ამოცნობის საიმედოობის შესაფასებლად გვექნება:

$$P_{ij}^{-} = \frac{|\,Rank(X_i;X_j) - (Rank(X_i;X_j) + Rank(X_j;X_l))\,|}{Rank(X_i;X_j)} \qquad (9)$$

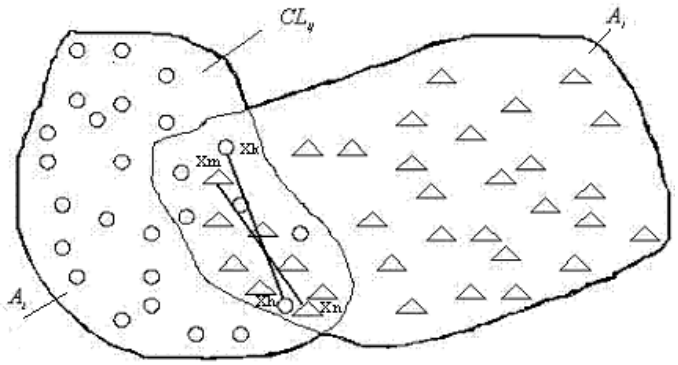
სადაც $i, j = \overline{1;I}, i \neq j$.

(7) გამოსახულებით ვისარგებლოთ მაშინ, როდესაც კლასტერირება არაა შესრულებული რანგული კავშირებით, ხოლო გამოსახულება (9)-ით – რანგული კავშირებით კლასტერირებისას.

3. ამოცნობის საიმედოობის პროგნოზირება არაკომპაქტური სახეებისათვის

არაკომპაქტური სახეები, განსაზღვრა 2-დან გამომდინარე, გვაქვს მაშინ, როდესაც ერთ კლასტერში გაერთიანებულია ერთზე მეტი სახის რეალიზაციები. დავუშვათ, რომ A_i და A_j სახეები არაკომპაქტურია, რომელთა შესაბამისი არაკომპაქტური კლასტერი აღვნიშნოთ CL_{ij}-თი. (ნახ.3). აქ ორმაგი კონტურული ხაზით გამოსახულია CL_{ij} კლასტერი, ხოლო ერთმაგი კონტურული ხაზით და პატარა წრეწირებით A_i სახის, ხოლო სამკუთხედებით მონიშნული არით A_j სახის რეალიზაციები. ამ შემთხვევაში აუცილებელია თანაკვეთის არის განსაზღვრა, ანუ თანაკვეთაში მოთავსებული რეალიზაციების (წერტილების) ნუსხის შედგენა. ამ ამოცანის გადაწყვეტა შესაძლებელია რანგული კავშირებით კლასტერირების მეთოდში შემუშავებული მეზობლების [1,3], პრინციპის გამოყენებით, რომელიც ემყარება კლასტერის აგების რანგის ცნებას.

მოცემულ შემთხვევაში, კლასტერის აგების რანგი, აღვნიშნოთ r_{ij}-ით, საშუალებას მოგვცემს კლასტერში გაერთიანებული ყველა წერტილისთვის განვსაზღვროთ ის რეალიზაციები, რომელიც მოცემულ წერტილთან იმყოფება r_{ij}-ის ტოლ ან ნაკლებ ჩაკეტილ რანგულ კავშირში.



ნახ.3

ასეთი მეთოდით შევადგენთ ქვეკლასტერებს თითოეული რეალიზაციისათვის და მათგან შევარჩევთ მხოლოდ ისეთებს, რომლებშიც გაერთიანებულია ორივე A_i და A_j სახის რეალიზაციები. ასეთი ქვეკლასტერების გაერთიანება მოგვცემს თანაკვეთის არეს მასში შემავალი რეალიზაციების ნუსხით.

თანაკვეთაში შემავალი A_i სახის რეალიზაციებისათვის განვსაზღვროთ მაქსიმალური მანძილი, რომელიც მე-3 ნახაზზე მოცემული მაგალითისათვის ხორციელდება X_h და X_k წერტილებზე. იგივე ოპერაცია ჩავატაროთ A_j სახის რეალიზაციებისათვის თანაკვეთის არიდან; შესაბამისად, მივიღებთ X_m და X_n წერტილებს. ამ წერტილებიდან შემოვხაზოთ ჰიპერსფეროები რადიუსებით X_kX_h და X_mX_n. X_kX_h რადიუსით შემოსაზღვრული ჰიპერსფეროებისათვის განვსაზღვროთ A_i სახის იმ რეალიზაციების რაოდენობა, რომლებიც მოთავსდა ამ

ჰიპერსფეროებში (ნახ.3). აქ მხედველობაში უნდა მივიღოთ ის გარემოება, რომ რეალიზაციები დათვლილი უნდა იყოს მხოლოდ ერთჯერ, იმ შემთხვევებში, როდესაც ის მოთავსდა ორივე ჰიპერსფეროში. აღვნიშნოთ დათვლილი რეალიზაციების რაოდენობა M_{ji}, ხოლო $X_k X_h$ რადიუსით შემოსაზღვრული ჰიპერსფეროებში გაერთიანებული X_j სახის რეალიზაციების რაოდენობა - M_{ij}-ით. შეცდომით ამოცნობის ალბათობისთვის გვექნება:

$$P_{ij}^{-} = \frac{M_{ij}}{M_i};\ P_{ji}^{-} = \frac{M_{ji}}{M_j} \qquad (10)$$

სადაც, $i, j = \overline{i, I}$, $i \neq j$.

ამოცნობის საიმედოობის შეფასებისათვის გვექნება:

$$P_{ij}^{+} = 1 - P_{ij}^{-};\ P_{ji}^{+} = 1 - P_{ij}^{-}$$

4. დასკვნა

ამოცნობის საიმედოობის შესაფასებლად გამოყენებულია ე.წ. სახეთა გავლენის ზონები, რომლებიც შეიძლება განვსაზღვროთ კლასტერირების შედეგებით. კლასტერირების ალგორითმები უნდა აკმაყოფილებდეს დეტერმინირებულობის პირობებს. ნაშრომში გამოყენებულია რანგული კავშირებით კლასტერირების მეთოდი, მაგრამ განხილულია ამოცნობის საიმედოობის პროგნოზირებისთვის სხვა ალგორითმების გამოყენების შესაძლებლობაც.

ლიტერატურა:

1. Verulava O. Clustering Analysis by “Rank of Links” Method. Transactions of GTU, #3(414), Tbilisi, 1997.
2. Verulava O., Khurodze R. Clustering Analysis and Decision-making by “Rank of Links”. Mathematical Problems in Engineering, 2002, vol. 8(4-5), Machine Intelligence Center #347, Tbilisi.
3. ვერულავა ო., ხუროძე რ. რანგული კავშირების თეორია – ამოცნობის პროცესების მოდელირება. სტუ, თბილისი.

FORECASTING OF RELIABILITY OF RECOGNITION ON THE BASIS OF RESULTS OF CLUSTERING

Verulava Otar, Todua Tea, Verulava Lasha
Georgian Technical University

Summary

There are considered problem of prediction of recognition process on the basis of the results of clustering. Introduced notion of influence zones for patterns (clusters). By means of influence zones might be prediction of possible recognition errors. In this article are considered cases: 1. Influence zones are disjointed; 2. influence zones are intersected (as special case of this is discussed non-compact clusters).

ПРОГНОЗИРОВАНИЕ НАДЕЖНОСТИ РАСПОЗНАВАНИЯ НА ОСНОВЕ ИТОГОВ КЛАСТЕРИЗАЦИИ

Верулава О., Тодуа Т., Верулава Л.
Грузинский Технический Университет

Резюме

Рассмотрена проблема прогнозирования процессов распознавания на основе итогов кластеризации. В частности, введено понятие зоны «влияния образов», с помощью которого определяется возможные ошибки распознавания. Рассматриваются случай, когда зоны влияния не пересекаются, пересекаются и как частный случай последнего, рассматриваются некомпактные кластеры.