მანქანური სწავლების კლასტერიზაციის ალგორითმებში გამომავალი ინფორმაციის ფორმირება

ზურაბ ბოსიკაშვილი, დავით ჭოხონელიძე
საქართველოს ტექნიკური უნივერსიტეტი

რეზიუმე

მანქანური სწავლების ერთ–ერთი დანიშნულებაა რაიმე სისტემაზე დაკვირვება. არსებობს სხვადასხვა ტიპის სისტემები: მათემატიკური, ბიოლოგიური, კომპიუტერული და ა.შ. მისი ერთ-ერთი სახეა ინტელექტუალური სისტემა. ისინი მოიცავენ სხვადასხვა დარგებს. მანქანური სწავლება ინტელექტუალური სისტემის ერთ-ერთი მნიშვნელოვანი ნაწილია, რაც თავისმხრივ შეიცავს ისეთ ანალიზურ საკითხებს როგორიცაა: შემავალი და გამომავალი ინფორმაცია, სისტემის მუშაობის ძირითადი პროცესები და პრინციპები. ასეთი სისტემების მანქანურ სწავლებაში განსაზღვრულია უამრავი სხვადასხვა ტიპის ალგორითმი, რომელთაგან ერთ-ერთია კლასტერიზაცია. მნიშვნელოვანია ვიცოდეთ, როგორ და რა პრინციპით დავადგინოთ გამომავალი ინფორმაცია ასეთი ტიპის ალგორითმებისათვის. მიმდინარე სტატია განიხილავს კლასტერიზაციის ტიპის ალგორითმებისათვის გამომავალი ინფორმაციის ფორმირებას.

საკვანძო სიტყვები: ინტელექტუალური სისტემა. მანქანური სწავლება. კლასტერიზაციის ალგორითმები.

1. შესავალი

ინტელექტუალური სისტემის ერთ-ერთი უმნიშვნელოვანესი ნაწილია მანქანური სწავლება, საშუალებით ხდება აღნიშნულ სისტემაზე გარკვეული დაკვირვებები. ინტელექტუალური სისტემა ჩვეულებრივი სისტემისაგან განსხვავებით ხასიათდება თვითგანვითარებით, ე.ი ახლის ათვისების უნარით, ახალი ინფორმაციის სწავლით. ასეთი ტიპის სისტემები გამოირჩევიან გარკვეული სპეციფიკაციით, კერძოდ ის შეიძლება იყოს დახურული ან ღია ტიპის. დახურული ტიპის სისტემებში უცნობია ის პრინციპები და ალგორითმები, რომლითაც აღნიშნული სისტემა მუშაობს, შესაბამისად დაკვირვებები ხორციელდება მხოლოდ შემავალ და გამომავალ ინფორმაციაზე. მისგან განსხვავებით ღია ტიპის სისტემებში ცნობილია არა მხოლოდ შემავალი და გამომავალი ინფორმაცია, არამედ ის ალგორითმები და პრინციპები, რომელსაც იყენებს აღნიშნული სისტემა. მანქანურ სწავლებაში გამოყენებული ალგორითმები დაჯგუფებულია გარკვეული სახით. ზოგიერთი მათგანია: კლასიფიკაციის, კლასტერიზაციის და ასოციაციის ალგორითმები. როგორც წესი, ასეთი ტიპის ალგორითმები ცალკე აღებული ვერ იძლევა ეფექტურ შედეგს, ამიტომ საჭირო ხდება მათი გარკვეული წესით შერწყმა. ამ ალგორითმების კომბინაციისათვის ერთ-ერთ მნიშნელოვან ფაქტორს წარმოადგენს შემავალი და გამომავალი ინფორმაცია. ერთი ალგორითმის გამომავალი ინფორმაცია შეიძლება წარმოადგენდეს შემავალ ინფორმაციას მეორე ალგორითმისთვის. კონკრეტული ალგორითმის შემთხვევაში აუცილებელია დავადგინოთ გამომავალი ინფორმაციის დაყვანის/წარმოდგენის წესი, რომელსაც შემდეგ უკვე გამოვიყენებთ როგორც:

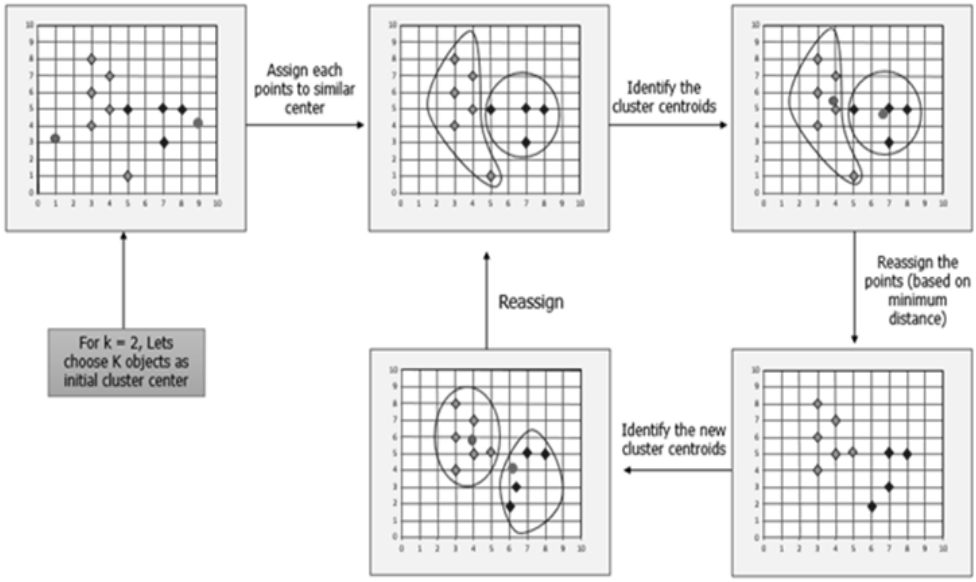
- შემავალი ინფორმაცია სხვა ნებისმიერი ალგორითმისთვის - საჭიროა იმ შემთხვევაში თუ სისტემა აწყობილია გარკვეული ალგორითმების კომბინაციით;
- საბოლოო შედეგების წარმოდგენისთვის.

გამომავალი ინფორმაციის დაყვანის/წარმოდგენის წესი შეიძლება გამოყენებულ იქნას არა მხოლოდ ისეთ სისტემებში სადაც მანქანურ სწავლებაში ჩადებულია ალგორითმების კომბინაცია, არამედ ცალსახად, ნებისმიერი ცალკე აღებული ალგორითმის შემთხვევაში.

2. კლასტერიზაციის ერთ-ერთი ალგორითმი

კლასტერიზაციის ტიპის ალგორითმები მოიცავენ სხვადასხვა ალგორითმებს, რომელთაგანაც ერთ-ერთია K-Means ალგორითმი. მოცემული გვაქვს k კოეფიციენტი და გარკვეულ წერილთა სიმრავლე, რომელთაგანაც უნდა აიგოს კლასტერი. ალგორითმი შეგვიძლია წარმოვადგინოთ როგორც შემდეგი ბიჯების ერთობლიობა:

- დავყოთ ობიექტები k რაოდენობის არა ცარიელ ქვესიმრავლეებად;
- ვიპოვოთ დაყოფილი ქვესიმრავლეების/კლასტერების ცენტროიდები;
- თითოეული წერტილი/ობიექტი მივაკუთვნოთ კონკრეტულ კლასტერს;
- გამოვთვალოთ მანძილები თითოეული წერილიდან და ამ კლასტერისთვის გამოყოფილი სხვა მრავალი წერტილიდან კლასტერამდე, სადაც მანძილი ცენტროიდამდე არის მინიმალური;
- წერტილების გადანაწილების შემდეგ (წინა პუნქტის შედეგად ხდება გარკვეული წერტილების სხვა კლასტერზე მინიჭება, რადგან მის ცენტროიდთან უფრო ახლოს არის ეს წერტილი) ვიპოვოთ ცენტროიდი ახალ კლასტერში [1].



ნახ.1. თითოეული ბიჯი

შეგვიძლია წარმოვადგინოთ შესაბამისი მათემატიკური მოდელიც:

1. მოვახდინოთ კლასტერის ცენტროიდების ინიციალიზაცია;
2. ვიმეორებთ ქვემოთ მოყვანილ ბიჯებს მანამ სანამ არ მივაღწევთ ოპტიმალურ განაწილებას:
   ა. ყოველი i სთვის ვპოულობთ $C^{(i)} := \arg min_j ||x^i - \mu_j||^2$;
   ბ. ყოველი j სთვის კი

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}};$$

ამ მათემატიკური მოდელის საფუძველზე საბოლოო ჯამში ვიპოვით კლასტერებს, სადაც თითოეული მოცემული წერტილი განკუთვნილი იქნება კონკრეტული კლასტერისთვის. ეს წერტილები რა თქმა უნდა განსაზღვრული იქნება რაიმე სიდიდით (მაგ. რიცხვითი სიდიდით), თუმცა უნდა მოხდეს ამ შედეგების გარკვეულ ფორმატში წარმოდგენა რათა შემდეგ ეს

ინფორმაცია გამოყენებულ იქნას როგორც სხვა ალგორითმის შემავალი ინფორმაცია ან/და საბოლოო შედეგების საჩვენებლად/წარმოსადგენად.

3. გამომავალი ინფორმაციის ფორმირება

ზემოაღნიშნული მათემატიკური მოდელის საფუძველზე მივიღებთ კონკრეტულ შედეგებს (მოცემულს რაიმე სიდიდის სახით) თუმცა შედეგების ასეთი პირდაპირი წარმოდგენა ყოველთვის შესაძლოა არ იყოს სწორი. როგორც უკვე აღვნიშნეთ, ინტელექტუალური სისტემა არის უამრავი სახის, მასში თითოეული ალგორითმი შესაძლოა ერთმანეთთან იყოს დაკავშირებული და აქედან გამომდინარე გამომავალი ინფორმაცია შეიძლება გამოყენებულ იქნას როგორც ან სხვა ალგორითმის შემავალ ინფორმაციად ან საბოლოო შედეგების ფორმირებისთვის/წარმოდგენისთვის.

აუცილებელია წინასწარ იყოს ცნობილი რა ტიპის უნდა იყოს გამომავალი ინფორმაცია. ასეთი ტიპები შეიძლება იყოს: ლოგიკური ტიპი (0 ან 1), რიცხვითი ტიპი, რაიმე სხვა ობიექტური ტიპი. ხშირ შემთხვევაში საჭიროა ლოგიკური ტიპის გამომავალი ინფორმაცია. ვთქვათ მოცემულია d კოეფიციენტი (რიცხვითი სახის) და უნდა ვიპოვოთ რაიმე r_i^j ელემენტები (i=0,1....N, j=1....M) მისთვის შეგვიძლია გამოვიყენოთ შემდეგი მიმართებითი დამოკიდებულება:

$$r_i^j = \begin{cases} 0 \; if \; r_i^j \le d \\ 1 \; if \; r_i^j > d \end{cases}$$

აღნიშნულ დამოკიდებულებაში საქმე გვაქვს ორი შესაძლო მნიშვნელობის მიღებასთან (0 ან 1) თუმცა იმავე გამომავალი ინფორმაციის ბინარული სახით ფორმირებისთვის შეგვიძლია გამოვიყენოთ ლოგიკური მიმართების უფრო კომპლექსური მეთოდები, მაგალითად რაიმე ლოგიკური ფორმულა ((A&B)||(A&C)). ასეთი კომპლექსური ლოგიკური ფორმულა შეიძლება მოიცავდეს ერთზე მეტ ოპერანდს და სხვადასხვა მიმართებებს მათ შორის.

რიცხვითი ტიპის შემთხვევაშიც შეიძლება გამოყენებულ იქნას მიმართება იგივე d კოეფიციენტთან მიმართებაში, მაგრამ ეს მიდგომა კარგი იქნება იმ შემთხვევაში თუ მისაღები რიცხვები არის წინასწარ განსაზღვრული სიმრავლის ელემენტები (მაგ. თუ სიმრავლეში არის რიცხვები {1,2,3,7,12}, მიმართებით მიღებული რიცხვიც უნდა იყოს ამ სიმრავლის ელემენტი). იმ შემთხვევაში თუ რიცხვის მნიშვნელობა შეიძლება იყოს ნებისმიერი, აუცილებელია განისაზღვროს ფორმულა, რომელიც მოგვცემს ოპტიმალურ მნიშვნელობას. ასეთი ფუნქციის განსაზღვრა დამოკიდებულია იმაზეც ეს მიღებული ინფორმაცია როგორი ტიპის/კონკრეტულად რომელი ალგორითმისთვის იქნება გამოყენებული როგორც შემავალი ინფორმაცია, რასაც აქ არ განვიხილავთ, რაგდან ის ცილდება მიმდინარე სტატიის კვლევის საგანს.

4. კონკრეტული მაგალითი

განვიხილოთ კონკრეტული მაგალითი. ვთქვათ მოცემულია მონაცემები, რომლებიც შეიცავს ორ ცვლადზე რაიმე რიცხვითი ტიპის ინფორმაციას [2]:

ინდექსი	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0

6	4.5	5.0
7	3.5	4.5

მოცემული მონაცემები დავაჯგუფოთ ორ კლასტერად, პირველად ავიღოთ 1 და მე 4 ინდექსის მქონე სტრიქონები და ვიპოვოთ ცენტროიდები (გამოვიყენოთ ევკლიდეს ფორმულა):

	ინდექსი	ცენტროიდი
ჯგუფი 1	1	(1.0, 1.0)
ჯგუფი 2	4	(5.0, 7.0)

თითოეული კლასტერული ჯგუფისთვის განვსაზღვროთ ინდექსების მიმდევრობები და გამოვთვალოთ ცენტროიდები (ისევ ევკლიდეს მეთოდით):

	კლასტერული ჯგუფი 1		კლასტერული ჯგუფი 2	
ბიჯი	ინდექსები	ცენტროიდი	ინდექსები	ცენტროიდი
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

საწყისი კლასტერული განაწილება შეიცვალა, აქედან გამომდინარე გვექნება:

	ინდექსი	ცენტროიდი
კლასტერი 1	1, 2, 3	(1.8, 2.3)
კლასტერი 2	4, 5, 6, 7	(4.1, 7.0)

ამის შემდეგ საბოლოოდ არ ვართ დარწმუნებულნი რომ თითოეული ობიექტი სწორ კლასტერულ ჯგუფშია, ამიტომ ვადარებთ თითოეული ინდექსის შესაბამისი ობიექტის საკუთარ კლასტერთან მანძილი ნაკლებია თუ არა მეორე კლასტერთან მანძილისა და თუ არა (ე.ი უფრო ახლოსაა მეორე კლასტერთან) უნდა გადავიტანოთ მეორე კლასტერში:

ინდექსი	მანძილი პირველ კლასტერამდე	მანძილი მეორე კლასტერამდე
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

საბოლოო ჯამში თითოეული ინდექსის შესაბამისი ობიექტი მოხვდება იმ კლასტერში, რომელთანაც უფრო ახლოსაა (მანძილი შესაბამის კლასტერამდე არის ნაკლები ვიდრე სხვა კლასტერამდე). აქედან გამომდინარე გვექნება:

	ინდექსი	ცენტროიდი
კლასტერი 1	1, 2	(1.3, 1.5)
კლასტერი 2	3, 4, 5, 6, 7	(3.9, 5.1)

საბოლოო ჯამში იტერაციული გზით მივედით ოპტიმალურ განაწილებამდე, თუმცა არის შემთხვევები როდესაც ასეთი იტერაციების რაოდენობა ან ძალიან დიდია, ან უსასრულოდ გრძელდება. ასეთ შემთხვევაში შეგვიძლია დავეყრდნოთ იტერაციათა კონკრეტულ რაოდენობას, რომელიც მიახლოებით ან ზუსტად არის განსაზღვრული კონკრეტული ამოცანისთვის.

გამომავალი ინფორმაცია საბოლოოდ არის ინდექსების შესაბამისი ობიექტები. ამ შემთხვევაში ვიცით, რომ პირველ კლასტერში არის 1 და 2- ე ინდექსის შესაბამისი ობიექტები, ხოლო დანარჩენი ეკუთვნის მეორე კლასტერს. იმისათვის რომ აღნიშნული მონაცემები გადავიყვანოთ ბინარულ ფორმატში, განვსაზღვროთ მიმართება d კოეფიციენტის მონაწილეობით, კერძოდ:

$$r_i^j = \begin{cases} 0 \; if \; r_i^j \le d \\ 1 \; if \; r_i^j > d \end{cases}$$

თუ d=1.1 მაშინ:

ინდექსი	A	B
1	0	0
2	1	1

5. დასკვნა

აღნიშნული მეთოდოლოგია წარმოადგენს ერთ-ერთ ვარიანტს თუ როგორ შეიძლება ერთი ალგორითმის შედეგების გამოყენებით ინფორმაცია მიეწოდოს მეორე ალგორითმს (თუ სისტემა ისეთნაირადაა მოწყობილი, რომ ერთი ალგორითმი დამოკიდებულია მეორეზე და ა.ს). რა თქმა უნდა შესაძლებელია გამომავალი ინფორმაციის ფორმირების განსხვავებული მეთოდოლოგიის შემუშავება, მაგრამ აღნიშნული მიდგომა წარმოადგენს მცდელობას ინტელექტუალურ სისტემაში (სადაც გამოყენებულია მანქანური სწავლების ალგორითმები) დაიხვეწოს სისტემის წარმადობა და გახდეს უფრო ეფექტური და სრულყოფილი.

ლიტერატურა:

1. http://www.edureka.co/blog/introduction-to-clustering-in-mahout/
2. http://mnemstudio.org/clustering-k-means-example-1.htm

OUTPUT INFORMATION GENERATION IN MACHINE LEARNING CLUSTERING ALGORITHMS

Bosikashvili Zurab, Chokhonelidze David
Georgian Technical University

Summary

One of the main purpose of machine learning is observating the system. There exists many kinds of system: Mathmetical, Biological, Informational system and etc. One kind of this system is intelligence system. Many industry uses these systems. Machine learning is one of the main part of intelligence system which includes such questions: Input and output information, main processes of system. Such systems' machine learning defines many kind of algorithms. One kind of algorithm of this system is clustering. It's important to know how to generate output information for that kind of algorithms. Current article discusses how to generate output information of clustering algorithms.

ФОРМИРОВАНИЕ ВЫХОДНОЙ ИНФОРМАЦИИ В АЛГОРИТМАХ КЛАСТЕРИЗАЦИИ МАШИННОГО ОБУЧЕНИЯ

Босикашвили З., Чохонелидзе Д.
Грузинский Технический Университет

Резюме

Одно из назначений машинного обучения - наблюдение за какой нибудь системой.Существуют разного типа системы:математические, биологические, компьютерные и т. д.Один из видов систем- интеллектуальная система.Эти системы используются в разных отраслях. Машинное обучение является важной частью интеллектуальной системы, которая включает в себя такие вопросы анализа, как : входящая и выходящая информация, основные процессы и принципы работы системы. В машинном обучении такого типа систем определены многие типы алгоритмов, один из которых - кластеризация. Важно знать, как и каким принципом установить выходящую информацию для такого типа алгоритмов. Представленная статья рассматривает формирование выходящей информации для алгоритмов типа кластеризации.