მანქანური სწავლების ალგორითმებში შემავალი და გამომავალი ინფორმაციის ნორმალიზაცია

ზურაბ ბოსიკაშვილი, დავით ჭოხონელიძე
საქართველოს ტექნიკური უნივერსიტეტი

რეზიუმე

მანქანური სწავლების ძირითადი მიზანია რაიმე სისტემაზე დაკვირვება. არსებული სისტემები შეიძლება იყოს მრავალნაირი: მათემატიკური, ბიოლოგიური, კომპიუტერული და ა.შ. მისი ერთ-ერთი სახეა ინტელექტუალური სისტემა. ეს სისტემები მოიცავს სხვადასხვა დარგებს. მანქანური სწავლება ინტელექტუალური სისტემის ერთ-ერთი მნიშვნელოვანი ნაწილია, რაც თავის მხრივ მოიცავს შემდეგ ძირითად ანალიზურ საკითხებს: შემავალი და გამომავალი ინფორმაცია, სისტემის მუშაობის ძირითადი პროცესები და პრინციპები. ასეთი სისტემების მანქანურ სწავლებაში განსაზღვრულია მრავალი სხვადასხვა ტიპის ალგორითმი. მათთვის, ისევე როგორც მთლიანი სისტემისათვის აუცილებელია ინფორმაციის ადექვატურად მიწოდება. სისტემამ ადექვატურ ფორმატში უნდა მოგვაწოდოს გამომავალი ინფორმაციაც. ამ ყოველივედან გამომდინარეობს ინფორმაციის ნორმალიზაციის საჭიროება. მიმდინარე სტატია განიხილავს ინფორმაციის ნორმალიზაციის ერთ-ერთ ალგორითმს.

საკვანძო სიტყვები: ინტელექტუალური სისტემა. მანქანური სწავლება. ნორმალიზაცია.

1. შესავალი

მანქანური სწავლება ინტელექტუალური სისტემის ერთ-ერთი უმთავრესი კომპონენტია. მისი საშუალებით ხდება აღნიშნულ სისტემაზე გარკვეული დაკვირვებები. ასეთი ტიპის სისტემები იყენებს სხვადასხვა ალგორითმებს [1]. გამოყენებული ალგორითმები შეიძლება იყოს სხვადასხვა ტიპის, ასევე შეიძლება წარმოდგენილი იქნას რამდენიმე ალგორითმის კომბინაციაც, ეს ყველაფერი კი მხოლოდ იმიტომ, რომ გაუმჯობესდეს ინტელექტუალური სისტემის წარმადობა. როგორიც არ უნდა იყოს მანქანური სწავლების ალგორითმები, მათ ახასიათებთ ორი ძირითადი ფაქტორი [2,3]:

1. **შემავალი ინფორმაცია** – ნებისმიერი სახის ინფორმაცია, რომელიც ადექვატურად მიეწოდება შესაბამის ალგორითმს;

2. **გამომავალი ინფორმაცია** – ნებისმიერი სახის ინფორმაცია. აუცილებელია აღნიშნული ინფორმაცია ადექვატურად იქნას წარმოდგენილი, რადგან ის შეიძლება გამოყენებულ იქნას როგორც სხვა ალგორითმის შემავალი მონაცემები ან/და საბოლოო შედეგი, რომელზე დაყრდნობითაც ადამიანმა უნდა მიიღოს გარკვეული გადაწყვეტილება.

საწყის ეტაპზე ხდება ინფორმაციის კატეგორიის იდენტიფიკაცია (ინფორმაცია შეიძლება იყოს მრავალი სახის, აქედან გამომდინარე პირველ რიგში აუცილებელია იმის გარკვევა თუ რომელ კატეგორიას ეკუთვნის ის). შემავალი იქნება ინფორმაცია თუ გამომავალი, აუცილებელია ის გარდაიქმნას კონკრეტულ ფორმატში. აღნიშნული ფორმატს გამოიყენებს ალგორითმი ინფორმაციის იდენტიფიკაციისთვის. მონაცემთა წარმოდგენა ისე უნდა განხორციელდეს, რომ საბოლოო შედეგის აღქმა შეეძლოს არა მხოლოდ ინტელექტუალურ სისტემას, არამედ ადამიანსაც, რომელიც იყენებს აღნიშნულ სისტემას გარკვეული ამოცანების შესასრულებლად. კონკრეტულ ფორმატში გარდაქმნა სხვა არაფერია თუ არა მონაცემების ნორმალიზაცია/დენორმალიზაცია, რომელსაც განვიხილავთ. ცალკე საკითხია რამდენად კარგია შერჩეული ნორმალიზაცია/დენორმალიზაციის ალგორითმები თუმცა ზოგადად ნორმალიზაცია/დენორმალიზაციის გარეშე ვერ მოვახდენთ მთლიანი სისტემის მუშაობის ადექვატურობას.

2. მონაცემთა ტიპების ძირითადი კატეგორიები

არსებობს მონაცემთა სხვადასხვა ტიპები, რომლებიც გამოიყენება ნებისმიერი ალგორითმისთვის. შემავალი და გამომავალი ინფორმაცია მიეკუთვნება ამ ტიპებიდან ერთ-ერთს. ძირითადად განსაზღვულია შემდეგი მონაცემთა ტიპები:

ნომინალური – მოიცავს ისეთი ტიპის ინფორმაციას, რომლის ყველა შესაძლო მნიშვნელობა წინასწარ არის ცნობილი, შესაბამისად მიმდინარე მნიშვნელობა აუცილებლად იქნება ამ ჩამოთვლილი მნიშვნელობებიდან ერთ-ერთი. აღნიშნულის კლასიკური მაგალითია სქესი, რომელსაც გააჩნია ორი შესაძლო მნიშვნელობა: მამრობითი ან მდედრობითი.

ორდინალური – მოიცავს ისეთი ტიპის ინფორმაციას, რომლის ყველა შესაძლო მნიშვნელობა წინასწარ არის ცნობილი, მიმდინარე მნიშვნელობა კი ხასიათდება ცვალებადობით რაღაც პირობებიდან გამომდინარე. აღნიშნულის კლასიკური მაგალითია ტემპერატურის სამი შესაძლო კატეგორია: "ცხელი", "თბილი", "ცივი". პირველ ჯერზე მიმდინარე მნიშვნელობა შეიძლება იყოს "ცხელი", გარკვეული პერიოდის/პირობების შედეგად კი მისი მნიშვნელობა შეიძლება შეიცვალოს და გახდეს "თბილი" და ა.შ. აქედან გამომდინარე ძირითადი განსხვავება ნომინალურსა და ორდინალურ ინფორმაციას შორის არის მიმდინარე მნიშვნელობის ცვალებადობა, (რაიმე პირობის/მდგომარეობის გათვალისწინებით) რომელიც ახასიათებს ორდინალური ტიპის ინფორმაციას.

ინტერვალური – მოიცავს რიცხვითი ტიპის ინფორმაციას, რომელსაც გააჩნია საწყისი 0. აღნიშნულის მაგალითია რაიმე წელიწადი, წარმოდგენილი რიცხვით ფორმატში (მაგალითად, 2010 წელი და ა.შ).

ფარდობითი – მოიცავს რიცხვითი ტიპის ინფორმაციას, რომელსაც არ გააჩნია 0 ზე დაბალი ან ტოლი მნიშვნელობა. აღნიშნულის კლასიკური მაგალითია ასაკი (მაგალიტად, 10 წელი, არ არსებობს 0 წელი ან თუნდაც -1 წელი).

როგორიც არ უნდა იყოს მონაცემები, მათ ახასიათებთ გარკვეული ოპერაციების მხარდაჭერა. მაგ: შეკრება, გამოკლება, გამრავლება და გაყოფა. ასევე შეიძლება ახასიათებდეთ მიმართებითი დამოკიდებულებებიც, მაგ: მეტია, ნაკლებია, ტოლია და ა.შ.

მონაცემთა ტიპები **ცხრ.1**

	ნომიწალური	ორდიწალური	იწტერვალური	ფარდობითი
*აწ /	არა	არა	არა	დიახ
+აწ -	არა	არა	დიახ	დიახ
<აწ>	არა	დიახ	დიახ	დიახ
=აწ ! =	დიახ	დიახ	დიახ	დიახ
მაგალითი	გენდერი	ცხელი/თბილი/ცივი	წელიწადი	ასაკი

მოცემულ ცხრილში წარმოდგენილია ოთხი სხვადასხვა მონაცემთა ტიპი თუმცა საბოლოო ჯამში შეიძლება განვსაზღვროთ ორი ძირითადი ტიპი, რომელთაგანაც პირველია რაოდენობრივი, ხოლო მეორე

ხარისხობრივი. რაოდენობრივს განეკუთვნება ყველა ისეთი ტიპის ინფორმაცია, რომელიც რიცხვითი სახით წარმოდგება, ხოლო ხარისხობრივს კი სხვა დანარჩენი. ინფორმაციის ნორმალიზაცია/დენორმალიზაცია გულისხმობს რაიმე კონკრეტული ალგორითმით მიმდინარე ინფორმაციის გარდაქმნას. არსებობს უამრავი სხვადასხვა ალგორითმი, რომელთაგანაც ერთ-ერთია Equilateral Encoding.

3. Equilateral Encoding ალგორითმი

აღნიშნული ალგორითმისათვის გენერირებული ინფორმაცია წარმოდგენილია მატრიცის სახით, რომლის განზომილებებია NxN-1 ზე, სადაც N არის კატეგორიების რაოდენობა (N შეიძლება იყოს 1 ან მეტი, აღნიშნულის განხილვისთვის ვიგულისხმოთ რომ N>1), ხოლო თითოეული სტრიქონი შეიცავს ე.წ encoding ს მიმდინარე კატეგორიისათვის [4]. ალგორითმისათვის აუცილებელია რაიმე წინასწარგანსაზღვრული შუალედი, რომელშიც მოვახდენთ მონაცემთა გრადაციას (როგორც წესი ეს შუალედი არის -1 დან 1 მდე თუმცა შეიძლება აღებულ იქნას რაიმე განსხვავებულიც. აღნიშნულის განხილვისთვის გამოვიყენოთ შუალედი -1 დან 1 მდე). ალგორითმი შეიძლება წარმოვადგინოთ შემდეგი ბიჯების სახით:

1. საწყისი მატრიცის პირველი სტრიქონის პირველ ელემენტში უნდა ჩაიწეროს შუალედის მინიმალური მნიშვნელობა, ხოლო მეორე სტრიქონის პირველ ელემენტში კი შუალედის მაქსიმალური მნიშვნელობა. შესაბამისად გვექნება:

$$result[\![0]\!][\![0]\!] = -1;$$
$$result[\![1]\!][\![0]\!] = 1;$$

2. ვახდენთ იტერაციას მე 2 კატეგორიიდან ზემოთ N მდე. (უგულებელვჰყობთ პირველს, რადგან არ გვაქვს განხილული ისეთი შემთხვევა როდესაც მოცემულია მხოლოდ 1 კატეგორია):

$$for\ k\ from\ 2\ to\ N\ \{$$

3. ვითვლით გრადაციის კოეფიციენტს. ყოველი ახლად აგებული მატრიცის მიღება ხდება წინა მატრიცის გამოყენებით (ათვლა მეორედან დავიწყეთ, მესამე კატეგორიისათვის მატრიცის მიღება შეგვიძლია წინა მატრიციდან, რომელიც უკვე გენერირებულია)

$$f = \frac{sqrt(N * N - 1)}{r};$$

4. ვახდენთ მატრიცის იმ ნაწილზე იტერაციას, რომელიც უკვე გამოვთვალეთ და მოვახდინეთ მისი გრადაცია:

$$for\ i\ from\ 0\ to\ k\ \{$$
$$for\ j\ from\ 0\ to\ k-1\ \{$$
$$result[\![i]\!][\![j]\!] *= f$$
$$\}$$
$$\}$$

5. გამოვთვალოთ მატრიცის ზღვარი (ვექტორ-სვეტების გარკვეული მნიშვნელობები)

$$r = -1/N;$$
$$for\ j\ from\ 0\ to\ k\ \{$$
$$result[\![i]\!][\![k-1]\!] = r$$
$$\}$$

6. მატრიცის ბოლო მნიშვნელობაში ჩავწეროთ 1 და გავაგრძელოთ მე 2 ბიჯზე აღწერილი ციკლი

$$result[\![k]\!][\![k-1]\!] = 1;$$

7. როდესაც ციკლი მთლიანად დასრულდება , უნდა მოვახდინოთ მთლიანი მატრიცის გრადაცია [-1 , 1] შუალედში.

$$dataLow = -1;$$
$$dataHigh = 1;$$
$$for\ row\ from\ 0\ to\ N\ \{$$
$$for\ col\ from\ 0\ to\ N-1\ \{$$
$$result[row][col] = \left(\frac{result[row][col]=dataLow}{dataHigh-dataLow}\right) * (normalizedHigh - normalizedLow) + normalizedLow;$$
$$\}$$
$$\}$$

კონკრეტული კატეგორიის სანახავად (რომელიც უკვე ნორმალიზებულია), უნდა ამოვიღოთ შესაბამისი სტრიქონი მიღებული/გენერირებული მატრიციდან. დეკოდირებისთვის (დენორმალიზაციისათვის) კი უნდა ვიპოვოთ მატრიცაში ისეთი სტრიქონი, რომელსაც გააჩნია უმცირესი ევკლიდური მანძილი გამომავალ ვექტორთან (ნორმალიზებულ ვექტორთან) მიმართებაში.

4. კონკრეტული მაგალითი

ზემოთ აღწერილი ალგორითმის საფუძველზე შეგვიძლია მოვიყვანოთ კონკრეტული მაგალითი: პირობითად ავიღოთ შემდეგი პარამეტრები: კატეგორიების/კლასების რაოდენობა - 5, შუალედი - [-1,1]. ალგორითმის საფუძველზე მიღებული მატრიცა შემდეგნაირად გამოიყურება:

Class	Output 0	Output 1	Output 2	Output 3
Class #1	-0.7905694150420949	-0.4564354645876385	-0.32274861218395134	-0.25
Class #2	0.790569415042095	-0.4564354645876385	-0.32274861218395134	-0.25
Class #3	0	0.912870929175277	-0.32274861218395134	-0.25
Class #4	0	0	0.9682458365518543	-0.25
Class #5	0	0	0	1

ნახ.2. მიღებული მატრიცა (კონკრეტული მაგალითი)

5. დასკვნა

კვლევის საფუძველზე განვიხილეთ ალგორითმი, რომლის დახმარებითაც მოვახდენთ შემავალი და გამომავალი ინფორმაციის ნორმალიზაცია/დენორმალიზაციას. აღნიშნულს ძალიან დიდი ფუნქცია აკისრია ინტელექტუალურ სისტემაში, რადგან ნორმალიზაციის არ არსებობამ ან არასწორმა ნორმალიზაციამ შესაძლოა გამოიწვიოს მთლიანი სისტემის არა ადაექვატური მუშაობა, აქედან გამომდინარე აუცილებელია ნორმალიზაციის გამოყენება და შესაბამისი ალგორითმის სწორად შერჩევა.

ლიტერატურა:

1. Bell J. (2015). Copyright ©John Wiley & Sons, Inc. http://onlinelibrary.wiley.com-/book/10.1002/9781119183464.

2. ბოსიკაშვილი ზ., ჭოხონელიძე დ. (2015). მანქანური სწავლების კლასიფიკაციის ალგორითმებში შემავალი ინფორმაციის ფორმირება. სტუ-ს შრ.კრ. „მართვის ავტომატიზებული სისტემები", №2(20). გვ.31-35.

3. ბოსიკაშვილი ზ., ჭოხონელიძე დ. (2015). მანქანური სწავლების კლასტერიზაციის ალგორითმებში გამომავალი ინფორმაციის ფორმირება. სტუ-ს შრ.კრ. „მართვის ავტომატიზებული სისტემები", №2(20). გვ.36-41.

4. Heaton J. (2013). Artificial Intelligence for Humans Volume 1 Fundamental Algorithms. https://www.youtube.com/watch?v=Ftk0iTdGlN4.

NORMALIZATION OF INPUT AND OUTPUT INFORMATION IN MACHINE LEARNING 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 such system is intelligence system. These systems are used in many industries. 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. Such systems' machine learning defines many kind of algorithms. For them, as well as the entire system it's necessary to adequately supply the information. System must also adequately generate output information. Therefore information normalization is required. The following article discusses one of the algorithms of normalization.

НОРМАЛИЗАЦИЯ ВХОДЯЩЕЙ И ВЫХОДЯЩЕЙ ИНФОРМАЦИИ В АЛГОРИТМАХ МАШИННОГО ОБУЧЕНИЯ

Босикашвили З., Чохонелидзе Д.
Грузинский Технический Университет

Резюме

Основная цель машинного обучения - это наблюдение за какой нибудь системой.Существующие системы могут быть многообразными: математические, биологические, компъютерные и т.д. Один из видов -интеллектуальная система. Эти системы включают в себя разные отрасли. Машинное обучение одна из основных частей интеллектуальной системы, что, в свою очередь, включает в себя аналитические вопросы: входящая и выходящая информации, основные процессы и принципы работы системы. В машинном обучении для таких систем предусмотрены разного типа алгоритмы. Для них, как и для всей системы, обязательным условием является предоставление адекватной информации. Система в адекватном формате должна предоставить и выходящую информацию. Изо всего этого вытекает необходимость нормализации информации. В статье рассматривается один из алгоритмов нормализации информации.