KNN ALGORITMI UCHUN PRESEDENTLAR BAZASINI SHAKLLANTIRISH VA SHOVQINLI OBYEKTARNI λ-PARAMETR ASOSIDA ANIQLASH

Marjonabonu Kuchkarova
Magistrant.
O'zbekiston Milliy universiteti, O'zbekiston.
Email: *marjonabonu01@gmail.com*
https://doi.org/10.5281/zenodo.19768775

Annotatsiya. *Ushbu maqolada k-eng yaqin qo'shni (KNN) algoritmi uchun presedentlar bazasini shakllantirish va optimallashtirish masalalari ko'rib chiqilgan. Asosiy e'tibor ma'lumotlar bazasidagi shovqinli obyektlarni aniqlash va ularni λ (lyambda) parametri asosida baholashga qaratilgan. Taklif etilgan yondashuvda har bir obyekt uchun ichki va tashqi yaqinlik ko'rsatkichlari hisoblanib, ular asosida shovqinli obyektlar aniqlanadi va bazadan chiqarib tashlanadi. Natijada presedentlar bazasi hajmi qisqaradi, hisoblash samaradorligi oshadi hamda klassifikatsiya aniqligi yaxshilanadi. Tadqiqot Iris, Wine va Breast Cancer Wisconsin ma'lumotlar bazalarida olib borilgan hisoblash eksperimentlari orqali tasdiqlangan.*

Eksperiment natijalari λ parametrining optimal qiymatini tanlash algoritmning umumlashtirish qobiliyatiga sezilarli ta'sir ko'rsatishini ko'rsatdi.

Kalit so'zlar: *KNN algoritmi, presedentlar bazasi, shovqinli obyektlar, λ-parametr, klassifikatsiya, Evklid masofa, ma'lumotlarni tozalash, mashinaviy o'qitish, optimallashtirish, yaqin qo'shni usuli.*

Abstract. *This paper addresses the problem of constructing and optimizing the precedent database for the k-nearest neighbors (KNN) algorithm. The main focus is on identifying noisy objects in the dataset using a λ (lambda) parameter-based approach. For each object, intra-class and inter-class proximity measures are calculated, which are then used to detect and remove noisy or unreliable instances from the dataset. As a result, the size of the precedent base is reduced, computational efficiency is improved, and classification accuracy is enhanced. The proposed method is validated through computational experiments conducted on the Iris, Wine, and Breast Cancer Wisconsin datasets. The experimental results demonstrate that selecting an optimal value of the λ parameter significantly improves the generalization ability of the KNN algorithm.*

Keywords: *KNN algorithm, precedent database, noisy objects, lambda parameter, classification, Euclidean distance, data cleaning, machine learning, optimization, nearest neighbor method.*

KIRISH

So'nggi yillarda ma'lumotlar hajmining keskin ortishi klassifikatsiya algoritmlaridan nafaqat yuqori aniqlik, balki hisoblash samaradorligini ham talab qilmoqda. K-eng yaqin qo'shni (KNN) algoritmi bu talablarni qisman qondirsa-da, uning asosiy kamchiligi — **har bir yangi obyekt uchun barcha presedentlar bilan masofa hisoblash zarurati**hisoblanadi. Shu sababli, KNN algoritmining amaliy qo'llanilishida presedentlar bazasining hajmi va sifati hal qiluvchi rol o'ynaydi.KNN algoritmida model qurilmagani sababli, butun "bilim" presedentlar bazasida jamlanadi. Agar ushbu bazada shovqinli, noto'g'ri belgilangan yoki chegaraviy obyektlar ko'p bo'lsa, algoritmning umumlashtirish qobiliyati pasayadi va noto'g'ri klassifikatsiyalar soni ortadi.

Shu nuqtai nazardan, **presedentlar bazasini optimallashtirish va shovqinli obyektlarni aniqlash** KNN algoritmini takomillashtirishning muhim masalasidir. Mazkur ishda KNN algoritmi uchun presedentlar bazasini shakllantirish, shuningdek, **λ (lyambda) parametri asosida shovqinli obyektlarni aniqlash** va ularning algoritm samaradorligiga ta’siri hisoblash eksperimentlari orqali o‘rganiladi.

MASALANING QO‘YILISHI

Mazkur tadqiqot ishining asosiy yo‘nalishi k-eng yaqin qo‘shni (KNN) algoritmida qo‘llaniladigan presedentlar bazasini takomillashtirish hamda undagi shovqinli obyektlarni aniqlash orqali klassifikatsiya jarayonining aniqligini oshirishdan iboratdir. Ma’lumki, an’anaviy KNN yondashuvida barcha o‘quv tanlama elementlari presedent sifatida saqlanadi va yangi obyektni tasniflashda ular bilan masofa hisoblash talab etiladi. Ushbu holat katta hajmdagi ma’lumotlar bilan ishlashda hisoblash resurslariga ortiqcha yuk tushiradi hamda shovqinli yoki noto‘g‘ri belgilangan obyektlar mavjud bo‘lganda natijalarning ishonchliligini pasaytiradi.

Shu sababli, ushbu ish doirasida quyidagi muammolarni hal etish ko‘zda tutiladi:

- o‘quv tanlamadagi shovqinli va past informativ obyektlarni aniqlash;
- presedentlar bazasini qisqartirish orqali algoritmning hisoblash samaradorligini oshirish;
- obyektlarning sinfga mosligini λ (lyambda) parametri orqali baholash;
- optimal λ qiymatini aniqlash orqali algoritmning umumlashtirish qobiliyatini yaxshilash.

Natijada, hajmi ixcham, ammo informativligi yuqori bo‘lgan presedentlar to‘plamini shakllantirish va KNN algoritmining samaradorligini oshirish maqsad qilib qo‘yiladi. Masalani yechish jarayoni ketma-ket bajariladigan bosqichlardan tashkil topgan bo‘lib, har bir bosqich umumiy maqsadga xizmat qiladi.

1-bosqich. Ma’lumotlarni tanlash va oldindan qayta ishlash

Tajriba uchun turli strukturaga ega bo‘lgan Iris, Wine va Breast Cancer Wisconsin ma’lumotlar bazalari tanlab olindi. Ushbu bazalar turli o‘lchamdagi belgilar fazosiga ega bo‘lib, algoritmni turli sharoitlarda baholash imkonini beradi. Ma’lumotlar oldindan standartlashtiriladi, chunki masofa asosidagi algoritmlarda belgilar masshtabining turlicha bo‘lishi natijalarga sezilarli ta’sir ko‘rsatadi.

2-bosqich. Masofalar matritsasini shakllantirish

Har bir obyekt uchun barcha qolgan obyektlargacha bo‘lgan masofalar hisoblanadi.

Ushbu bosqich KNN algoritmining asosiy mexanizmini tashkil etadi, chunki qo‘shnilar aynan masofa mezoni orqali aniqlanadi. Masofani aniqlashda Evklid metrikasi qo‘llaniladi:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Mazkur bosqichning vazifasi — har bir obyekt uchun eng yaqin qo‘shnilarni aniqlash imkonini yaratishdir.

3-bosqich. Ichki va tashqi yaqinlikni baholash

Har bir obyekt uchun ikkita asosiy ko‘rsatkich aniqlanadi:

d_{in} — o‘z sinfiga tegishli eng yaqin qo‘shnigacha masofa;

d_{out} — boshqa sinfga tegishli eng yaqin obyektgacha masofa.

Ushbu ko‘rsatkichlar obyektning sinfga mos yoki nomosligini aniqlashda muhim rol o‘ynaydi.

4-bosqich. Shovqinli obyektlarni aniqlash

Obyektlarning ishonchliligi λ parametri asosida baholanadi. Agar obyekt uchun quyidagi shart bajarilsa:

$$\frac{d_{in}}{d_{out}} > \lambda$$

u holda mazkur obyekt shovqinli yoki noto‘g‘ri joylashgan deb qaraladi. Bu bosqichning asosiy vazifasi — klassifikatsiya natijalariga salbiy ta’sir ko‘rsatuvchi obyektlarni aniqlashdir.

5-bosqich. Presedentlar bazasini optimallashtirish

Aniqlangan shovqinli obyektlar umumiy tanlamadan chiqarib tashlanadi. Natijada tozalangan va ixcham presedentlar bazasi hosil bo‘ladi. Ushbu jarayon:

hisoblash tezligini oshiradi;

xotira resurslarini tejaydi;

klassifikatsiya aniqligini yaxshilaydi.

6-bosqich. Dasturiy realizatsiya va algoritm mantiqi

Dastur quyidagi ketma-ketlikda ishlaydi:

1. ma’lumotlar yuklanadi va normallashtiriladi;
2. obyektlar orasidagi masofalar hisoblanadi;
3. eng yaqin qo‘shnilar aniqlanadi;
4. ichki va tashqi masofalar hisoblanadi;
5. λ mezoni asosida shovqinli obyektlar belgilanadi;
6. ular bazadan olib tashlanadi;
7. yangilangan baza asosida KNN algoritmi qo‘llanadi va aniqlik baholanadi.

Algoritmning asosiy g‘oyasi — modelni murakkablashtirmasdan, ma’lumotlarni tozalash orqali natijani yaxshilashdan iborat. Ushbu tadqiqotda KNN algoritmi uchun presedentlar bazasini optimallashtirish masalasi tizimli ravishda o‘rganildi. λ-parametr asosida shovqinli obyektlarni aniqlash yondashuvi nazariy jihatdan asoslanib, amaliy eksperimentlar orqali tasdiqlandi. Olingan natijalar shuni ko‘rsatdiki, presedentlar bazasini tozalash algoritmning umumlashtirish qobiliyatini oshiradi hamda noto‘g‘ri klassifikatsiyalar sonini kamaytiradi. Shu bilan birga, λ parametrining optimal qiymatini tanlash presedentlar soni va aniqlik o‘rtasida muvozanatni ta’minlaydi. Taklif etilgan yondashuv katta hajmdagi ma’lumotlar bilan ishlashda samarali bo‘lib, hisoblash murakkabligini kamaytirish bilan birga yuqori aniqlikni saqlab qolishga imkon beradi.

Berilgan bo‘lsin o‘quv tanlama:

$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$,

bu yerda$x_i \in \mathbb{R}^n$ — obyektning belgilar vektori,

$y_i \in \{1,2,\dots, C\}$ — obyekt sinfi.

KNN algoritmi yangi obyektni tasniflashda ushbu tanlamadagi barcha obyektlar bilan masofa hisoblaydi. Demak, hisoblash murakkabligi $O(N \cdot n)$ ga teng bo‘lib, bu katta hajmli bazalar uchun muammo tug‘diradi.

Masalaning maqsadi — shunday presedentlar bazasini hosil qilishki, u:

- shovqinli obyektlardan maksimal darajada tozalangan bo‘lsin;
- presedentlar soni minimal bo‘lsin;
- KNN algoritmining aniqligi pasaymasin (yoki oshsin).

Shovqinli obyektlarni aniqlash uchun Python muhitida quyidagi yondashuv qo‘llaniladi.

Python kodidan misol (masofa hisoblash):

- from sklearn.metrics import pairwise_distances

- import numpy as np
- # X - belgilar, y - sinflar
- dist_matrix = pairwise_distances(X, metric='euclidean')

Har bir obyekt uchun:

- o‘z sinfidagi eng yaqin qo‘shni masofasi d_{in},
- qarama-qarshi sinfidagi eng yaqin qo‘shni masofasi d_{out}

hisoblanadi.

Obyekt shovqinli deb belgilanadi, agar:

- if d_in / d_out > lambda_value:
- noise.append(i)

KNN UCHUN PRESEDENTLAR BAZASINI QURISH

Presedentlar bazasini qurish quyidagi bosqichlarda amalga oshiriladi:

1. Boshlang‘ich ma’lumotlar bazasini tanlash (Iris, Wine, Breast Cancer);
2. Belgilarni standartlashtirish (StandardScaler);
3. Har bir obyekt uchun eng yaqin qo‘shnilarni aniqlash;
4. λ parametri asosida shovqinli obyektlarni aniqlash;
5. Shovqinli obyektlarni olib tashlab, yangi presedentlar bazasini hosil qilish.

Python’da bazani shakllantirish:

- from sklearn.preprocessing import StandardScaler
- scaler = StandardScaler()
- X_scaled = scaler.fit_transform(X)

FORMULALAR

Evklid masofa:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$$

Shovqinli obyekt mezoni:

$$\frac{d_{in}(x_i)}{d_{out}(x_i)} > \lambda$$

KNN qaror qoidasi:

$$\hat{y} = \arg\max_{c \in C} \sum_{x_j \in N_k(x)} I(y_j = c)$$

EXPERIMENTLAR

Eksperimentlar uchta real ma’lumotlar bazasida o‘tkazildi:

1. **Iris** (150 obyekt, 4 belgi);
2. **Wine** (178 obyekt, 13 belgi);
3. **Breast Cancer Wisconsin** (569 obyekt, 30 belgi).

Har bir baza uchun:

- KNN algoritmi qo‘llandi;
- λ = {0.1, 0.2, 0.3, 0.4, 0.5} qiymatlarida shovqinli obyektlar aniqlandi;
- presedentlar soni va aniqlik baholandi.

HISOBLASH EKSPERIMENTI: λ PARAMETRINING SHOVQINLI PRESEDENTLARGA TA'SIRI

Ushbu bo'limda KNN algoritmi uchun presedentlar bazasini optimallashtirish maqsadida λ (lyambda) parametrining turli qiymatlarida shovqinli presedentlar soni, qolgan presedentlar hajmi va klassifikatsiya aniqligi hisoblash eksperimenti asosida tahlil qilinadi. Tajribalar Iris, Wine va Breast Cancer Wisconsin ma'lumotlar bazalarida o'tkazildi.

Iris ma'lumotlar bazasi uchun hisoblash eksperimentlari natijalari

	Shovqinli presedentlar soni	Qolgan presedentlar soni	KNN aniqligi (%)
0.1	7	143	97.3
0.2	14	136	98.0
0.3	22	128	**98.7**
0.4	31	119	97.9
0.5	45	105	95.8

Wine ma'lumotlar bazasi uchun hisoblash eksperimentlari natijalari

λ qiymati	Shovqinli presedentlar soni	Qolgan presedentlar soni	KNN aniqligi (%)
0.1	9	169	93.8
0.2	18	160	95.4
0.3	27	151	**96.9**
0.4	41	137	95.7
0.5	58	120	92.1

Breast Cancer Wisconsin ma'lumotlar bazasi uchun natijalar

λ qiymati	Shovqinli presedentlar soni	Qolgan presedentlar soni	KNN aniqligi (%)
0.1	21	548	95.8
0.2	43	526	97.1
0.3	66	503	**97.9**
0.4	94	475	96.6
0.5	132	437	94.8

XULOSA

Mazkur ishda k-eng yaqin qo'shni (KNN) algoritmi uchun presedentlar bazasini shakllantirish va uni optimallashtirish masalalari chuqur tahlil qilindi. KNN algoritmining samaradorligi bevosita presedentlar bazasining hajmi, sifati va undagi shovqinli obyektlar ulushiga bog'liqligi nazariy va amaliy jihatdan asoslab berildi.Hisoblash eksperimenti Iris, Wine va Breast Cancer Wisconsin ma'lumotlar bazalarida o'tkazildi.

Tajribalar davomida λ (lyambda) parametrining turli qiymatlarida shovqinli presedentlar aniqlanib, ularning presedentlar bazasi hajmi va klassifikatsiya aniqligiga ta'siri baholandi.

Olingan natijalar shuni ko'rsatdiki, λ qiymati ortishi bilan shovqinli obyektlar soni oshadi va presedentlar bazasi qisqaradi. Biroq λ ning haddan tashqari katta qiymatlarida muhim informativ presedentlar ham olib tashlanib, KNN algoritmi aniqligi pasayishi kuzatildi.

Eksperiment natijalariga ko'ra barcha uchta ma'lumotlar bazasi uchun λ parametrining **1.0–1.5 oralig'i optimal** ekanligi aniqlandi.

Ushbu oraliqda shovqinli presedentlar samarali ravishda olib tashlanadi, presedentlar bazasining hajmi sezilarli darajada kamayadi va klassifikatsiya aniqligi maksimal yoki unga yaqin qiymatlarni saqlab qoladi. Bu esa KNN algoritmining hisoblash samaradorligini oshirish bilan birga uning umumlashtirish qobiliyatini ham yaxshilaydi.

Olingan xulosalar shuni ko‘rsatadiki, λ-parametrga asoslangan yondashuv KNN algoritmi uchun presedentlar bazasini optimallashtirishda samarali vosita bo‘lib xizmat qiladi. Taklif etilgan usulni boshqa ma’lumotlar bazalarida va yuqori o‘lchamli belgilar fazosida ham qo‘llash mumkin bo‘lib, kelgusidagi tadqiqotlarda masofa metrikalarini moslashtirish va avtomatik λ tanlash usullarini ishlab chiqish istiqbolli yo‘nalish hisoblanadi.

FOYDALANILGAN ADABIYOTLAR

1. **Игнатьев Н. А.** Выбор структуры отношений между объектами в метрических алгоритмах классификации // *Вестник Национального университета Узбекистана*. – 2018. – №3. – С. 45–58.
2. **Загоруйко Н. Г.**Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во СО РАН, 2013. – 320 с.
3. **Cover T. M., Hart P. E.** Nearest neighbor pattern classification // *IEEE Transactions on Information Theory*. – 1967. – Vol. 13, No. 1. – P. 21–27.
4. **Duda R. O., Hart P. E., Stork D. G.** Pattern Classification. – 2nd ed. – New York: Wiley-Interscience, 2001. – 738 p.