

METRIK ALGORITMLARDA OBYEKT LARNING KOMPAKTLIGI VA UMUMLASHTIRISH QOBILIYATI BOG'LIQLIGI

Tursunmurotov Davrbek Xudayorovich

Muhammad al-xorazmiy nomidagi

Toshkent axborot texnologiyalari universiteti

Dasturiy injiniring fakulteti, dasturlash texnologiyalari

kafedrası katta o'qituvchisi.

ANNOTATSIYA

Ushbu maqolada obrazlarni tanish va mashinali o'qitishda keng qo'llaniladigan metrik algoritmlar, xususan NN va KNN algoritmlarining umumlashtirish qobiliyati kompaktlik tushunchasi asosida tadqiq etildi. Klass ichidagi obyektlar yaqinligini tavsiflovchi kompaktlik mezonlari hamda ularning klasslararo masofa bilan bog'liqligi tahlil qilindi. Kompaktlik gipotezasiga tayangan holda, kompakt sinflarda NN va KNN algoritmlarining qaror barqarorligi va aniqlik ko'rsatkichlari o'rganildi. Tajriba sinovlari Iris tanlanmasi asosida o'tkazilib, turli metrikalar uchun kompaktlik baholari va klassifikatsiya aniqligi o'rtasidagi bog'liqlik ko'rsatildi. Olingan natijalar kompaktlik oshishi bilan umumlashtirish qobiliyati yaxshilaydi hamda xatolikni kamaytiradi.

KIRISH

Obrazni tanish (pattern recognition) va mashinali o'qitish (machine learning) sohalarida ob'ektlar o'rtasidagi o'xshashlikni baholash muhim masalalardan biridir. Ushbu vazifani bajarishda **metrik algoritmlar** markaziy o'rin tutadi. Metrik yondashuvlar ob'ektlar orasidagi masofani aniqlash orqali sinflashtirish, regressiya va klasterlash masalalarini samarali yechishga imkon beradi.

Metrik algoritmlar ob'ektlar fazosida aniqlangan masofa funksiyasi asosida ishlaydi. Bu masofa Evklid, Manhattan, Chebyshev yoki maxsus o'rganilgan metrik bo'lishi mumkin. Asosiy g'oya – o'xshash ob'ektlar bir-biriga yaqin joylashadi, turli sinfga mansub obyektlar esa uzoq bo'ladi.

Obrazni tanish masalalarida (tasvir, signal, matn, belgilarni tanish) metrik algoritmlar quyidagi afzalliklarga ega:

- Ob'ektlar o'rtasidagi **strukturaviy va statistik o'xshashlikni** hisobga oladi;
- Belgilar fazosida **geometrik talqin** berishga imkon yaratadi;
- Kichik hajmdagi tanlovlarda ham samarali ishlaydi.

Metrik algoritmlar obrazni tanish va mashinali o'qitish tizimlarining nazariy va amaliy asosini tashkil etadi. Ularning samaradorligi ma'lumotlar fazosining kompaktligi va to'g'ri metrik tanlash bilan chambarchas bog'liq. Shu bois metrik algoritmlar umumlashtirish qobiliyatini tahlil qilish va takomillashtirish bugungi kunda dolzarb ilmiy yo'nalishlardan biri hisoblanadi.

Metrik algoritmlar asosan yaqin qo'shni(NN) va k ta yaqin qo'shni (KNN) algoritmlari hisoblanadi. NN (Nearest Neighbor) va KNN (k-Nearest Neighbors) algoritmlari obrazni tanish va mashinali o'qitishda keng qo'llaniladigan metrik algoritmlar sinfiga kiradi. Ushbu algoritmlar ob'ektlar o'rtasidagi o'xshashlikni baholash uchun belgilar fazosida aniqlangan masofa funksiyasiga tayanadi. Asosiy g'oya shundan iboratki, bir-biriga yaqin joylashgan ob'ektlar odatda bir xil sinfga mansub bo'ladi.

KNN algoritmi NN algoritmining umumlashtirilgan shakli bo'lib, qaror qabul qilishda eng yaqin k ta qo'shni hisobga olinadi. Klassifikatsiyada obyekt sinfi qo'shnilar orasidagi ko'pchilik ovozi asosida aniqlanadi, regressiya masalalarida esa qo'shni qiymatlarning o'rtacha yoki vaznli o'rtacha qiymati olinadi.

KNN algoritmi k parametri orqali model murakkabligini boshqarish imkonini beradi. K ning kichik qiymatlari lokal xususiyatlarni yaxshi aks ettiradi, katta qiymatlari esa umumiy tendensiyalarni hisobga oladi.

Umumlashtirish qobiliyati — bu algoritmnining **o'quv tanlamada ko'rilmagan yangi ob'ektlar** uchun ham to'g'ri qaror qabul qila olish xususiyatidir. Metrik algoritmlarda umumlashtirish qobiliyati belgilar fazosida ob'ektlarning **kompakt joylashuvi** va sinflararo masofaning yetarli darajada ajratilganligi bilan belgilanadi.

NN va KNN algoritmlarida umumlashtirish lokal xarakterga ega bo'lib, qaror qabul qilish yaqin qo'shnilar asosida amalga oshiriladi. Shu sababli umumlashtirish qobiliyati **tanlama zichligi, shovqin darajasi** va **tanlangan metrikaga** kuchli bog'liq.

Empirik xato — bu algoritmnining **o'quv tanlamadagi** xatolar ulushi bo'lib, modelning tanlanmaga qanchalik moslashganini ko'rsatadi. Haqiqiy xato esa algoritmnining **asosiy (noma'lum) taqsimotdan olingan yangi ma'lumotlardagi** xatolar ulushidir. NN algoritmda empirik xato ko'pincha **nolga teng** bo'lishi mumkin, chunki har bir o'quv ob'ekti o'ziga eng yaqin qo'shni hisoblanadi. Biroq bu holat haqiqiy xatoning kichik bo'lishini kafolatlamaydi. KNN algoritmda esa k oshirilishi empirik xatoni biroz oshirishi mumkin, ammo haqiqiy xatoni kamaytirishga yordam beradi.

Me'yordan ortiq o'rganish(Overfitting) — bu algoritmnining o'quv tanlamaga haddan tashqari moslashib, shovqin va tasodifiy tebranishlarni ham o'rganib olishi natijasida umumlashtirish qobiliyatining pasayishidir. NN algoritmi overfittingga eng moyil algoritmlardan biri bo'lib, har bir yangi obyekt bo'yicha qaror bitta qo'shni asosida qabul qilinadi. KNN algoritmda esa k parametrini oshirish orqali qaror silliqashtiriladi va overfitting ta'siri kamaytiriladi. Biroq juda katta k qiymatlari underfittingga olib kelishi mumkin.

Kompaktlik (compactness) — bu bir sinfga mansub obyektlarning belgilar fazosida **bir-biriga yaqin joylashganlik darajasini** ifodalovchi tushunchadir. Agar sinf obyektlari o'zaro kichik masofalarda joylashgan bo'lsa, bunday sinf **kompakt**

deb ataladi. Metrik algoritmlar nuqtayi nazaridan, sinf ichidagi obyektlar yaqinligi qanchalik yuqori bo'lsa, yangi obyekt uchun eng yaqin qo'shnilarni topish shunchalik ishonchli bo'ladi va klassifikatsiya xatosi kamayadi.

Kompaktlik tushunchasi sinflararo ajralish (separability) bilan chambarchas bog'liq. Samarali tanish sharti sifatida quyidagi holat talab etiladi:

- klass ichidagi masofalar **kichik** bo'lishi;
- klasslararo masofalar **katta** bo'lishi.

Agar sinflar ichki jihatdan kompakt bo'lib, bir-biridan yetarlicha uzoq joylashgan bo'lsa, metrik algoritmlar yuqori umumlashtirish qobiliyatiga ega bo'ladi. Aksincha, sinflararo masofa kichik bo'lsa, hatto kompakt sinflar ham chalkashib ketishi mumkin.

Kompaktlik gipotezasi shuni ko'rsatadiki **agar alomatlar fazosida har bir sinf obyektlari kompakt joylashgan bo'lsa va sinflararo masofa yetarli darajada katta bo'lsa, u holda metrik algoritmlar (NN, KNN) past haqiqiy xato bilan umumlashtira oladi.**

Bu gipoteza metrik yondashuvlarning nazariy asosini tashkil etadi va umumlashtirish qobiliyatini tushuntirishda muhim rol o'ynaydi. Kompaktlik gipotezasi shovqin, tanlama hajmi va metrika tanlashning ahamiyatini ham asoslab beradi.

Kompaktlikni miqdoriy baholash uchun klass ichidagi o'rtacha masofa keng qo'llaniladi.

$$Comp(C) = \frac{2}{n(n-1)} \sum_{i < j} \rho(x_i, x_j) \quad (1)$$

C — sinf, $|C| = n$, $\rho(x_i, x_j)$ — masofa funksiyasi.

Markazga nisbatan kompaktlik:

$$Comp(C) = \frac{1}{n} \sum_{i=1}^n \rho(x_i, \mu_C) \quad (2)$$

bu yerda μ_C sinf markazi (o'rtacha vector)

Umumiy tanlama uchun nisbiy kompaktlik

$$Q = \frac{\sum_{k=1}^K Comp(C_k)}{\sum_{k < l} \rho(\mu_{C_k}, \mu_{C_l})} \quad (3)$$

Bu yerda:

surat – sinf ichidagi tarqoqlik

maxraj – sinflararo ajralish

Q qanchalik kichik bo'lsa, tanlanma tuzilishi metrik algoritmlar uchun shunchalik qulay bo'ladi.

Kompaktlik tushunchasi metrik algoritmlarning umumlashtirish qobiliyatini tushuntiruvchi asosiy nazariy omillardan biridir. Klass ichidagi obyektlar yaqinligi va klasslararo masofaning yetarliligi NN va KNN algoritmlarining samaradorligini belgilaydi. Kompaktlikni miqdoriy baholash mezonlari esa tanlama sifatini tahlil qilish va algoritm tanlashda muhim vosita hisoblanadi.

Kompaktlik darajasi oshgan sari sinf ichidagi obyektlar zichligi ortadi, bu esa empirik va haqiqiy xatolar o'rtasidagi tafovutni kamaytiradi. Nazariy jihatdan, agar klass ichidagi maksimal masofa sinflararo minimal masofadan kichik bo'lsa, NN va KNN algoritmlarida klassifikatsiya xatosi nolga yaqinlashadi.

Bu holat umumlashtirish qobiliyatining kuchayishini anglatadi. Ya'ni, tanlama kompakt bo'lsa, metrik algoritmlar o'quv tanlamadan tashqaridagi obyektlar uchun ham to'g'ri qaror qabul qilish ehtimoli yuqori bo'ladi.

Kompaktlik va umumlashtirish qobiliyati o'rtasidagi bog'liqlik metrik algoritmlar nazariyasining asosiy tushunchalaridan biridir. Sinflarning kompaktligi ta'minlangan sharoitda NN va KNN algoritmlari shovqinlarga kam sezgir, qarorlari barqaror va umumlashtirish qobiliyati yuqori bo'ladi. Shu sababli kompaktlikni baholash va oshirish usullarini ishlab chiqish metrik algoritmlarni takomillashtirishda muhim ilmiy vazifa hisoblanadi.

HISOBLASH EKSPERIMENTI

Ushbu tajriba sinovlari Iris tanlanmasi asosida o'tkazildi. Iris tanlanmasi 150 ta obyekt va 3 ta klassdan (*Setosa*, *Versicolor*, *Virginica*) iborat bo'lib, har bir obyekt 4 ta miqdoriy alomat (gulkosa va gulbarg uzunligi hamda eni) bilan tavsiflanadi. Tanlanma mashinali o'qitishda klassifikatsiya algoritmlarini sinash uchun keng qo'llaniladi va kompaktlik hamda umumlashtirish qobiliyatini tahlil qilish uchun qulay hisoblanadi. Tajribada obyektlar orasidagi masofa turli metrikalar (Evklid, Manxetten va Chebyshev) asosida hisoblendi hamda NN va KNN ($k = 5$) algoritmlarining aniqlik ko'rsatkichlari baholandi.

Jadval 1. Kompaktlik va aniqlik o'rtasidagi bog'liqlikni tasvirlash

Metrika	(1) bo'yicha kompaktlik	(2) bo'yicha kompaktlik	(3) bo'yicha kompaktlik	Aniqlik (NN bo'yicha)	Aniqlik (KNN bo'yicha)
Yevklid	1.350	0.944	0.491	95.6	93.3
Manhattan	2.128	1.507	0.402	95.6	93.3
Chebyshev	1.133	0.785	0.709	95.6	95.6

Jadvalda keltirilgan natijalardan ko'rinadiki, metrika turiga qarab kompaktlik qiymatlari sezilarli darajada farqlanadi. Kompaktlik qiymatlari kichik bo'lgan holatlarda NN va KNN algoritmlarining aniqlik ko'rsatkichlari nisbatan yuqori bo'lib, bu kompaktlik va umumlashtirish qobiliyati o'rtasidagi bog'liqlikni

tasdiqlaydi. Ayniqsa, nisbiy kompaktlik (3) kamayishi bilan klassifikatsiya aniqligining barqarorlashuvi kuzatiladi. Bu holat kompakt sinflarda qaror chegaralari kamroq buzilishini va chegaraviy obyektlar ta'siri kamayishini ko'rsatadi.

ADABIYOTLAR RO'YXATI:

1. **Cover T., Hart P.** Nearest neighbor pattern classification. — *IEEE Transactions on Information Theory*, 1967, Vol. 13, No. 1, pp. 21–27.
2. **Duda R. O., Hart P. E., Stork D. G.** *Pattern Classification*. — 2nd ed., Wiley-Interscience, New York, 2001.
3. **Vorontsov K. V.** Лекции по метрическим алгоритмам классификации. — М.: МГУ, 2010.
4. **Bishop C. M.** *Pattern Recognition and Machine Learning*. — Springer, New York, 2006.
5. **Hastie T., Tibshirani R., Friedman J.** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. — 2nd ed., Springer, 2009.
6. **Fix E., Hodges J. L.** Discriminatory analysis. Nonparametric discrimination: Consistency properties. — *USAF School of Aviation Medicine*, Technical Report, 1951.
7. **Devroye L., Györfi L., Lugosi G.** *A Probabilistic Theory of Pattern Recognition*. — Springer, New York, 1996.
8. **Fukunaga K.** *Introduction to Statistical Pattern Recognition*. — 2nd ed., Academic Press, 1990.
9. **Ripley B. D.** *Pattern Recognition and Neural Networks*. — Cambridge University Press, 1996.
10. **Zhang H., Zhou Z.-H.** When does nearest neighbor work well? — *Pattern Recognition*, 2007, Vol. 40, No. 2, pp. 688–702.