

MASHINALI O'QITISHDA METRIK ALGORITMLARNING UMUMLASHTIRISH QOBILIYATINI OSHIRISH

Tursunmurotov Davrbek Xudayorovich

Muhammad al-xorazmiy nomidagi

Toshkent axborot texnologiyalari universiteti

Dasturiy injiniring fakulteti, dasturlash texnologiyalari

kafedrası katta o'qituvchisi

Annotatsiya: Ushbu ishda eng yaqin qo'shni usuli algoritmlarini amalga oshirish xususiyatlarini hisobga olgan holda o'quv tanlanmalarini senzuralash masalasi ko'rib chiqiladi. Senzuralash jarayoni berilgan metrika bo'yicha sinflarning chegaraviy obyektlari to'plamidan foydalanish bilan bog'liq bo'lib, u quyidagi maqsadlarni ko'zlaydi: shovqinli obyektlarni aniqlash va olib tashlash hamda o'quv tanlanmaning bog'langanlik nuqtayi nazaridan klaster tuzilmasini tahlil qilish.

Shovqinli obyektlarni olib tashlash va algoritmlarni o'qitish uchun pretsedentlar bazasini shakllantirishning maxsus shartlari tadqiq etiladi. Bunday baza asosida obyektlarni tanib olish jarayoni boshlang'ich tanlanmaga nisbatan hisoblash resurslari minimal sarflangan holda yuqoriroq aniqlikni ta'minlashi lozim.

Kirish

O'quv tanlanmalarini senzuralash mashinali o'qitish jarayonining muhim qismi sifatida qaraladi. Senzuralashning asosiy maqsadlari algoritmlarning umumlashtirish qobiliyatini oshirish va murakkabligini kamaytirishdan iborat. Ushbu maqsadlarni amalga oshirish, odatda, shovqinli obyektlar va belgilarni aniqlash hamda olib tashlash orqali o'quv tanlanmalarini shakllantirish bilan bog'liq.

Metrik tanib olish algoritmlari uchun senzuralashning o'ziga xos jihatlarini [2] ishda sinf obyektlarining va umuman tanlanmaning kompaktlik o'lchovi orqali ko'rib chiqish taklif etilgan ishda senzuralash jarayoni tibbiyot sohasiga oid ma'lumotlar tanlanmalari misolida diagnostika qilinayotgan obyektlarni korreksiyalashni nazarda tutgan. Korreksiyaning mazmuni ayrim obyektlarni olib tashlash yoki diagnostika qilinayotgan (maqsadli) belgidagi xatolikni tuzatishdan iborat bo'lgan. Ushbu protsedura raqobatli o'xshashlik funksiyasidan foydalanilgan holda, tuzatishlar kiritilishidan oldin va keyin sinf obyektlarining ajraluvchanligi bahosidagi o'zgarishlarni tahlil qilishga asoslangan. Shikastlangan obyektlarni filtrlashdan voz kechish bunday harakatlar o'quv tanlanmaning representativligiga salbiy ta'sir ko'rsatishi mumkinligi bilan asoslangan.

Ishda shovqinli obyektlarning metrik algoritmlarda o'quv tanlanmalari obyektlari o'rtasidagi munosabatlar tuzilmasiga ta'siri mavjudligi qayd etilgan.

Shovqinli obyektlar to'plami berilgan metrika bo'yicha chegaraviy obyektlarning quyi to'plami ekanligi ta'kidlangan. Chegaraviy obyektlarga qiziqish ularni sinf obyektlari orasidagi bog'langanlik munosabatlarini giperkurlar (gipersharlar) tizimi orqali hisoblashda qo'llash bilan bog'liq.

Bog'langanlik munosabati sinf obyektlarini o'zaro kesishmaydigan guruhlar ajratish va ular asosida o'quv tanlanmani etalonlar yordamida minimal qoplashni hisoblash uchun qo'llaniladi. Etalonlar tanlanma obyektlari bilan ifodalanadi.

Baholash uchun ikki xil kompaktlik o'lchovi taklif etilgan:

1. obyektlar munosabatlari tuzilmasini $(0; 1]$ oralig'ida baholovchi o'lchov;
2. tanib olish algoritmlarining umumlashtirish qobiliyatini baholovchi o'lchov.

Birinchi o'lchovning $(0; 1]$ oralig'idagi ruxsat etilgan qiymatlari har bir sinfdagi guruhlar soni va ularning quvvatiga bog'liq. Ikkinchi o'lchov esa minimal qoplashdagi bitta etalon tomonidan "tortiladigan", ya'ni shovqinli obyektlar chiqarib tashlangandan keyin qoladigan tanlanma obyektlarining o'rtacha soni sifatida aniqlanadi. Ushbu o'lchov algoritmlarning umumlashtirish qobiliyatini baholashda mashhur kross-tekshiruv (cross-validation) usuliga muqobil sifatida taklif etiladi.

Shovqinli obyektlar olib tashlangandan so'ng ularning soni va tarkibi chegaraviy obyektlar konfiguratsiyasini, natijada qoplovchi etalonlar to'plamining quvvatini o'zgartiradi. Qoplovchi etalonlar soni o'quv tanlanmaning representativligi ko'rsatkichi bo'lib xizmat qiladi.

Ishda minimal qoplashni izlash jarayoni ochko'z (greedy) algoritm yordamida amalga oshirilgan. Shu sababli, shovqinli obyektlar etalon sifatida tanlanishi va natijada tanib olish algoritmlarining umumlashtirish qobiliyatiga salbiy, ya'ni uni kamaytiruvchi ta'sir ko'rsatishi mumkin.

Tanib olish masalalari noaniq (nekorrekt) masalalar sinfiga mansub bo'lgani sababli, optimal yechim mavjudligi, ya'ni kompaktlik o'lchovi qiymatining maksimumiga erishish masalasi yuzaga keladi. Optimal yechimni izlash minimal qoplashdagi etalonlar soni bilan olib tashlanadigan shovqinli obyektlar tarkibi va soni o'rtasidagi nisbat orqali amalga oshirilgan. Obyektning shovqinli obyektlar to'plamiga mansubligini tekshirish uchun qo'shimcha mezon — regularizatoridan foydalanish taklif etiladi.

Masalaning qo'yilishi

Tanib olish masalasi standart qo'yilishda ko'rib chiqiladi. l ta o'zaro kesishmaydigan $K_1, \dots, K_l$ sinflarga mansub $E_0 = \{S_1, \dots, S_m\}$ obyektlar to'plami berilgan deb hisoblanadi. Obyektlar n xil $X(n) = (x_1, \dots, x_n)$ turdagi belgilar yordamida tavsiflanadi:

O'quv tanlanmani etalonlar yordamida minimal qoplashni shakllantirish jarayoni ishda quyidagi bosqichlarni ketma-ket bajarish orqali amalga oshiriladi:

berilgan $\rho(x,y)$ metrikasi bo'yicha sinflarning chegaraviy obyektlari to'plami $B(E_0, \rho)$ ni ajratib olish;

chegaraviy obyektlar to'plamidan $T \subseteq B(E_0, \rho)$ bo'lgan shovqinli obyektlarni izlash va olib tashlash;

$E = E_0 \setminus T$ to'plamdagi chegaraviy obyektlar asosida sinf obyektlarini bog'langanlik munosabati bo'yicha o'zaro kesishmaydigan guruhlariga ajratish;

har bir guruh bo'yicha etalon obyektlardan iborat minimal qoplashni shakllantirish.

E_0 to'plamni etalonlar bilan minimal qoplashni va formula orqali aniqlanadigan kompaktlik o'lchovi qiymatini hisoblash "eng yaqin qo'shni" usuli algoritmilarini amalga oshirish uchun pretsedentlar bazalarining ko'p xilma-xilligini shakllantirishga asos bo'lib xizmat qiladi. Tanlanma obyektlari o'rtasidagi munosabatlarning o'zgarish shartlari, ularning bir qismi ishda keltirilgan bo'lib, aynan shu holatlar bunday xilma-xillikning yuzaga kelish sababini tushuntiradi.

Hisoblash eksperimenti

O'qitish uchun pretsedent sifatida etalon obyektlarni tanlash samaradorligini tekshirish maqsadida **Spambase** to'plamidagi 4204 ta obyekt ikkita quvvati teng tanlanmaga ajratildi. Bunda har bir sinfdagi obyektlar indekslarining **juft va toq tartibda kelish ketma-ketligidan** foydalanildi.

Har bir tanlanma (**Juft** va **Toq**) navbatma-navbat o'qitish va nazorat uchun qo'llanildi. Ikki tanlanma bo'yicha pretsedentlarni tanlash natijalari **1-jadvalda** keltirilgan.

Jadval1. Yevklid metrikasi bo'yicha olingan natijalar

Pretsedentlar bazasi	Nazorat tanlanma	
	Chet	Nechet
Chet	—	86.54
Nechet	86.39	—

Adabiyotlar ro'yxati:

1. Bishop C. M. *Pattern Recognition and Machine Learning*. – Springer, 2006.
2. Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning*. – Springer, 2009.
3. Cover T., Hart P. Nearest neighbor pattern classification // *IEEE Transactions on Information Theory*, 1967.
4. Vorontsov K. V. *Mashinnoye obucheniye*. – M.: MCCME, 2020.
5. Goodfellow I., Bengio Y., Courville A. *Deep Learning*. – MIT Press, 2016.