



MUHIM OBYEKTlar GURUHINI ANIQLASH UCHUN TIMSOLLARNI ANIQLASHNING METRIK ALGORITMLARI HAQIDA

Yuldashev Sherzod Ismoiljonovich

Toshkent amaliy fanlar universiteti v.b. dotsent (PhD)

Annotatsiya. Ushbu maqola k yaqin qo'shnilar algoritmini ko'plab modifikatsiyalashgan variantlari mavjudligi, uni takomillashtirish jarayonida tarkibiga qo'shimcha parametrlar ham kiritilganligi. k yaqin qo'shnilar algoritmining modifikatsiyalashning bir usuli kam sonidagi tayanch obyektlar tashkil topgan qism to'plamni saralab olish va aynan shu tayanch obyektlar qism to'plami bilan nazorat obyektlarini tanib olishdan iborat ekanligi ko'rib chiqilgan. Tayanch obyektlar qism to'plamini saralab olishning, bir qator afzalliklari keltirib o'tilgan. Timsollarni aniqlashning aksariyat usullari o'quv tanlanmani barcha obyektlaridan emas, balki ma'lum bir qism obyektlaridan tanib olish jarayonida foydalanish jarayonlari keltirilgan hamda Вапник ni "Tayanch vektorlar mashinasi" (Support Vector Machines yoki SVM) misol qilib olingan. Maqolada tanib olishni amaliy masalarini yechishda sinflashtirishning metrik algoritmlaridan va pretsendent axborotlar bilan bog'liqlikni tiklashning kombinatorik nazariyasidan keng foydalangan holda, eng yaqin qo'shnilar algoritmi o'rganib chiqilgan.

Tayanch so'zlar: yaqin qo'shnilar, algoritm, axborot-kommunikatsion texnologiyalar, modifikatsiya, tayanch obyektlar, halaqit obyektlar, nazorat obyektlar, tanlanma, o'qitish masalasi, kompaktlik gipotezasi.

I. KIRISH

Tanib olishni amaliy masalarini yechishda sinflashtirishning metrik algoritmlaridan keng foydalaniladi. Bunday algoritmlar asosida kompaktlik gipotezasi yotadi, ya'ni ma'lum bir toifadagi obyektlar bitta sinfda yotadi. Kompaktlik gipotezasining eng yaqqol ko'rinishi bu eng yaqin qo'shnilar algoritmi (kNN) bo'lib, aksariyat tanib olinayotgan obyekt o'quv tanlanmadagi obyektlarga k yaqinlikda bo'ladi.

Bugungi kunda k yaqin qo'shnilar algoritmini ko'plab modifikatsiyalashgan variantlari mavjud bo'lib, uni takomillashtirish jarayonida tarkibiga qo'shimcha parametrlar ham kiritilgan. k yaqin qo'shnilar algoritmining modifikatsiyalashning bir usuli kam sonidagi tayanch obyektlar tashkil topgan qism to'plamni saralab olish va aynan shu tayanch obyektlar qism to'plami bilan nazorat obyektlarini tanib olishdan iboratdir. Tayanch obyektlar qism to'plamini saralab olish, bir muncha afzalliklarni hosil qiladi:

- halaqit obyektlaridan o'quv tanlanmani tozalash orqali tanib olish sifatini oshiradi;
- nazorat obyektlarini tanib olishda yaqin qo'shnilarni topish orqali amalga oshirish mashina vaqtini qisqartiradi. Bu yondashuv juda katta xajmdagi ma'lumotlarni tanib olishda muhim ahamiyatga ega;
- o'quv tanlanmani siqilgan (qisqartirilgan) ifodasini shakllantirish tayanch obyektlar qism to'plamida qolgan obyektlarni sinfning vakili sifatida qoraladi.

Shunga e'tibor berish kerakki, ayrim obyektlarni o'quv tanlanmadan chiqarib yuborish axborotlarni yo'qotish degani emas. Bu jarayon o'quv tanlanma to'liq shakllantirilguncha davom etadi.

II MASALANING QO'YILISHI

Timsollarni aniqlash usullari va metrik algoritmlar. Timsollarni aniqlashning aksariyat usullari o'quv tanlanmani barcha obyektlaridan emas, balki ma'lum bir qism obyektlaridan tanib olish jarayonida foydalanadi. Bunga Вапник ni "Tayanch vektorlar mashinasi" (Support Vector Machines yoki SVM) yaqqol misol bo'ladi. SVM – da tayanch obyektlar sinfni chegara qismidagi obyektlar

hisoblanadi. O'quv tanlanma uchun tayanch obyektlarni tanlash turli algoritmlar uchun turlichadir. Sabab, har bir algoritmnı o'ziga xos ishlash tamoyiliga egaligidir. Masalan, SVM uchun tayanch hisoblangan obyektlar k yaqin qo'shnilar algoritmi uchun informativ bo'lmagan obyektlar hisoblanadi. Chunki, k yaqin qo'shnilar algoritmidagi o'quv tanlanma obyektlari sinf markazida yoki markaz yaqinida joylashgan obyektlardir.

Ayni vaqtda metrik algoritmlar orasida STOLP, λ -STOLP, FRIS- STOLP kabi algoritmlar vujudga keldi. STOLP – metrik klassifikatorlar uchun etalon obyektlarni saralab beruvchi degan ma'noni anglatadi. Bu turdagi algoritmlar xasis (жадный) strategiyaga asoslangan bo'lib, tayanch obyektlarni ketma – ketlikda qo'shib boradi. Algoritmnı har bir qadamida tayanch obyektlar e'lon qilib boriladi. Sinfni chegarasida joylashgan obyektlar o'z – o'zidan ma'lumki, SVM algoritmgı tayanch algoritm bo'lib xizmat qiladi.

Mazkur maqolada moslashuvchan nazoratni amalga oshirishda kombinatorika formulalaridan samarali foydalanish mezonı va tanlanmani kompaktlik profili tushunchasi taklif qilingan. Taklif qilinayotgan algoritm STOLP toifasidagi algoritmlar bilan solishtirishda qarama – qarshi tomonda turadi. Dastlab, bu algoritm tanlanmadagi xalaqıt beruvchi obyektlarni chiqarib tashlash orqali funksionalni ish faoliyatini yaxshilaydi. Undan so'ng unchalik informativ bo'lmagan obyektlarni tanlanmadan chiqarib yuboriladi. Bu jarayon funksionalni qiymati turg'un holga kelguncha davom ettiriladi. Jarayon nihoyasida o'quv tanlanmadagi obyektlar uch toifaga bo'linadi: halaqıt beruchshi, informativ bo'lmagan va tayanch obyektlar.

Moslashuvchan nazoratli kombinatorik funksional algoritmnı o'qıtish bilan xarakterlanadi [1]. Shuning uchun saralangan tayanch obyektlar yordamida tanib olishni amalga oshiruvchi algoritm sinflashtirish sifatini bir muncha oshiradi.

Yirik xajmdagi axborotlarni sinflashtirishning amaliy masalalarida juft – jufti bilan obyektlarni qiyoslashni tezkor va aniq amalga oshirishga xizmat qiladi.

Sinflashtirishning metrik algoritmlari uchun obyektlarlarni muhimlik darajasiga nisbatan filtratsiya qilish orqali o'quv tanlanmani shakllantirish, tayanch obyektlarni saralab olishda kompaktlik profilini tahlil qilish imkoniyatiga ega kombinatorik baholarni umumlashtirish orqali amalga oshirish. Xususiyl holda eng yaqin qo'shnilar algoritmidan foydalaniladi.

Timsollarni aniqlash masalasini yechishdan avval quyidagicha belgilashlar kiritib olamiz. X obyektlar to'plami, Y javoblar to'plami va A algoritmlar to'plami. Ayrim hollarda A algoritmlarni $y^*: X \rightarrow Y$ ko'rinishda ham ifodalanadi. $y^*(x)$ akslantirish qiymati chekli miqdordagi obyektlar to'plamiga tegishlidir va buni o'quv tanlanma deb yuritiladi. O'quv tanlanmadagi i -obyekt javobini $y_i = y^*(x_i)$ orqali ifodalanadi.

Pretsedent bo'yicha o'qıtish masalasi $a^* \in A$ algoritm uchta talab asosida qurish bilan yakunlanadi.

Birinchi, o'quv tanlanmaga oldindan berilgan javob qiymatlari bilan berilgan bo'lishi shart, ya'ni $a^*(x_i) = y_i, i = \overline{1, l}$. Bu yerda tenglikni konkret masalaga yaqinlashuvchi yoki aniq masala deb tushunish kerak. Bu turdagi talabni lokal chegaralangan deb nomlanadi, ya'ni chekli miqdordagi o'quv tanlanma va samarali chekli miqdordagi qadamlar bilan amalga oshirilishidir.

Ikkinshi, a^* algoritm umumiy xususiyatiga qo'shimcha cheklanishlarni kiritadi, qaysiki, X ni Y akslantirishni qanoatlantirishi shart. Masalan, simmetriklik, uzluksizlik, sillqlik, monotonlik kabi cheklanishlar bo'lishi mumkin.

Uchinchi, qidirilayotgan a^* algoritm umumlashtirish qobiliyatiga ega bo'lishi zarur, u holda nafaqat y^* ni o'quv tanlanmaga bog'liqligini tiklash mumkin, balki X to'plamni barcha elementlariga bog'liqlikni o'rnatish mumkin. Ushbu talabni turli sifat funksionallari orqali shakllantirish mumkin.

III. YECHISH USULI

O'quv tanlanmani shakllantirish. O'quv tanlanmani shakllantirish jarayoni qurilayotgan algoritmnı o'qıtish deb yuritiladi. Uni tadbiqi yangi obyektı sinflashtirishda tanib olish deb ataladi. O'quv tanlanmadagi obyektlarni pretsedent deb ataladi.

I-ta'rif. $X^p = \{x_1, x_2, \dots, x_p\} \subset X$ tanlanma bo'yicha a algoritmning xatoliklar chastotasi quyidagicha hisoblanadi:

$$v(a, X^p) = \frac{1}{p} \sum_{i=1}^p I(x_i, a(x_i))$$

bu yerda $I(x, y)$ –xatoliklar indikatori 1 qiymat qabul qiladi, agar javob x obyekt uchun xatolik mavjud bo'lsa, aks hollarda "0" qiymat qabul qiladi. Indikatorni tanlash aniq masaladan kelib chiqilgan holda amalga oshiriladi. Bunda eng avval Y to'plamni xususiyatiga bog'liq. Sinflashtirish masalasida Y yakuniy to'plam bo'lib, obyektning tushadigan sinfini raqamini belgilab beradi.

II-ta'rif. $a = \mu(X^l)$ algoritmgaga mos keluvchi l uzunlikdagi X^l tanlanmadan ixtiyoriy ravishda olingan obyektning o'z o'rniga qo'yishni ta'minlovchi μ akslantirishga o'qitish usuli deb yuritiladi. Bir nechta $A \subseteq A_u$ algoritmlar oilasidan tanlab olgan holda μ aniq bir algoritmnini qurish usuli hisoblanadi.

Agar $v(a, X^l) = 0$ bo'lsa, a algoritmi tanlanmaga korrekt deyiladi.

Agar $\mu(X^l)$ algoritmi X^l da korrekt bo'lsa, u holda $v(\mu(X^l), X^l) = 0$ tenglik o'rinli bo'ladi. Bu esa μ usulni X^l tanlanmada korrekt deyiladi.

O'quv tanlanmadagi kam chastotadagi $v(\mu(X^l), X^l)$ xatoliklari umumiy holda kafolatlamaydi. Chunki, qurilgan algoritmi X to'plamni o'quv tanlanmadan tashqaridagi obyektlarida kam xato qiladi. Umumlashtirish qobiliyatiga ega usul o'quv tanlanma tarkibiga kirmaydiganlar orqali baholash imkoniyatiga ega. Shuning uchun, moslashuvchan nazorat tamoyiliga asoslanib, sifat funksionali kiritiladi.

l va q - ixtiyoriy fiksirlangan miqdorlar hamda $X^L = (x_1, x_2, \dots, x_L)$ berilgan bo'lsin, bu yerda $L = l + k$. X^L tanlanmani l va k uzunligidagi o'quv va nazorat tanlanmalariga mos holda mumkin bo'lgan barcha guruhlash (X_n^l, X_n^k) orqali ifodalanadi, $n = \overline{1, N}$.

IV. NATIJALAR VA NAMUNALAR

Moslashuvchan nazoratning turfa xilligi bosh to'plamni guruhlashni amalga oshirish imkoniyati bilan bog'liq.

Agar to'plamni guruhlashtirish bitta elementli bo'lsa, u holda gap alohida test tanlanma haqida ketadi (hold-out estimate),

$$Q_{HO}(\mu, X^l, X^k) = v(\mu(X^l), X^k)$$

Q_{HO} funksionalni kamchiligi saralashni o'quv va nazorat tanlanmalarga bog'lanib qolganligidir.

Moslashuvchan nazoratda yetarlicha keng tarqalgan funksional ***k-fold***. Bu funksional ishga tushganda X^L tanlanmani bir xil quvvatdagi k ta qism to'plamlarga ajratadi yoki saralaydi. Har bir qism to'plamdan algoritmi nazorat tanlanma sifatida foydalanadi, qolgan ($k-1$) qism to'plam o'quv tanlanma sifatida qaraladi va bu algoritmnini xatolik chastotasini baholashga xizmat qiladi. Funksional natijalar k ta algoritmlarni har birining o'rtacha qiymati sifatida qaraladi.

Moslashuvchan funksional nazoratni aniqligini yanada oshirish uchun X^L tanlanmani l va k uzunlikdagi mumkin bo'lgan barcha o'quv va nazorat tanlanmalarga ajratish zarur. Bunda vujudga keladigan guruhlashtirishlar soni $N = C_L^l$ teng. Bundan kelib chiqib, moslashuvchan nazorat funksionali quyidagicha ko'rinishga ega bo'ladi:

$$Q_C^{lk}(\mu, X^L) = \frac{1}{N} \sum_{n=1}^N v(\mu(X_n^l), X_n^k)$$

Q_C^{lk} funksionalni kamchiligi k ning kichik miqdorlarida ham amalda sinash murakkab masaladir. Ma'lumki, [2] da Ваппик-Червоненкис bir xil baholar bilan o'z sinfiga kirish nazariyasi va [3] da funksionalni yuqori bahosi nazariyalarida keltirilgan. Biroq, [2,3,4] da keltirilgan nazariya ham amaliy masalaga jiddiy bog'lanib qolgan. Funksionalni maxsus xususiyl holda analitik aniq ifodasini shakllantirish mumkin. Bu formulani k yaqin qo'shnilar va uning modifikatsiya qilingan variantlari uchun samarali hisoblash imkoniyatini vujudga keltiradi.



V. XULOSA

Ushbu maqolada tanib olishni amaliy masalalarini yeshishda sinflashtirishning metrik algoritmlaridan va pretsedent axborotlar bilan bog'liqlikni tiklashni kombinatorik nazariyasi keng foydalangan xolda, eng yaqin qo'shnilar algoritmi o'rganib chiqilgan.

FOYDALANILGAN ADABIYOTLAR RO'YXATI

1. Воронсов К.В. Комбинаторные отценки качества обучения по претцедентам // Докл. РАН.- 2004. -Т.394, №2. –С.175-178.
2. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. –М.: Наука, 1979.
3. Воронцов К.В. Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики. -2004.-Т 477, № 4. –С. 231-238.
4. Загуройко Н.Г. Прикладные методы анализа данных и знаний. –Новосибирск: ИМ СО РАН, 1999.