**O‘ZBEKISTON MILLIY UNIVERSITETI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
DSc.03/30.12.2019.FM.01.02 RAQAMLI ILMIY KENGASH**

O‘ZBEKISTON MILLIY UNIVERSITETI

TURSUNMUROTOV DAVRBEK XUDAYOROVICH

KOMPAKTLIK O‘LCHOVLARI BO‘YICHA OBYEKTLAR MUNOSABATLARI TUZILMALARINING TAHLILI

05.01.11 – Raqamli texnologiyalar va sun’iy intellekt

**FIZIKA MATEMATIKA FANLARI
BO‘YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI
AVTOREFERATI**

Toshkent – 2025

O‘UK: 519.95

Fizika-matematika fanlari bo‘yicha falsafa doktori (PhD) dissertatsiyasi avtoreferati mundarijasi

Оглавление автореферата диссертации доктора философии (PkhD) по физико-математическим наукам

Contents of dissertation abstract of doctor of philosophy (PhD) on physical-mathematical sciences

O‘ZBEKISTON MILLIY UNIVERSITETI
HUZURIDAGI ILMIY DARAJALAR BERUVCHI
DSc.03/30.12.2019.FM.01.02 RAQAMLI ILMIY KENGASH

O‘ZBEKISTON MILLIY UNIVERSITETI

TURSUNMUROTOV DAVRBEK XUDAYOROVICH

KOMPAKTLIK O‘LCHOVLARI BO‘YICHA OBYEKTLAR MUNOSABATLARI TUZILMALARINING TAHLILI

05.01.11 – Raqamli texnologiyalar va sun’iy intellekt

FIZIKA MATEMATIKA FANLARI BO‘YICHA FALSAFA DOKTORI (PhD) DISSERTATSIYASI AVTOREFERATI

Toshkent – 2025

Fizika-matematika fanlari bo‘yicha falsafa doktori (Doctor of Philosophy) dissertatsiyasi mavzusi O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi huzuridagi Oliy attestasiya komissiyasida №B2025.1.PhD/FM1271 raqam bilan ro‘yxatga olingan.

Dissertatsiya Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universitetida bajarilgan.

Dissertatsiya avtoreferati uch tilda (o‘zbek, rus, ingliz (rezyume)) Ilmiy kengash veb-sahifasi (http://ik-fizmat.nuu.uz/) va “Ziyonet” ta’lim axborot tarmog‘ida (www.ziyonet.uz) joylashtirilgan.

Ilmiy rahbar:	**Ignatev Nikolay Aleksandrovich** fizika-matematika fanlari doktori, professor
Rasmiy opponentlar:	**Kabulov Anvar Vasilovich** texnika fanlari doktori, professor **Xudaybergenov Qabul Qadirbergenovich** fizika-matematika bo’yicha PhD, dotsent
Yetakchi tashkilot:	**Raqamli texnologiyalar va sun’iy intellektni rivojlantirish ilmiy-tadqiqot instituti**

Dissertatsiya himoyasi O‘zbekiston Milliy universiteti huzuridagi DSc.03/30.12.2019.FM.01.02 raqamli Ilmiy kengashning “___”_____ 2025-yil soat____ dagi majlisida bo‘lib o‘tadi (Manzil: 100174, Toshkent sh., Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (+99871) 227-12-24, faks: (+99871) 246-53-21, 246-02-24, e-mail: nauka@nuu.uz).

Dissertatsiya bilan O‘zbekiston Milliy universitetining Axborot-resurs markazida tanishish mumkin (___ raqami bilan ro‘yxatga olingan). Manzil: 100174, Toshkent sh., Olmazor tumani, Universitet ko‘chasi, 4-uy. Tel.: (+99871) 246-02-24.

Dissertatsiya avtoreferati 2025-yil “____” _____kuni tarqatildi.
(2025-yil “____” _____ dagi _____ raqamli reyestr bayonnomasi).

M.M.Aripov
Ilmiy darajalar beruvchi ilmiy
kengash raisi, f.-m.f.d., professor

Z.R.Raxmonov
Ilmiy darajalar beruvchi ilmiy
kengash ilmiy kotibi, f.-m.f.d., professor

D.T.Muxamediyeva
Ilmiy darajalar beruvchi ilmiy
kengash huzuridagi ilmiy seminar
raisi, t.f.d., professor

KIRISH (falsafa doktori (PhD) dissertatsiyasi annotatsiyasi)

Dissertatsiya mavzusining dolzarbligi va zarurati. Jahon miqyosida olib borilayotgan ko‘plab ilmiy-amaliy tadqiqotlar, hozirgi kunda axborot texnologiyalarini rivojlanishida berilganlarni intellektual tahlil (BIT) qilish usullari muhim ekanligini ko‘rsatmoqda. BIT sonli algoritmlarini ishlab chiqishda katta berilganlar muammolari, o‘lcham la’nati tufayli obyektlar o‘rtasidagi munosabatlarni yuvilib ketishi, texnik qurilmalarda amalga oshirishning kombinator murakkabligi, shovqin obyektlar va alomatlar mavjud bo‘lgan holatlarda maqbul vaqt ichida natijani olishning imkonsizligi bilan bog‘liq muammolarga duch kelinadi. Berilganlarni tahlil qilish va anglash algoritmlarini tanlashni asoslash vositalaridan biri kompaktlik o‘lchovidir. O‘rgatuvchi tanlanma obyektlari o‘rtasidagi turli munosabatlarni ushbu o‘lchovlarning qiymatlarini tadqiq qilish anglash algoritmlarining umumlashtirish qobiliyatini oshirish masalalari tobora dolzarb bo‘lib bormoqda. Mazkur yo‘nalishlarda samarali yondashuvlar ishlab chiqish berilganlarni intellektual tahlil qilish oldida turgan muhim vazifalaridan biri bo‘lib qolmoqda.

Hozirgi kunda jahonda berilganlarda turli xatoliklar paydo bo‘lish xavfi mavjud bo‘lganda berilganlar asosida olingan xulosalarning sifatini sezilarli darajada yomonlashishini oldini olish va aniqlangan qonuniyatlar ishonchliligini oshirishga imkon beradigan usullarni yaratishga alohida e’tibor qaratilmoqda. Bunda, amaliy masalalarni yechishda berilganlarni tahrirlash va tozalash modellari muhim rol o‘ynaydi. Shu sababli, o‘rgatuvchi tanlanma obyektlari o‘rtasidagi turli munosabatlarda ushbu o‘lchovlarning qiymatlarini kuzatish va izohlash, anglash algoritmlarining umumlashtirish qobiliyatini oshirish maqsadli ilmiy tadqiqotlardan hisoblanadi.

Mamlakatimizda axborot texnologiyalari, sun’iy intellekt va katta berilganlarni tahlil qilish sohalarida ilmiy va amaliy tadqiqotlarga alohida e’tibor qaratilmoqda. Berilganlarni intellektual tahlil qilishning samarali usullarini ishlab chiqish, ularni amaliyotda qo‘llash va umumlashtirish qobiliyatini oshirish bugungi kunda dolzarb masalalardan biri hisoblanadi. So‘nggi yillarda berilganlarning intellektual tahlili usullari va kriteriyalarini ishlab chiqish, tibbiy tashhislashda qabul qilinayotgan qarorlar samaradorligini oshirish uchun berilganlardan yashirin qonuniyatlarni izlash uslublarini ishlab chiqishda salmoqli natijalarga erishildi. Sun’iy intellektni mashina yordamida o‘qitish uchun katta hajmda davlat tilidagi raqamli ma’lumotlarni shakllantirish, shuningdek, davlat tilidagi nutqni tahlil va sintez qilishni qo‘llovchi dasturiy mahsulotlarni ishlab chiqish “Raqamli O‘zbekiston — 2030” Strategiyasida ustuvor yo‘nalishlardan etib belgilandi[1]. Qaror ijrosini ta’minlashda o‘rgatuvchi tanlanma obyektlarining zichlik va kompaktlik o‘lchovlarini hisoblash orqali “yaqin qo‘shnilar” turidagi usullarni ishlab chiqish, obyektlar xususiyatlarini baholash uchun k yaqin qo‘shnilarning optimal sonini tanlash metodikasini ishlab chiqish muhim ahamiyatga ega hisoblanadi.

[1] O‘zbekiston Respublikasi Prezidentining 2021-yil 17-fevraldagi “Sun’iy intellekt texnologiyalarini jadal joriy etish uchun shart-sharoitlar yaratish to‘g‘risida”gi PQ-4996-sonli qarori.

O‘zbekiston Respublikasi Prezidentining 2017-yil 7-fevraldagi PF-4947-sonli "O‘zbekiston Respublikasini yanada rivojlantirish bo‘yicha Harakatlar strategiyasi to‘g‘risida", 2019-yil 8-oktyabrdagi PF-5847-sonli "O‘zbekiston Respublikasining 2030-yilgacha bo‘lgan kompleks ijtimoiy-iqtisodiy rivojlanish konsepsiyasi to‘g‘risida"gi farmonlarida, shuningdek, O‘zbekiston Respublikasi Prezidentining 2017-yil 17-fevraldagi PQ-2789-sonli "Fanlar akademiyasi faoliyatini yanada takomillashtirish, ilmiy-tadqiqot ishlarini tashkil etish, boshqarish va moliyalashtirish chora-tadbirlari to‘g‘risida", 2018-yil 27-apreldagi PQ-3682-sonli "Innovatsion g‘oyalar, texnologiyalar va loyihalarni amaliyotga tatbiq etish tizimini yanada takomillashtirish chora-tadbirlari to‘g‘risida"gi qarorlarida, shuningdek, 2019-yil 24-may kuni O‘zbekiston Milliy universitetida Prezidentning fan va ta’lim sohasi vakillari bilan uchrashuvdagi nutqi va boshqa normativ-huquqiy hujjatlarda belgilangan vazifalarni amalga oshirishda ushbu dissertatsiya tadqiqoti muayyan darajada xizmat qiladi.

Tadqiqotning respublika fan va texnologiyalari rivojlanishining ustuvor yo‘nalishlariga mosligi. Dissertatsiya respublika fan va texnologiyalar rivojlanishining IV. “Matematika, mexanika va informatika” ustuvor yo‘nalishi doirasida bajarilgan.

Muammoning o‘rganilganlik darajasi. Berilganlarni intellektual tahlil qilish (BIT) usullaridan foydalanish nazariyasi va amaliyotini rivojlantirishga mashhur xorijiy va mahalliy olimlar katta hissa qo‘shgan. BIT sohasida tadqiqot olib borgan xorijiy olimlar orasida Yu. I. Jurovlev, K. V. Vorontsov, V.N. Vapnik, N. G. Zagoruyko, A. B. Petrovkiy, R.E.Bellman, J.Goodfellow, E.Nadaraya, G.S.Watson va K. V. Rudakovni qayd etish mumkin. BIT sohasida ishlayotgan mahalliy olimlar orasida M. M. Kamilov, Sh. F. Madrahimov, F. T. Adilova, T.F.Bekmuratov, N.S.Mamatov, D. T. Muhammadieva va Sh. X. Fazilov va boshqalarning ishlarida tadqiq qilingan.

O‘rgatuvchi tanlanmada shovqin obyektlar va sachratqilar mavjud bo‘lgan holatlarda anglash algoritmlarini o‘rgatish jarayoni turlicha bo‘ladi. Qarorlar daraxtlarini qurilishida sachratqi obyektlar ishlatilganligi sababli past statistik ishonchlilikka ega daraxt ostilari o‘chiriladi. Boshqa algoritmlarda berilganlarga dastlabki ishlov berish ko‘zda tutilgan bo‘lib bu jarayonda qandaydir mezonlar asosida shovqin obyektlar aniqlanadi va filtrlanadi. Ayrim holatlarda sachratqini tipik obyektga o‘girish maqsadida ayrim alomatlarni sozlashga harakat qilinadi. Berilganlarni bunday turdagi shovqinlardan tozalash uchun konkurent o‘xshashlik funksiyasidan foydalanishga asoslangan yondashuv taklif etiladi. Tanlanmaning ajratilish bahosi maksimal qiymatga erishganda yoki qolgan tanlanmada sachratqi obyektlar yo‘q bo‘lganda, senzuralash jarayoni to‘xtatiladi. Umumlashtirish qobiliyati muammosi mashinali o‘rgatishda asosiy masalalardan biri hisoblanadi. Algoritmlarni o‘rgatish jarayonida cheklangan miqdordagi pretsedentlarga asoslangan noma’lum bog‘liqlik tiklanadi. Yangi pretsedentlardan tashkil topgan tanlanmani test tanlamasida algoritmning aniqligini bashorat qilish talab etiladi.

Optimal murakkablikdagi algoritmlarni tanlashni baholash mezonlarini izlash davom etmoqda. Yaqin qo‘shnilar toifasidagi usullar uchun bunday baholarni olish

usullaridan biri kompaktlik mezoni bo‘yicha sifat funksionalidan foydalanish hisoblanadi. Mashinali o‘rgatish katta hajmdagi berilganlar bazasini ishlov berish bilan bog‘liqligi sababli, ular uchun dastlab qayta ishlov berish usullarini tanlash muhim ahamiyatga ega. Dastlabki ishlov berish natijalari bo‘yicha kutilgan natija bu o‘rgatuvchi tanlanmani obyektlar shuningdek alomatlar bo‘yicha seleksiya hisoblanadi. Dastlabki ishlov berish tasniflash algoritmlarini qo‘llashga sezilarli ravishda ta’sir qilishi mumkin. Berilganlarning intellektual tahlili usullari murakkab tizimlarni adaptiv boshqarishda optimal qarorlarni tanlashda kompaniyalar va davlat idoralari tomonidan keng qo‘llanilmoqda.

Dissertatsiya tadqiqotining dissertatsiya bajarilgan oliy ta’lim muassasasining ilmiy tadqiqot ishlari bilan bog‘liqligi. Dissertatsiya tadqiqoti Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universitetiining ilmiy-tadqiqot ishlari rejalariga muvofiq «HBV-infeksiyasi sababli kelib chiqadigan jigar sirrozi va gepatokartsinomaga oldindan tashxis qo‘yishning matematik modelini yaratish» doirasida bajarilgan.

Tadqiqotning maqsadi o‘rgatuvchi tanlanma obyektlarining zichlik va kompaktlik o‘lchovlarini hisoblash orqali “yaqin qo‘shnilar” turidagi usullarni ishlab chiqish va ularni asoslashdan iborat.

Tadqiqotning vazifalari quyidagilardan iborat:

sinflar kompaktlik o‘lchovlarining ekstremumini izlash uchun mezonlar regulyarizatorlardan foydalangan holda, o‘rgatuvchi tanlanmaning etalonlar bilan minimal qoplamasidan pretsedentlar bazasini shakllantirish usullarini ishlab chiqish;

“yaqin qo‘shni” usullari algoritmlarini umumlashtirish qobiliyatini oshirish uchun sinflarning chegaraviy obyektlari o‘rtasidagi nisbiy chekinish bo‘yicha shovqin obyektlarni tanlashning zarur va yetarli shartini aniqlash;

obyektlar xususiyatlarini baholash uchun *k* eng yaqin qo‘shnilarning optimal sonini tanlash metodikasini ishlab chiqish;

obyektlarni ularning kesishuvchi gipersharlar tizimi bo‘yicha bog‘langanligi asosida guruhlarga ajratish metodikasini asoslash.

Tadqiqot obyekti. Berilganlarni intellektual tahlil qilish usullarini ishlab chiqish va asoslash hamda ulardan bilimlarga asoslangan axborot modellarini shakllantirishda qo‘llash.

Tadqiqot predmeti. Pretsedentlar bo‘yicha qaror qabul qilish masalalarida o‘rgatuvchi tanlanmani seleksiya qilish va obyektlarni guruhlash uchun kompaktlik va taqsimot zichligi o‘lchovlarini hisoblash usullari.

Tadqiqot usullari. Tadqiqot ishida diskret matematika, amaliy statistika, berilganlarni intellektual tahlil qilish usullari hamda algoritmik tillarda dasturlash texnologiyalaridan foydalanildi.

Tadqiqotning ilmiy yangiligi quyidagilardan iborat:

o‘rgatuvchi tanlanma etalonlari yordamida minimal qoplama masalasini hal qilish asosida pretsedentlar bazasini shakllantirish usuli yaratilgan;

sinflarning chegaraviy obyektlari orasidagi nisbiy chekinish kattaligi bo‘yicha shovqin obyektlarni tanlash uchun zarur va yetarli shartlar ishlab chiqilgan;

obyektlarni ularning taqsimot zichligiga asoslangan holda kategoriyalarga ajratishda, kesishuvchi gipersharlar tizimi orqali bog‘liqlik munosabati yordamida klasterlash sifatini baholash uchun kompaktlik o‘lchovini qo‘llash zarurligi isbotlangan;

sinflar chegarasida joylashgan shovqin obyektlarni aniqlash va ularni olib tashlash uchun yangi regulyarizator ishlab chiqilgan, mezon sifatida o‘rgatuvchi tanlanmaning etalonlar yordamida minimal qoplanishiga asoslangan kompaktlik o‘lchovini maksimal qiymatidan foydalanish usuli yaratilgan.

Tadqiqotning amaliy natijalari quyidagilardan iborat:

pretsedentlar bazasini shakllantirish orqali anglash algoritmlari murakkabligini kamaytirish va umumlashtirish qobiliyatini oshirish uchun tanlanma etalonlari yordamida minimal qoplama masalasini hal etish usuli ishlab chiqilgan.

shovqin obyektlarni aniqlash va olib tashlash, shuningdek, informativ alomatlarni tanlash bu tanlanmadagi ortiqcha ma’lumotlarni kamaytirib, modelning aniqligi va barqarorligini hamda samaradorligini oshirish uchun kompaktlik o‘lchoviga asoslangan usul ishlab chiqilgan.

Tadqiqot natijalarining ishonchliligi natijalarni izchil solishtirish, mavzuga oid matematik apparat va nazariy yondashuvlardan foydalanish, matematik modellarni qo‘llashning qat’iyligi bilan asoslangan.

Tadqiqot natijalarining ilmiy va amaliy ahamiyati. Tadqiqot natijalarining ilmiy ahamiyati anglash algoritmlarining umumlashtirish qobiliyatini oshirish uchun berilganlarni intellektual tahlil qilishning yangi usullari ishlab chiqilganligi bilan izohlanadi.

Tadqiqot natijalarining amaliy ahamiyati tanlanmadan etalon obyektlarni aniqlash, shovqin obyektlarni ajratib olish hamda informativ alomatlarni kompaktlik o‘lchovi asosida tanlash metodologiyasi ishlab chiqilganligi bilan izohlanadi.

Tadqiqot natijalarining joriy qilinishi. Dissertatsiya ishida ishlab chiqilgan kompaktlik o‘lchovlari bo‘yicha minimal qoplama etalonlarini shakllantirish, obyektlar munosabatlari tuzilmalarining tahlili, kesishuvchi gipersharlar bo‘yicha guruhlarga ajratish ko’rinishdagi ilmiy natijalar asosida:

kompaktlik o‘lchovlari bo‘yicha obyektlar munosabatlarini tahlil qilish metodologiyasidan IL-52421091471 “Suv resurslarini monitoring qilish uchun apparat-dasturiy vositalarni ishlab chiqish” innovatsion loyihasida murakkab tizim jarayonlarini tahlil qilishda foydalanilgan (Toshkent axborot texnologiyalari universitetining 2025-yil 24-martdagi 1027/05-2-sonli ma’lumotnomasi). Ilmiy natijalarning qo‘llanilishi hisoblash resurslarini kamaytirish va aniqlikni oshirish orqali samaradorlikni baholash imkonini bergan;

kesishuvchi gipersharlar tizimlari bo‘yicha tanlanma obyektlarini guruhlarga ajratish algoritmidan IL–7823051524 “PARATRANSLATOR parallel korpusi asosida kontekstual tarjima uchun lug‘aviy platforma ishlab chiqish” innovatsion loyihasida so‘zlarni semantikasiga ko‘ra guruhlarga ajratishda foydalanilgan (Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universitetining 2025-yil 3-maydagi 04/11-5787-sonli ma’lumotnomasi). Ilmiy natijalarning qo‘llanilishi klaster tahlil qilish

natijalariga ko‘ra, atamalarning predmet sohalarga semantik yaqinligi ko‘rsatildi. Buning natijasida ma’no jihatdan yaqin so’zlarni aniqlash imkonini bergan.

Tadqiqot natijalarining aprobatsiyasi. Mazkur tadqiqot natijalari 6 ta, jumladan, 2 ta xalqaro va 4 ta respublika ilmiy-amaliy anjumanlarida muhokamadan o‘tkazilgan.

Tadqiqot natijalarining e’lon qilinganligi. Tadqiqot mavzusi bo‘yicha jami 14 ta ilmiy ish chop etilgan, shulardan, O‘zbekiston Respublikasi Oliy Attestatsiya komissiyasining doktorlik dissertatsiyalari asosiy natijalarini chop etish tavsiya etilgan ilmiy nashrlarda 6 ta maqola, jumladan, 2 tasi xorijiy, 1 tasi scopus bazasida indekslangan va 4 tasi respublika jurnallarida nashr etilgan hamda xalqaro va respublika ilmiy amaliy konferensiyalarda 8 ta tezis chop etilgan. Shuningdek, yaratilgan kompyuter dasturiy mahsulotlari uchun 2 ta mualliflik guvohnomasi olingan.

Dissertatsiyaning tuzilishi va hajmi. Dissertatsiya kirish qismi, uchta bob, xulosa, foydalanilgan adabiyotlar ro‘yxati va ilovalardan tashkil topgan. Dissertatsiyaning hajmi 107 betdan iborat.

DISSERTATSIYANING ASOSIY MAZMUNI

Kirish qismida dissertatsiya mavzusining dolzarbligi va muhimligiga asoslangan, tadqiqotning O‘zbekiston Respublikasida ilm-fan va texnologiyalar rivojlanishining ustuvor yo‘nalishlariga muvofiqligi belgilangan, dissertatsiya mavzusiga oid xorijiy ilmiy tadqiqotlar sharhi va muammoning o‘rganilgan darajasi keltirilgan, maqsad va vazifalar shakllantirilgan, tadqiqot obyekti va predmeti aniqlangan, ilmiy yangiligi va tadqiqot natijalarining amaliy natijalari bayon qilingan, olingan natijalarning nazariy va amaliy ahamiyati ochib berilgan, tadqiqot natijalarini joriy qilish, chop etilgan ishlar va dissertatsiya tuzilishi to‘g‘risida ma’lumotlar keltirilgan.

I - bobda “**anglashning metrik algoritmlari**” sinf obyektlarini kompaktlik gipotezasi asoslash bilan bog‘liq savollar tadqiq qilinadi.

§1.1 da **yaqin qo‘shni (NN) va *k* yaqin qo‘shni (KNN)** algoritmlari qaralgan. NN bilan bog‘liq muammolar - to‘plamning minimal qoplamasi texnologiyasi bo‘yicha umumlashtirish qobiliyati nuqtayi nazaridan optimal sondagi shovqin obyektlarni izlash va pretsedentlar bazasini shakllantirish hisoblanadi. KNN ga nisbatan tadqiqotlar esa yaqin qo‘shnilar soni k ni tanlash, metrikalar, baholar, obyektlar turg‘unligi va informativ alomatlarni tanlashga qaratilgan.

Sinflar obyektlarining bog‘langanlik munosabatlari §1.2 da ko‘rilib, sinf obyektlarini o‘zaro kesishmaydigan guruhlarga bo‘lishda qo‘llanilgan. Guruh obyektlari kompaktlik o‘lchovlarini hisoblashda qo‘llaniladi.

Masalaning qo‘yilishi

Standart anglash masalasi qaraladi. Obyektlarning $E_0=\{S_1,\ldots,S_m\}$ to‘plami berilgan bo‘lib, u l ($l > 2$) o‘zaro kesishmaydigan $K_1,\ldots,K_l$ to‘plam ostilariga (sinfga) ajratilgan deb hisoblanadi. Obyektlar n ta turli toifadagi $X(n)=(x_1,\ldots,x_n)$ alomatlar bilan tavsiflanib, ularning ξ tasi interval shkalada (n - ξ) tasi nominal shkalalarda o‘lchanadi. Obyektlarning E_0 to‘plamida $\rho(x,y)$ metrika berilgan.

Obyektlarning E_0 to‘plamida $\rho(x, y)$ metrika bo‘yicha aniqlangan sinflarning chegaraviy obyektlarining to‘plami ostisini $L(E_0,\rho)$ orqali belgilaymiz. Tanlanmaning $S_i,S_j \in K_t$, obyektlari o‘zaro bog‘langan $(S_i \leftrightarrow S_j)$ hisoblanadi agar $\{S \in L(E_0,\rho) | \rho(S,S_i) < r_i$ и $\rho(S,S_j) < r_j\} \neq \varnothing$ bo‘lsa, bu yerda $r_i(r_j)$ $\rho(x,y)$ metrika bo‘yicha $S_i(S_j)$ dan CK_t ($CK_t = E_0 \setminus K_t$) ga eng yaqin obyektgacha bo‘lgan masofa $t=1,\ldots,l$ bir biri bilan bog‘langan hisoblanadi, agar quyidagi shart bajarilsa:
bu yerda.

$G_{tv} = \{S_{v_1},\ldots,S_{v_c}\}$, $c \geq 2$, $G_{tv} \subset K_t$, $v < |K_t|$ to‘plam bog‘langan obyektlar to‘plami bo‘lib, u K_t sinfidagi bog‘langan obyektlarni ifodalaydi, agar ixtiyoriy $S_{v_i}, S_{v_t} \in G_{tv}$ obyektlar uchun yo‘l mavjud bo‘lsa $S_{v_i} \leftrightarrow S_{v_k} \leftrightarrow \ldots S_{v_l}$. Obyekt $S_i \in K_t$, $t=1,\ldots,l$ bir elementdan iborat guruhga tegishli bo‘lib va u bog‘lanmagan deb hisoblanadi, agar har bir $S_j \neq S_i$ и $S_j \in K_t$ obyekt uchun $S_i \leftrightarrow S_j$ yo‘li mavjud bo‘lmasa. Har bir K_t, $t=1,\ldots,l$ sinfi uchun bog‘langan va bog‘lanmagan obyektlar to‘plamidan eng kichik kesishmaydigan guruhlar sonini aniqlash talab qilinadi.

Minimal guruhlar sonini aniqlashda, bog‘langan va bog‘lanmagan obyektlar sinflari uchun $\rho(x,y)$ metrika bo‘yicha berilgan $L(E_0,\rho)$ – chegaraviy obyektlar to‘plam ostisini va obyektlarni binar fazoda tasvirlashdan foydalaniladi. Sinf qobiqlarini ajratish uchun har bir $S_i \in K_t$, $t=1,\ldots,l$ obyekt uchun $\rho(x,y)$ metrika bo‘yicha tartiblangan ketma-ketlik quriladi.

$$S_{i_0}, S_{i_1}, \ldots, S_{i_{m-1}}, S_i = S_{i_0}. \quad (1)$$

Deylik $S_{i_\beta} \in CK_t$, S_i ga K_t sinfga kirmaydigan (1) dan eng yaqin obyekt. $O(S_i)$ orqali markazi S_i da bo‘lgan $r_i = \rho(S_i, S_{i_\beta})$ radiusdagi barcha $\rho(S_i, S_{i_\tau}) < r_i$, $\tau = 1,\ldots,\beta$-1 obyektlarni belgilaymiz. $O(S_i)$ da doim bo‘sh bo‘lmagan obyektlar to‘plam ostisi mavjud

$$\Delta_i = \left\{ S_{i_\alpha} \in O(S_i) \,|\, \rho(S_{i_\beta}, S_{i_\alpha}) = \min_{S_{i_\tau} \in O(S_i)} \rho(S_{i_\beta}, S_{i_\tau}) \right\} \quad (2)$$

(2) bo‘yicha sinf qobiq obyektlariga tegishliligi $L(E_0,\rho) = \bigcup_{i=1}^{m} \Delta_i$ da aniqlanadi.

Obyektlar o‘rtasidagi bog‘lanish munosabatlarining xususiyatlari quyidagilardir:

- E_0 tanlanma faqat bitta va aniq sonli kesishmaydigan obyektlar guruhlariga bo‘linadi;
- biron guruhdagi ixtiyoriy S_i, S_j obyektlar o‘rtasida har doim zanjir qurish mumkin $S_i \leftrightarrow S_k \leftrightarrow \ldots \leftrightarrow S_c$.

Obyektlarni bog‘laydigan zanjirlarni xilma-xil to‘plamlarini $R(S_i,S_j)$ orqali belgilaymiz $S_i, S_j \in G \cup K_2$, $|G| \geq 2$. Eng qisqa yopiqmas yo‘lni(EQYY) tanlash funksionalni optimallashtirish sifatida qaraladi

$$z(S_i, S_j) = \min_{\rho(S_i,S_j)} \sum_{S_u,S_v \in R(S_i,S_j)} \rho(S_u, S_v). \quad (3)$$

Sinf obyektlarini o‘zaro kesishmaydigan guruhlarga bo‘lishdan maqsadlari quyidagilar:

- sinflar va to‘liq tanlanma obyektlarining kompaktligini hisoblash va tahlil qilish;
- o‘rgatuvchi tanlanmani obyektlar - etalon bilan minimal qoplamasini izlash;
- klaster tahlilning sifatini baholash.

Quyidagi keltirilgan 1- jadvalda “*Australian*” berilganlarida kompaktlikka qo‘shgan hissasi va eng qisqa yopiqmas yo‘l (EQYY) bo‘yicha (3) guruhdagi masofalarni hisoblash bo‘yicha eksperiment natijalari keltirilgan.

1-jadval. Dastlabki alomatlar fazosida berilganlarni tahlili natijalari

Metrika	Guruhdagi obyektlar soni	Sinf nomeri i =	Kompaktlikni hissasi $\frac{\lvert G\rvert^2}{\lvert K_i\rvert^2}$	EQYY da hajmi
Evklid	18	1	0,1258	883,18
	13	1	0,0656	1108,33
	32	2	0,1604	281,54
	10	2	0,0156	227,33
	15	1	0,0873	1384,88
Juravlyov	22	1	0,1724	441,82
	23	1	0,1884	98686,39
	145	2	0,8941	2582,58
	37	2	0,0582	652,24
	9	2	0,0034	682,09

1- jadval tahlil natijalariga bo‘yicha quyidagi xulosalarni qilish mumkin

- Kompaktlikka qo‘shilgan hissaning qiymatlari to‘g‘ridan-to‘g‘ri guruhdagi obyektlar soniga bog‘liq;
- EQYY uzunligi guruhlardagi obyektlar soniga bevosita bog‘liq emas. Masalan, K_1 sinfdagi (1-jadvalga qarang) 18 va 13 ta obyektdan iborat ikki guruh uchun kompaktlikka qo‘shilgan hissa qiymatlari mos ravishda 0,1258 va 0,0656 ga teng, EQYY uzunligi esa mos ravishda 883,18 va 1108,33 ga teng bo‘ldi.

Kompaktlik o‘lchovlari va ularni hisoblash usullari §1.3da tavsiflangan. Taklif etilgan kompaktlik o‘lchovi va algoritmlarning umumlashtirish qobiliyati o‘rtasidagi umumiylik, o‘rgatuvchi tanlanmadagi obyektlar to‘plamida klaster tuzilmalarni aniqlash va qo‘llashdan iborat bo‘ladi. O‘lchov qiymatining yagonaligi sinf obyektlarini o‘zaro kesishmaydigan guruhlar usuli bilan kafolatlanadi.

O‘rgatish sifati nuqtayi nazaridan klaster tuzilmani qo‘llashga talab quyidagilardan iborat:

– shovqin obyektlarni aniqlash va ularni o‘chirish;

– tanlanmani sinflarga korrekt bo‘lishni ta‘minlaydigan, shovqin obyektlarsiz tanlanmaning obyektlar etalonlar bilan minimal qoplamasini ajratish.

Aytaylik, o‘rgatuvchi tanlanmani bog‘langanlik munosabati bo‘yicha har bir $K_i \subset E_0$ sinf uchun, o‘zaro kesishmaydigan $G_{i1},...,G_{i\mu}$ guruhlarga bo‘lingan, bunda sinflardan $D_i \subset K_i$ shovqin obyektlari o‘chirilgan bo‘lsin. Shovqin obyektlarni o‘chirishdan maqsad – anglash algoritmlarining umumlashtirish qobiliyatini oshirishdir.

Quyidagicha belgilash kiritamiz $m_{ij} = \left|G_{ij}\right|$, $j = 1,...,\mu$, $\sum_{j=1}^{\mu} m_{ij} = m_i$. Sinfni o‘zaro kesishmaydigan guruhlarga bo‘lish natijalarini ularning soni, vakilligi (obyektlar soni bo‘yicha) va shovqin obyektlarni o‘chirishni hisobga olgan holda tahlil qilish uchun kompaktlik bahosi kabi strukturaviy xususiyatdan foydalanishni taklif qilinadi

$$\Theta_i = \frac{\sum_{j=1}^{\mu} m_{ij}^2}{m_i^2}. \qquad (4)$$

Ma’lumki , (4) bo‘yicha Θ_i ni mumkin bo‘lgan qiymatlari $\left[\frac{1}{m_i};1\right]$ intervalda yotadi.

Agar G_{i1} guruh $K_i \cap \left(E_0 \setminus \bigcup_{j=1}^{l} D_j\right)$ dan tashkil topgan bo‘lsa, u holda $\Theta_i = 1$. O‘rgatuvchi tanlanmaning umumiy kompaktligini o‘rtacha baholash, umuman olganda quyidagi formula bilan aniqlanib

$$R\left(E_0 \setminus \bigcup_{i=1}^{l} D_i, \rho\right) = \left(\frac{\left|E_0 \setminus \bigcup_{i=1}^{l} D_i\right|}{m}\right) \frac{\sum_{i=1}^{l} m_i \Theta_i}{\left|E_0 \setminus \bigcup_{i=1}^{l} D_i\right|} = \frac{\sum_{i=1}^{l} m_i \Theta_i}{m} \qquad (5)$$

hisobdan chiqarilgan shovqin obyektlarning $\left(\frac{\left|E_0 \setminus \bigcup_{i=1}^{l} D_i\right|}{m}\right)$ hissasini inobatga olgan holda amalga oshiriladi. (4) va (5) qiymatlari o‘rgatuvchi tanlanma tuzilmasining bir xilligi (yoki bir xil emasligi) haqida vositali ma’lumot beradi. Guruhlar tarkibidagi obyektlar soni bo‘yicha o‘xshashlik qanchalik yaqin bo‘lsa, (4) qiymati $\frac{1}{m_i}$ ga, (5) esa $\frac{l}{m}$ ga shunchalik yaqinlashadi. $R\left(E_0 \setminus \bigcup_{i=1}^{l} D_i, \rho\right) = 1$ bo‘lsa $E_0 \setminus \bigcup_{j=1}^{l} D_j$ obyektlar guruhlarining soni sinflar soniga teng bo‘ladi. (4) va (5) bo‘yicha qiymatlar to‘plami mos ravishda $\left[\frac{1}{m_i};1\right]$ va $\left[\frac{l}{m};1\right]$ da sinflar va tanlanmaning umumiy kompaktlik o‘lchovi sifatida ko‘rib chiqilishi taklif etiladi. Yuqorida

ko‘rsatilgan intervallardagi kompaktlik o‘lchovlari qiymatlari ma’lumotlar bazalaridagi yashirin qonuniyatlarni aniqlash uchun ishlatilishi mumkin. Tasniflangan obyektlar to‘plami va anglash algoritmlari to‘plami o‘rtasidagi bog‘liqlik kompaktlik gipotezasi shaklida taxmin qilinadi. Kompaktlik tushunchasini noaniq talqin qilish odatda obyektlar o‘rtasidagi yaqinlik o‘lchovlarini tanlash va alomatlar fazosini o‘zgartirish bilan bog‘liq. Kompaktlikni o‘lchash uchun bir ma’noli talqinni ta’minlaydigan birliklar (o‘lchovlar) tanlash masalasi dolzarbdir.

Obyektning turg‘unligi tushunchasi tabiiy ravishda *k* eng yaqin qo‘shnilar asosida optimal yaqinlikni hisoblash protsedurasidan kelib chiqadi va bu holda u obyektning o‘zi bilan bir xil sinfdagi obyektlar orasidagi joylashuvini tavsiflovchi sifat ko‘rsatkichi sifatida tushuniladi. Obyektning λ_i^j turg‘unligi $S_i \in K_j, j = \overline{1,l}$, $i = \overline{1,m}$, K_j sinfda quyidagicha aniqlanadi:

$$\lambda_i^j = \frac{d_i^j}{2\min\limits_{1\le j\le l}\left|K_j\right| - 3}\ , \qquad (6)$$

bu yerda d_i^j — E_0 dagi S_i ga eng yaqin bo‘lgan *k* ta obyektlar soni hodisalar soni, $k = 1,...,2\min\limits_{1\le j\le l}\left|K_j\right| - 3$, ko‘pchiligi asosan $K_j \cap E_0$ sinfiga tegishli obyektlar tashkil qiladi. Ko‘rinib turganidek, $\lambda_i^j \subset [0,1]$.

Obyektlarning turg‘unligi informativ alomatlar to‘plamlarini tanlash, obyektlar orasidagi yaqinlik o‘lchovlarini aniqlash va alomatlar fazosini o‘zgartirish uchun foydalanish mumkin. Turli tuzilishga ega berilganlar mavjudligi muammosini hal qilishda mumkin bo‘lgan yondashuvlardan biri — berilganlarni o‘zgartirish orqali ularni berilgan algoritmga moslashtirishdir. Shubhasiz, berilganlarni shunday o‘zgartirishni topish kerakki, undan so‘ng ular anglash algoritmlari modeli talablariga maksimal darajada moslashadi.

2-jadvalda *Australian* tanlanmasi asosida turg‘unlik ko‘rsatkichi va kompaktlikning o‘zaro bog‘liqligini ko‘rib chiqamiz. Qavs ichida obyektlarning o‘zaro kesishmaydigan guruhlari soni ko‘rsatilgan.

2 - jadval . Dastlabki berilganlarda turg‘unlik va kompaktlik ko‘rsatkichlari

Metrika	Sinfdagi turg‘unlik		Sinfdagi kompaktlik	
	K_1	K_2	K_1	K_2
Yevklid	0,3006	0,9690	0,0273(126)	0,0435(105)
Chebishev	0,2733	0,9676	0,0293(113)	0,0882(91)
Juravlyov	0,3163	0,9755	0,0297(119)	0,1603(87)

2 - jadval natijalarini tahlil qilish orqali sinflardagi kompaktlik va turg‘unlik o‘rtasida bog‘liqlikni ko‘rishimiz mumkinki.

II-bobda **Algoritmni o‘rgatish uchun pretsedentlar bazasini shakllantirish** qaraladi. O‘rgatuvchi tanlanmani tsenzuralashtirish – bu mashinali o‘rgatish

modellari uchun ishlatiladigan o‘rgatuvchi tanlanmadan ma’lum obyektlarni cheklash yoki olib tashlash jarayonidir.

Kompaktlikni baholash uchun ikkita o‘lchov taklif etiladi:

1. Obyektlar orasidagi munosabatlar tuzilmasini (0;1] intervalida (5) bo‘yicha baholash;
2. Anglash algoritmlarining umumlashtirish qobiliyatini baholash.

(5) bo‘yicha ruxsat etilgan qiymatlar to‘plami (0;1] intervalida har bir sinfdagi guruhlar soni va ularning quvvatiga bog‘liq. 2 o‘lchov tanlanmadagi shovqin obyektlarni chiqarib tashlagan holda, minimal qoplamaga ega bitta etalon obyekt tomonidan jalb qilinadigan o‘rtacha obyektlar soni sifatida aniqlanadi. Ushbu o‘lchov algoritmlarning umumlashtirish qobiliyatini baholashda mashhur kross-validayatsiya usuliga muqobil sifatida taklif etiladi.

Shovqin obyektlar olib tashlangandan so‘ng natijada ularning soni va tarkibi chegaraviylar tuzilmasini, qoplama etalonlari to‘plamining quvvatini o‘zgartiradi. Qoplama etalonlarining soni o‘rgatuvchi tanlanmaning vakillik darajasi ko‘rsatkichi hisoblanadi.

§2.1 da **masalaning qo‘yilishi**. **Kompaktlik o‘lchovi asosida regulyarizatsiya qilish** muhokama qilinadi**.**

Standart shakldagi anglash masalasini qaraymiz. Faraz qilamizki, *m* obyektlar to‘plami berilgan $E_0=\{S_1,\ldots,S_m\}$ va ular *l* ta o‘zaro kesishmaydigan $K_1,...,K_l$ sinflarga bo‘lingan. Obyektlar tavsifi *n* xil turdagi alomatlar yordamida amalga oshiriladi, ulardan ξ alomat interval shkalalarida, n - ξ alomat esa nominal (nomlanadigan) shkalalarda o‘lchanadi. Obyektlar to‘plamida $\rho(x,y)$ metrika berilgan.

To‘plamlar uchun belgilashlarni kiritamiz:

$$B(E,\rho)=\left\{S\in E\middle|\rho(S_i,S)=\min_{S_i\in K_j,S_d\in CK_j}\rho(S_i,S_d)\right\}$$ – sinf chegaraviy obyektlari;

$T\subset B(E_0,\rho)$ – E_0 to‘plamda $\rho(x,y)$; $E=E_0\setminus T$ metrika bo‘yicha aniqlanadigan shovqin obyektlar. *E* to‘plamda aniqlangan, obyektlar bog‘liqligi munosabatlari strukturasini shakllantiramiz.

Faraz qilamizki, *E* to‘plamida minimal qoplama to‘plami E_{ob} ni shakllantirish va kompaktlik o‘lchovini hisoblash uchun ochko‘z algoritm aniqlangan.

$$\mu(E,\rho)=|E|/|E_{ob}| \qquad (7)$$

$S\in E_{ob}\cap K_t$ etaloniga yaqinlik $\rho_s(x,y)=\alpha_s\rho(x,y)$ lokal metrika bo‘yicha hisoblanadi, bu yerda α_s parametri $E\cap CK_t$ chegaraviy obyektlari asosida aniqlanadi. Kompaktlik o‘lchovi (7) *E* to‘plamidagi bitta etalon tomonidan tortib turadigan obyektlarning o‘rtacha soni sifatida ko‘rib chiqiladi. Bu etalonlar minimal qoplama to‘plami E_{ob} dan olinadi.

Shovqin obyektlar to‘plami *T* ning quvvatini va uning tarkibini aniqlash talab qilinadi, bunda

$$\mu(E,\rho)=\max_{T\subset E_0}\mu(E_0\setminus T,\rho). \quad (8)$$

O‘rgatuvchi tanlanmaning minimal qoplamasi etalonlarini shakllantirish jarayoni quyidagi bosqichlarni ketma-ket bajarish orqali amalga oshiriladi:

- Berilgan $\rho(x,y)$ metrika bo‘yicha sinflarning chegaraviy obyektlari $B(E_0,\rho)$ to‘plamini ajratib olish;
- Shovqin obyektlarni $T\subset B(E_0,\rho)$ chegaraviy obyektlar to‘plamidan topish va olib tashlash;
- $E=E_0\setminus T$ bo‘yicha chegaraviylar orqali bog‘liqlik munosabatiga asoslanib, sinf obyektlarini o‘zaro kesishmaydigan guruhlarga ajratish;
- Har bir guruh uchun etalonlar yordamida minimal qoplamani shakllantirish.

§2.2 da **Sinflar orasida nisbiy chekinish parametrini tanlash** anglash aniqligini oshirish yo‘llari tavsiflanadi. Chegaraviy obyektlar to‘plamida $B=B(E_0,\rho)$ juftliklar to‘plamini shakllantiriladi $BG=\{(S_i,S_j)\}$, $S_i\in K_t\cap B$, $t\geq 2$, $S_j\in CK_t\cap B$, $\rho(S_i,S_j)=\min\limits_{S_v\in B\cap CK_t}\rho(S_i,S_v)$. $(S_i,S_j)\in BG$ uchun belgilashni kiritamiz

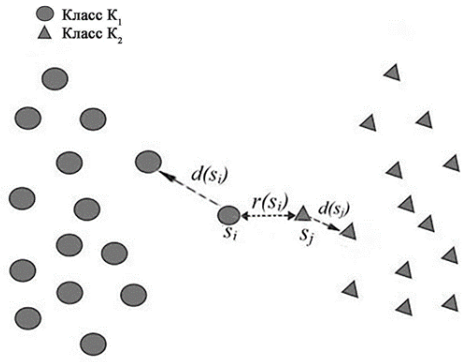
$$r(S_i)=\rho(S_i,S_j),\ d(S_i)=\rho(S_i,S_v),$$

bu yerda $S_\mu=\arg\min\limits_{S_a\in E_0\cap CK_t\setminus\{S_j\}}\rho(S_k,S_a)$.

$\frac{r(S_i)}{d(S_i)}<\lambda, 0<\lambda<1$ munosabati $S_i\in K_t\cap B$ obyektni shovqin obyektlar to‘plamiga tegishliligini zarur sharti sifatida qaraladi. Yetarli sharti quyidagicha

$$\frac{r(S_i)}{d(S_i)}<\lambda \quad and \quad \frac{r(S_j)}{d(S_j)}\geq\lambda\ . \quad (9)$$

Chegaraviy $S_i\in K_1$ obyektni $r(S_i)$, $d(S_j)$, $d(S_i)$ masofalar munosabatlari bo‘yicha $E_0=K_1 U K_2$ tanlanmada shovqin obyektlar to‘plamiga tegishliligini aniqlash 1-rasmda tasvirlangan.



1 - rasm. Chegaraviy $S_i\in K_1$ obyektni $r(S_i)$, $d(S_j)$, $d(S_i)$ masofalar munosabatlari bo‘yicha shovqin obyektlar to‘plamiga tegishliligi

Regulyarizator (koeffitsient) parametri sifatida (9) bo‘yicha aniqlanadigan λ qiymati, fiksirlangan faktorlarda kompaktlikni (8) bo‘yicha ekstremal qiymatlarini

izlash uchun qo‘llaniladi. Obyektlar munosabatlari tuzilmalarini o‘zgartiruvchi faktorlarni(obyektlar o‘rtasidagi masofalar o‘lchovi, normirovka usuli, alomatlar to‘plamlarining tarkibi v.h.k) tanlashning samaradorligi haqidagi qaror odatda hisoblash eksperimenti natijalari bo‘yicha qabul qilinadi.

3-jadval. German tanlanmasida Juravlyov metrikasi bo‘yicha regulyarizatsiya koeffitsientiga bog‘liq holda shovqin va etalon obyektlarni tanlash.

Regulyarizatsiya koeffitsienti	Obyektlar soni		(8) etalon bo‘yicha o‘rtacha qiymat
	shovqinlar	etalonlar	
0,5	42	267 (126, 141)	3,4373
0,6	60	259 (120, 139)	3,4116
0,7	54	259 (112, 147)	3,4553
0,8	42	260 (114, 146)	3,5299
0,9	27	277 (127, 150)	3,4178

German tanlanmasi uchun (3-jadvalga qarang) optimal yechim sifatida regulyarizatsiya koeffitsienti 0,8 bo‘lganda 42 ta shovqin obyektlarni o‘chirish 260 etalonni tanlash tavsiya etiladi.

Shovqin obyektlar to‘plamining xilma-xilligi (3-jadvalga qarang) regulyarizatsiya koeffitsienti tanlovi bilan bog‘liq. Bu tasdiqning namoyon qilish uchun 4-jadvalda koeffitsientning turli qiymatlarida olingan to‘plamlar kesishmasidagi umumiy obyektlarning mavjudligi haqidagi ma’lumotlar keltirilgan.

4 - jadval. To‘plamlar kesishmasida umumiy obyektlar soni

Regulyarizatsiya koeffitsienti	0,5	0,6	0,7	0,8	0,9
0,5	0	26	12	7	7
0,6	26	0	35	22	17
0,7	12	35	0	33	22
0,8	7	22	33	0	26
0,9	7	17	22	26	0

Spambase tanlanmasida o‘tkazilgan tajribada dastlabki 4601 ta obyekt soni 4204 taga kamaytirildi. Ulardan 2528 ta obyekt 1-sinfga, 1676 tasi esa 2-sinfga tegishli. Ikkala sinfda kesishgan obyektlar olib tashlandi va har bir sinfdagi bir-biriga o‘xshash obyektlardan faqat bittadan vakil qoldirildi.

Etalon obyektlarni tanlash samaradorligini tekshirish uchun Spambase ma’lumotlar to‘plamidagi 4204 ta obyekt teng kuchga ega ikkita tanlanmaga ajratildi. Bu ajratishda har bir sinfdagi obyektlarning indekslari juft va toq tartibda taqsimlandi. Har bir tanlanma (Chet va Nechet) o‘qitish va nazorat uchun ishlatildi. Ikkala tanlanma bo‘yicha pretsedentlarni tanlash natijalari 4-jadvalda keltirilgan. Ushbu pretsedentlar minimal qoplama etalonlari hisoblanadi. Ularni shakllantirishda [0;1] oralig‘ida ma’lumotlar tavsiflari bo‘yicha masofalarni hisoblash uchun lokal metrikalar qo‘llanilgan.

4 - jadval. Chet va Nechet tanlanmasi bo‘yicha pretsedentlarni tanlash natijalari

Metrika	Yevklid		Chebishev	
Tanlanma	Chet	Nechet	Chet	Nechet

Regulyarizatsiya koeffitsienti		0,7	0,5	0,9	0,8
Obyektlar soni	shovqinlar	41	15	55	65
	etalonlar	223 (113, 110)	246 (122, 124)	176 (159, 17)	210 (137, 73)
(8) bo‘yicha o‘rtacha etalon		9,0619	8,4232	11,3264	9,4000

5 - jadval. Yevklid metrikasi bo‘yicha anglash aniqligi

Tanlanma bo‘yicha pretsedentlar	Nazorat tanlanma	
	Chet	Nechet
Chet	–	88,20 (87,01)
Nechet	88,73(88,63)	–

5-jadvalning birinchi qatorida 246 ta etalondan iborat pretsedentlar bazasi (4-jadval) Chet tanlanmadan olingan 2102 ta obyektni test qilish uchun ishlatiladi. Xuddi shunday, ikkinchi qatorida Nechet tanlanmadan olingan 223 ta etalondan iborat pretsedentlar bazasi Chet tanlanmadan olingan 2102 ta obyektni test qilish uchun qo‘llaniladi. Qavs ichida keltirilgan 88,63% va 87,01% aniqlik ko‘rsatkichlari lokal metrikalar qo‘llangan holdagi pretsedentlar bazasiga nisbatan pastroq natijalardir. Ammo minimal qoplama etalonlari bo‘yicha algoritm hisoblash murakkabligi bir necha barobar kamaygan.

Ushbu ishda shovqin obyektlarni klaster tahlili uchun tanlash texnologiyasining maqsadi quyidagilardan iborat:

- sinflar tarkibini to‘g‘rilash uchun shovqin obyektlarni aniqlash;
- bog‘liqlik munosabatlariga asoslanib sinflar obyektlarini guruhlarga ajratish;
- dastlabki berilganlarda yashirin qonuniyatlarni izlash.

Chegaradagi obyektni shovqin obyektlar to‘plamiga kiritish (yoki kiritmaslik) to‘g‘risidagi qaror, ushbu obyektning o‘z sinfidagi va sinfdan tashqaridagi obyektlarga bo‘lgan eng yaqin ikki masofasining nisbati asosida qabul qilinadi.

Tahlil qilish uchun bunday qaror qabul qilishga asos bo‘ladigan shartlarni aniqlash talab qilinadi. Klaster tahlilida chegaradagi obyekt bir guruhdan boshqasiga o‘tish uchun nomzod sifatida ko‘riladi. Tushuntirish uchun (9)-dagi kabi belgilashlar qo‘llaniladi. $\frac{r(S_i)}{d(S_i)} < 1$ munosabat $S_i \in K_t \cap B$ obyektni shovqin obyektlar to‘plamiga tegishli ekanligini zarur sharti sifatida qaraladi.

Yetarli sharti esa quyidagicha

$$\frac{r(S_i)}{d(S_i)} < 1 \quad and \quad \frac{r(S_j)}{d(S_j)} \geq 1. \qquad (10)$$

Chegaraviy $S_i \in K_1$ obyektni $r(S_i)$, $d(S_j)$, $d(S_i)$ masofalar munosabatlari bo‘yicha $E_0 = K_1 \cup K_2$ tanlanmada shovqin obyektlar to‘plamiga tegishliligini aniqlash

2 - rasmda tasvirlangan. Shart (10) ga ko‘ra S_i obyekt K_1 sinfdan K_2 sinfga o‘tishi kerak.

Deylik $J_{11},\ldots, J_{1u}$ ($J_{11} \cup \ldots \cup J_{1u} = K_1$) и $J_{21}, \ldots, J_{2v}$ ($J_{21} \cup \ldots \cup J_{2v} = K_2$) - K_1 va K_2 sinflarda obyektlar bog‘liqlik munusabatlari bo‘yicha olingan kesishmaydigan guruhlar to‘plami. Har bir guruh kesishmaydigan chegaraviy obyektlar $B(E_0,\rho)$ to‘plam ostisi bo‘yicha aniqlanadi: K_1 yoki K_2. Klasterlarning sinflar va umuman tanlanma bo‘yicha kompaktligi o‘lchovi klasterlash sifatini baholash uchun ishlatiladi va quyidagicha aniqlanadi

$$\Omega(K_1) = \frac{|J_{11}|^2 + \cdots + |J_{1u}|^2}{|K_1|^2} \tag{11}$$

$$\Omega(K_2) = \frac{|J_{21}|^2 + \cdots + |J_{2v}|^2}{|K_2|^2} \tag{12}$$

Obyektlarni sinflarga ajratish variantlaridan biri bu taqsimot zichligi qiymatlaridan foydalanishdir. Zichlik qiymatini k yaqin qo‘shnilar orqali yoki ε radiusli gipershar shaklidagi lokal soha alomatlar fazosi orqali hisoblash mumkin. Zichlik qiymatlarining xilma-xilligi, shu jumladan, k va **ε** parametrlariga bog‘liq bo‘ladi. Sinflar tarkibini aniqlashtirish (10) bo‘yicha shovqin obyektlarni izlash orqali amalga oshiriladi. Guruhlar ichida va guruhlar orasidagi munosabatlar tuzilmasining o‘zgarishi jarayoni shovqin obyektlarni ajratish va ularning bir guruh tarkibidan boshqasiga o‘tishi bilan bog‘liq. Obyektni shovqin obyektlar to‘plamiga kiritish natijalari taqsimot zichligini hisoblash uchun ε parametriga bog‘liq bo‘ladi, bunda $\varepsilon = \varepsilon(k)$.

Guruh ichidagi munosabatlar tuzilishini quyidagilar orqali tadqiq qilish mumkin: minimal qoplama etalonlar o‘rtasidagi munosabatlar va obyektlarning ularga bo‘lgan masofalari EQYY. EQYY qiymatlari 2 yoki undan ortiq etalon mavjud bo‘lgan holda hisoblanadi. Guruh konfiguratsiyasi bo‘yicha yashirin qonuniyatlarni aniqlash uchun quyidagi ma’lumot manbalaridan foydalaniladi:
- etalonlar va chegaradagi obyektlar to‘plamining kesishmasi;
- etalonlar soni va chegaradagi obyektlar soni o‘rtasidagi nisbat;
- minimal qoplama bitta etaloni tomonidan tortilgan guruh obyektlarining o‘rtacha soni yoki guruhning kompaktligi.

Qavs ichida *Australian* tanlanmasi asosida sinf tarkibiga kiritilgan shovqin obyektlar soni ko‘rsatilgan.

6-jadval. Dastlabki alomatlar fazosida Yevklid metrikasi bo‘yicha sinflar kompaktlik o‘lchovi

Radius $\varepsilon(k)$	Sinf K_1		Sinf K_2	
	Obyektlar soni	(11) bo‘yicha kompaktlik	Obyektlar soni	(12) bo‘yicha kompaktlik
ε(3)	345(2)	0,9942	345(2)	0,9318
ε(5)	345(5)	0,9884	345(5)	0,8521

ε(7)	345(5)	0,9884	345(5)	0,7918
ε(10)	345(6)	0,9827	345(6)	0,7867

§2.3 da **Yaqin qo‘shni usuli asosida axborot modeli** NN qo‘llashni quyidagi variantlari taklif qilinadi:

1. Minimal hisoblash resursi sarflagan holda obyektlarni anglash;
2. Obyektlarni klaster tuzilishi tahlili va minimal qoplama etalonlarining xususiyatlari;
3. Shovqin obyektlar paydo bo‘lishi sabablarini tadqiq qilish.

Etalon obyektlar guruhlarning tipik vakillari sifatida alohida tadqiqot mavzusini tashkil etadi. Shovqin (noan’anaviy) obyektlar (3-variant) esa empirik qonuniyatlardan chetga chiqish yoki og‘ish sifatida ko‘rib chiqiladi. Masalan, sotsiologik ma’lumotlarga ko‘ra, respondentning iste’mol darajasi uning e’lon qilgan daromad darajasidan sezilarli farq qilishi mumkin.

III bobda ***k* yaqin qo‘shnilarga asoslangan anglash modellari** *k* yaqin qo‘shnilarga asoslangan metrik algoritmlardan foydalangan holda ikki sinfli anglash muammosini hal qilish ko‘rib chiqiladi. Quyidagi yechimlar taklif etiladi: Klassik KNN algoritmi ehtimoliy taqsimot zichligining noparametrik baholash asosida; maqsadli alomat qiymatlari bo‘yicha regressiya bog‘liqligini hisoblash uchun Nadaraya-Watson (NW) usuli; obyektlarning lokal atrofidagi baholashlarni hisoblash va ularning qiymatlarini ikki sinfdan birining vakillari ustunlik qiladigan kesishmaydigan intervallarga ajratish. Optimallik mezonlari sifatida KNN uchun ko‘pchilik qoidasi, NW uchun maqsadli alomat qiymatlarining sonli o‘q bo‘ylab aniq ajratilishi va interval chegaralaridagi obyekt baholarining turg‘unligi xizmat qildi. Keltirilgan variantlar uchun *k* ning optimal qiymatlarini qidirishda kross-validatsiya usuli qo‘llanildi.

§3.1 da **Yaqin qo‘shnilar soni *k* ni tanlash kross-validatsiya usuli yordamida amalga oshiriladi** va bu tasniflash yoki regressiya aniqligiga erishish uchun muhim qadam hisoblanadi. Kross-validatsiya usuli (cross-validation) *k* ning optimal qiymatini tanlash uchun standart usuldir.

Yaqin *k* ta qo‘shni (KNN) usulida tanib olish bo‘yicha, pretsedentlar bazasini shakllantirish bo‘yicha tadqiqotlar o‘tkazilmagan. KNN ni amalga oshirishda shovqin, etalon va chegara obyektlari tushunchalari, lokal metrikaning parametrlari yoki sinflar orasidagi farq masalalari ko‘rib chiqilmagan.

Maqsadli alomat qiymatlarini hisoblash usullarini tanlash va asoslash alohida vazifa hisoblanadi. Sinflar orasidagi chegaralar berilganlar tabiati haqida hech qanday taxminlarsiz aniqlanadi. Odatda, ehtimollik zichligi gipotetik funksiyasining parametrlari noma’lum bo‘ladi. Ushbu zichlik funksiyasi dastlab ko‘p ekstremumli deb faraz qilinadi. Regressiya masalalarini hal qilish uchun turli usullar mavjud, jumladan, nochiziqli usullar. Nochiziqli usullarga Nadaraya-Watson usuli, Kolmogorov va Arnold teoremasidan kelib chiqadigan bog‘liqliklarni tiklashni amalga oshiruvchi KAN neyron tarmoqlari kiradi.

§3.2. Y**aqin *k* ta qo‘shnilar asosida lokal baholashlarni hisoblash**

Yaqin *k* ta qo‘shnidan iborat mahalliy atrof sinflar vakillarining soni bo‘yicha baholashlarni hisoblash uchun ishlatiladi. Baholash qiymatlari bir-biriga yaqin bo‘lgan obyektlarning mavjudligi haqidagi taxminlar o‘rganiladi va ular asosida kesishmaydigan guruhlarni shakllantirish mumkin. Obyektlarning sinflarga tegishliligini aniqlash uchun qaror qabul qilish qoidalarida guruhlarga tegishlilik funksiyasi qiymatlarini qo‘llash taklif etiladi. Sinflarga tegishlilik funksiyalari asosida optimal *k* qiymatini tanlashga bog‘liqligi ko‘rsatilgan.

Maqsadli alomat sifatida umumlashgan baholarni hisoblashning nochiziqli algoritmlari yordamida shakllantirilgan latent qiymatlardan foydalanish taklif etiladi.

7 - Jadval. German tanlanmasidan shakllantirilgan latent alomatlar

Latent alomat nomeri	Latent alomatni shakllantirish tartibi	Kompaktlik o‘lchovi	Intervallar chegaralari	Aniqlik %
1	$(((x_1,x_2),x_4),x_{20})$	0,3766	[-0,2796;0,0] (0,0;0,0971]	75,9
2	$((((x_3, x_5),x_6,x_8),x_{10}),x_{16})$	0,3427	[-0,0734;0,0] (0,0;0,2694]	72,3
3	$((((x_7,x_{12}),x_{13}),x_{14}),x_9,x_{18})$	0,2930	[-0,1330;0,0] (0,0;0,1601]	65,9
4	(x_{15},x_{17})	0,2749	[-0,0024;0,0] (0,0;0,2726]	63,6
5	(x_{11}, x_{19})	0,2523	[-0,1678;0.0] (0,0;0,085]	59,5

Nochiziqli kombinatsiya $(((x_1, x_2), x_4), x_{20})$ 75,9% aniqlik bilan 9-jadvaldagi tanlama bo‘yicha Nadaraya-Watson usuli uchun maqsadli alomat sifatida ishlatilgan. Regressiya masalalarida klassik variant sifatida maqsadli alomat qiymatlarini $\{-1,1\}$ to‘plamidan tanlash qabul qilingan.

S_i Obyektning bahosini $g(S_i,E_0\setminus\{S_i\})$ kross-validatsiya usuli yordamida optimallashtirish uchun mezon tanlash taklif etiladi. Quyidagicha predikat aniqlangan

$$\beta\left(S_i\right)=[S_i\in K_1\ and\ g\left(S_i;E_0\setminus\{S_i\}\right)>\Gamma\ or\ S_i\in K_2\ and\ g\left(S_i;E_0\setminus\{S_i\}\right)<\Gamma],$$

bu yerda Γ nol atrofidagi chegara qiymati (9 - jadvalga qarang). Kriteriya quyidagicha yoziladi

$$LOO\left(k,E_0\right)=\sum_{i=1}^{m}\beta\left(S_i\right)\to\max_k.$$

8 - jadval. Kross -validatsiya yordamida *k* ni optimal qiymatini hisoblash

Usul	Maqsadli alomat	*k* qiymati	Aniqlik %
KNN	-	11	74,7
Nadaray - Vatson	$(((x_1, x_2), x_4), x_{20})$	10	75,0
	{ -1,1}	15	73,1

Maqsadli alomat sifatida latent alomat uchun nisbatan yuqori aniqlik 75% ko‘rsatilgan (10-jadvalga qarang).

§ 3.3 da **sinfdagi baholarni turg‘unligi bo‘yicha shovqin obyektlar to‘plamini shakllantirish haqida masala**, **sinf turg‘unligi qiymatlari bo‘yicha *k* yaqin qo‘shnilarni optimal qiymatini tanlash bo‘yicha tadqiq qilinadi**

Latent alomat c(*k*) qiymatlari sifatida E_0 obyektlarining eng yaqin *k* ta qo‘shnilaridan iborat lokal atrofda olingan baholar to‘plami $\{z(S_i,k)\}$ qabul qilinadi. Tartiblangan baholar to‘plami c(*k*) uchun kesishmaydigan intervallarga ajratish mezoni qo‘llaniladi

$$\left|\frac{d_{tc}(u,v)}{|K_t|}-\frac{d_{3-t,c}(u,v)}{|K_{3-t}|}\right|\to\max, \tag{13}$$

bu yerda $d_{tc}(u,v)$, $d_{3\text{-}t,c}(u,v)$ - $[r_u;r_v]^i$, $i\in\{1,...,p_c\}$ intevaldagi K_t, $K_{3\text{-}t}$ sinf vakillari soni, p_c – qoplash uchun intervallar soni .

b_r orqali tegishlilik funksiyasining K_1 sinfiga tegishli bo‘lgan qiymatini formulaga muvofiq belgilaymiz, agar S_r obyektining c(*k*) belgisi (13) formulaga binoan intervalga tegishli bo‘lsa. c(*k*) alomatning turg‘unligi quyidagicha hisoblanadi

$$\varphi\big(c(k)\big)=\frac{1}{m}\sum_{r=1}^{m}\begin{cases} b_r, b_r>0.5,\\ 1-b_r, b_r<0.5.\end{cases} \tag{14}$$

(14) mumkin bo‘lgan qiymatlar to‘plami (0.5;1] ga tegishli. *k* ni optimal qiymati quyidagi kriteriya asosida tanlanadi

$$k_{opt}=\arg\max_k \varphi\big(c(k)\big).$$

$S\in K_i$, i=1,2 obyektni *k* yaqin qo‘shnilari atrofida baholarni hisoblashni 2 xil usulini qarab chiqamiz

$$\frac{d_i(k)}{|K_i|}-\frac{d_{3-i}(k)}{|K_{3-i}|}, \tag{15}$$

$$\frac{d_i(k)}{|K_i|}\left(1-\frac{d_{3-i}(k)}{|K_{3-i}|}\right) \tag{16}$$

$d_i(k)\geq 0$, $d_i(k)+d_{3\text{-}i}(k)=k$ shartda. Baholash (15) additiv, (16) multiplikativ.

Eksperiment natijalarini ko‘rsatish uchun *German* tanlanmasidan foydalanilgan. Tanlanmadagi har bir 1000 ta obyektdan iborat berilganlar $|K_1|:|K_2|$ = 700:300 sinflar nisbatiga muvofiq 7 ta miqdoriy va 13 ta nominal alomat bilan tavsiflangan. Miqdoriy alomatlar qiymatlari [0;1] oralig‘ida akslantirilgan.

11-jadval natijalarini tahlil qilishdan maqsad - *German* tanlanmasida sinflar obyektlarining yomon ajraluvchanligini (turg‘unlikning 0,5 ga yaqinligi orqali) ko‘rsatishdir. *num_dependents* va *residence_since* alomatlarning boshidan

informativ bo‘lmagan deb tasniflanishimiz mumkin. (14) ga ko‘ra *num_dependents* va *residence_since* alomatlari o‘chirilgan.

Obyektlarning (15) bo‘yicha baholarining turg‘unligi yaqin qo‘shnilarni turli soni uchun qanday ekanligini 12-jadvalda ko‘rish mumkin.

9 - jadval. *k* yaqin qo‘shnilar atrofida (16) bo‘yicha baholash turg‘unligi

k ni qiymati	Intervallar soni	Turg‘unlik (14)	Aniqlik %
2	3	0,9772	95,9
3	3	0,7538	89,1
4	5	0,9977	99,5
5	3	0,8135	85,5

(15) va (16) bo‘yicha informatsion modelni shakllantirish uchun muhim ma’lumot sifatida obyektlarning shovqin obyektlar to‘plamiga tegishliligini aniqlash mumkin. Ilgari taklif qilinganidek, bunday to‘plamlarning quvvati va tarkibining sinf obyektlari o‘rtasidagi munosabatlar strukturasidan qanday bog‘liq ekanligini o‘rganish tavsiya etiladi.

(15) bo‘yicha baho qiymati 0 ga teng bo‘lishi obyektning shovqinlar to‘plamiga kiritilishi sharti sifatida qabul qilinadi, bu yerda *k* - yaqin qo‘shnilar soni. Turli *k* qiymatlari uchun shovqin obyektlar sonining o‘zgarishi 10-jadvalda ko‘rsatilgan.

10 - jadval. (16) baho bo‘yicha shovqin obyektlarni tarqalishi

k =	Sinfga tegishlilik		
	K_1	K_2	Jami
2	44	102	146
3	16	70	86
4	5	52	57
5	3	36	39

(15) bo‘yicha baholarga ko‘ra shovqin obyektlar to‘plamiga tegishlilik ularning ikkita qiymatdan biriga $\frac{k}{|K_1|}$ yoki $\frac{k}{|K_2|}$ teng bo‘lishi orqali aniqlanadi. Shovqin obyektlar to‘plamining additiv tamoyillari bo‘yicha baholarni hisoblashdagi mosligi 14-jadvalda ko‘rsatilgan.

11 - jadval. (15) baho bo‘yicha shovqin obyektlarni tarqalishi

k =	Sinfga tegishlilik		
	K_1	K_2	Jami
2	44	102	146
3	16	70	86
4	5	52	57
5	3	36	39

10-jadval va 11-jadval mazmunini tahlil qilish natijasida, K_2 sinf obyektlarining K_1 sinf obyektlariga nisbatan kuchli aralashib ketganligi aniqlanadi.

Shovqin obyektlarni §2.2 dagi nisbiy farq qiymati bo‘yicha tanlash va *k* yaqin qo‘shnilar yordamida baholarni hisoblash natijalari bo‘yicha quyidagi xulosalar chiqarish mumkin:

1. §2.2 da keltirilgan usul sinflar soni $l \geq 2$ bo‘lgan holatlarda va lokal metrikalar bo‘yicha yaqinlik o‘lchovlariga nisbatan qo‘llanilishi mumkin.
2. *k* eng yaqin qo‘shnilar yordamida baholarni hisoblash obyektlarni ikkita sinfga ajratishga mo‘ljallangan.

XULOSA

Dissertatsiya ishi kompaktlik o‘lchovlarini qo‘llash asosida klassifikatsiya va klaster tahlil uchun metrik algoritmlarni ishlab chiqish va asoslashga bag‘ishlangan. Dissertatsiya tadqiqot ishini bajarish davomida quyidagi natijalar qo‘lga kiritildi.

1. Obyektlar o‘rtasidagi aloqalarning klaster strukturasini baholash metodikasi ishlab chiqildi, ular kategoriyalarga bo‘lingan. Obyektlarni kesishuvchi gipersharlar tizimi bo‘yicha bog‘liqlik munosabatlari orqali guruhlash usuli taklif etildi va ushbu usulning xususiyatlari o‘rganildi.

2. Anglash algoritmlarining umumlashtirish qobiliyatini kompaktlik o‘lchovi orqali baholashning to‘g‘riligi isbotlandi. Kompaktlik o‘lchoviga asoslangan baholar kross-validatsiya usullariga alternativ sifatida qo‘llaniladi.

3. Obyektlarning shovqin obyektlar to‘plamlarga tegishli bo‘lishi uchun zarur va yetarli shartlar shakllantirildi. Ushbu shartlar kompaktlik o‘lchovining maksimal qiymatini ta’minlaydigan regularizator mezonida minimal qoplama etalonlaridan pretsedentlar bazasini shakllantirish uchun ishlatilgan.

4. Yaqin qo‘shnilar usuli algoritmi uchun shovqin obyektlarni tanlashning ikki usuli ishlab chiqildi. Birinchi usulda sinflarning chegara obyektlari o‘rtasidagi nisbiy farq qiymatini regularizator yordamida optimallashtirish qo‘llanilgan. Ikkinchi usulda k yaqin qo‘shnilar yordamida baholarni hisoblash, k ning optimal qiymatini tanlash va shovqin obyektlarning sonini ularning sonli o‘qdagi turg‘unlik qiymatiga qarab aniqlash asos qilib olingan.

НАУЧНЫЙ СОВЕТ DSc.03/30.12.2019.FM.01.02
ПО ПРИСУЖДЕНИЮ УЧЕНЫХ СТЕПЕНЕЙ ПРИ
НАЦИОНАЛЬНОМ УНИВЕРСИТЕТЕ УЗБЕКИСТАНА

НАЦИОНАЛЬНЫЙ УНИВЕРСИТЕТ УЗБЕКИСТАНА

ТУРСУНМУРОТОВ ДАВРБЕК ХУДАЁРОВИЧ

АНАЛИЗ СТРУКТУР ОТНОШЕНИЙ ОБЪЕКТОВ ПО МЕРАМ КОМПАКТНОСТИ

05.01.11 – Цифровые технологии и искусственный интеллект

АВТОРЕФЕРАТ
ДИССЕРТАЦИИ ДОКТОРА ФИЛОСОФИИ (PHD)
ПО ФИЗИКО-МАТЕМАТИЧЕСКИМ НАУКАМ

Ташкент – 2025

Тема диссертации доктора философии (Doctor of Philosophy) по физико-математическим наукам зарегистрирована в Высшей аттестационной комиссии при Министерстве высшего образования, науки и инноваций Республики Узбекистан за №B2025.1.PhD/FM1271.

Диссертация выполнена в Национальном Университете Узбекистана имени Мирза Улугбека.

Автореферат диссертации на трех языках (узбекский, русский, английский (резюме)) размещен на веб-странице Научного совета (http://ik-fizmat.nuu.uz/) и на Информационно-образовательном портале «Ziyonet» (www.ziyonet.uz).

Научный руководитель:	**Игнатьев Николай Александрович** доктор физико-математических наук, профессор
Официальные оппоненты:	**Кабулов Анвар Василович** доктор технических наук, профессор
	Худайбергенов Кабул Кадирбергенович PhD по физико-математическим наукам, доцент
Ведущая организация:	**Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта**

Защита диссертации состоится «____» _____2024 года в ____ часов на заседании Научного совета DSc. 03/30.12.2019.FM.01.02 при Национальном университете Узбекистана. (Адрес: 100174, г. Ташкент, Алмазарский район, ул. Университетская, 4. Тел.: (+99871)227-12-24, факс: (+99871) 246-53-21, e-mail: nauka@nuu.uz).

С диссертацией можно ознакомиться в Информационно-ресурсном центре Национального университета Узбекистана (зарегистрирована за №_____). (Адрес: 100174, г. Ташкент, Алмазарский район, ул. Университетская, 4. Тел.: (+99871) 246-02-24).

Автореферат диссертации разослан «____» _____2025 года.
(протокол рассылки №_____ от «____» _____2025 года).

М.М.Арипов
Председатель Научного совета по присуждению ученых степеней, д.ф.-м.н., профессор

З.Р.Рахмонов
Ученый секретарь Научного совета по присуждению ученых степеней, д.ф.-м.н. профессор

Д.Т.Мухамадиева
Председатель научного семинара при Научном совете по присуждению ученых степеней, д.т.н., профессор

ВВЕДЕНИЕ (аннотация диссертации доктора философии(PhD))

Актуальность и востребованность диссертации. Многочисленные научно-практические исследования, проводимые в мировом масштабе, показывают, что методы интеллектуального анализа данных (ИАД) играют важную роль в современном развитии информационных технологий. При разработке численных алгоритмов ИАД возникают такие проблемы, как «проклятие размерности», размывание отношений между объектами, комбинаторная сложность реализации на технических устройствах, а также невозможность получения результатов в приемлемое время при наличии шумовых объектов и признаков. Одним из средств обоснования выбора алгоритмов анализа и интерпретации данных является мера компактности. Изучение значений этих мер в различных отношениях между объектами обучающей выборки становится всё более актуальным в задаче повышения обобщающей способности алгоритмов интерпретации. Разработка эффективных подходов в этих направлениях остаётся одной из важнейших задач интеллектуального анализа данных.

В настоящее время в мире риск появления различных ошибок в данных, как правило, приводит к существенному ухудшению качества получаемых на их основе выводов и снижению достоверности выявленных закономерностей. При решении различных прикладных задач активно исследуются модели редактирования и очистки данных. Во многих стандартных пакетах анализа данных применяются алгоритмы фильтрации шумовых объектов для повышения качества входных данных. Наблюдение и интерпретация значений мер в различных отношениях между объектами обучающей выборки являются предметом целевых научных исследований, направленных на повышение обобщающей способности алгоритмов интерпретации.

В нашей стране уделяется особое внимание научным и практическим исследованиям в области информационных технологий, искусственного интеллекта и анализа больших данных. Разработка эффективных методов интеллектуального анализа данных, их применение на практике и повышение способности к обобщению являются одной из актуальных задач сегодняшнего дня. В последние годы достигнуты значительные результаты в разработке методов и критериев интеллектуального анализа данных, а также в создании подходов для выявления скрытых закономерностей в данных с целью повышения эффективности принимаемых решений при медицинской диагностике. Формирование крупномасштабных цифровых данных на государственном языке для обучения машинного интеллекта, а также разработка программных продуктов, поддерживающих анализ и синтез речи на государственном языке, определены в «Стратегии “Цифровой Узбекистан — 2030» в качестве приоритетных направлений[1]. Для реализации данного решения важно разработать методы типа «ближайших соседей» на основе вычисления мер плотности и компактности обучающей выборки, а также

[1] Постановление Президента Республики Узбекистан от 17 февраля 2021 года № ПҚ-4996 «О создании условий для ускоренного внедрения технологий искусственного интеллекта».

методику выбора оптимального числа *k* ближайших соседей для оценки характеристик объектов.

Данное диссертационное исследование в определённой степени служит выполнению задач, предусмотренных Указом Президента Республики Узбекистан от 7 февраля 2017 года № ПФ-4947 «О Стратегии действий по дальнейшему развитию Республики Узбекистан», Указом Президента Республики Узбекистан от 8 октября 2019 года № ПФ-5847 «О Концепции комплексного социально-экономического развития Республики Узбекистан до 2030 года», а также Постановлением Президента Республики Узбекистан от 17 февраля 2017 года № ПҚ-2789 «О мерах по дальнейшему совершенствованию деятельности Академии наук, организации, управлению и финансированию научно-исследовательских работ», Постановлением Президента Республики Узбекистан от 27 апреля 2018 года № ПҚ-3682 «О мерах по дальнейшему совершенствованию системы внедрения в практику инновационных идей, технологий и проектов», выступлением Президента на встрече с представителями науки и образования в Национальном университете Узбекистана 24 мая 2019 года и другими нормативно-правовыми документами.

Соответствие исследования приоритетным направлениям развития науки и технологий республики. Данное исследование выполнено направлением развития науки и технологий Республики Узбекистан IV. “Информатизация и развитие информационных коммуникационных технологий”.

Степень изученности проблемы. В развитие теории и практики использования методов интеллектуального анализа данных (ИАД) значительный вклад внесли известные зарубежные и отечественные ученые. Среди зарубежных исследователей, работавших в области ИАД, можно отметить Ю. И. Журавлёва, К. В. Воронцова, В. Н. Вапника, Н. Г. Загоруйко, А. Б. Петровского, Р. Е. Беллмана, J. Goodfellow, Е. Надарай, G. S. Watson и К. В. Рудакова. Среди отечественных ученых, работающих в области ИАД, следует особо выделить М. М. Камилова, Ш. Ф. Мадрахимова, Ф. Т. Адилову, Т. Ф. Бекмуратова, Н. С. Маматова, Д. Т. Мухаммадиеву и Ш. Х. Фазилова.

Процесс обучения алгоритмов распознавания при наличии шумовых объектов и выбросов в обучающей выборке реализуется поразному. В алгоритмах построения решающих деревьев для уменьшения влияния шумовых объектов предусмотрена процедура редукции (pruning) - удаление поддеревьев, имеющих низкую статистическую надежность из-за того, что для их построения использовались объекты – выбросы. В других алгоритмах предусмотрена предобработка данных, в процессе которой шумовые объекты с помощью некоторого критерия выявляются и отфильтровываются. В некоторых случаях предпринимаются попытки корректировки отдельных признаков объекта – выброса с целью преобразовать его в типичный объект. Для очистки данных от такого рода шумов предлагался подход, основанный на использовании функции конкурентного сходства. Процесс цензурирования

останавливается при достижении максимального значения оценки разделимости выборки или при отсутствии объектов – выбросов в оставшейся части выборки. Проблема обобщающей способности являются ключевой в машинном обучении. При обучении алгоритмов восстанавливают некоторую неизвестную зависимость, построенную на конечной выборке прецедентов. Требуется предсказывать точность работы алгоритма на тестовой выборке, состоящей из новых прецедентов.

Идет поиск критериев для оценки выбора алгоритмов оптимальной сложности. Одним из способов получить такие оценки для методов типа ближайших соседей было использование функционала качества по критерию компактности.Поскольку машинное обучение связано с обработкой больших баз данных, принципиальное значение имеет выбор методов их предобработки. Желательным исходом по результатам предобработки является селекция обучающих выборок как по объектам, так и по признакам. Результаты предобработки могут существенно повлиять на применение алгоритмов классификации. Интеллектуальный анализ данных широко используется компаниями и государственными органами при выборе оптимальных решений для адаптивного управления сложными системами.

Связь темы диссертации с научно-исследовательскими работами учреждением высшего образования, где выполнялась диссертация. Диссертационное исследование выполнено в соответствии с планами научно-исследовательских работ Национального университета Узбекистана имени Мирзо Улугбека в рамках темы «Разработка математической модели для предварительной диагностики цирроза печени и гепатоцеллюлярной карциномы, вызванных HBV-инфекцией».

Целью исследования Целью диссертационной работы является разработка и обоснование методов типа «ближайших соседей» с помощью вычисления мер компактности и плотности распределения объектов обучающей выборки.

Задачи исследования состоят в следующем:

- разработать метод формирования баз прецедентов из минимального покрытия обучающей выборки эталонами с использованием критериев – регуляризаторов для поиска экстремума меры компактности классов;
- определить необходимые и достаточные условия выбора шумовых объектов по относительному отступу между граничными объектами классов для повышения обобщающей способности алгоритмов метода “ближайший сосед”;
- разработать методику выбора оптимального числа k ближайших соседей для оценки свойств объектов;
- обосновать методику разбиения объектов на группы по отношению их связанности по системе пересекающихя гипершаров.

Объект исследования Разработка и обоснование методов интеллектуального анализа данных и их использование для формирования информационых моделей, основанных на знаниях.

Предмет исследования Методы вычисления мер компактности и плотности распределения для селекции обучающих выборок и группировки объектов в задачах принятия решений по прецедентам.

Методы исследования. Для исследования использовались методы дискретной математики, прикладной статистики, интеллектуального анализа данных, технологии программирования на алгоритмических языках.

Научная новизна заключается в следующем**:**

- Разработан метод формирования базы прецедентов на основе решения задачи минимального покрытия с использованием эталонов обучающей выборки;

- разработаны необходимые и достаточные условия для отбора шумовых объектов по величине относительного отклонения между граничными объектами классов;

- доказана необходимость применения меры компактности для оценки качества кластеризации при категоризации объектов на основе плотности их распределения с использованием отношения связанности через систему пересекающихся гиперсфер;

- разработан новый регуляризатор для выявления и удаления шумовых объектов, расположенных на границах классов, создан метод, использующий в качестве критерия максимальное значение меры компактности, основанной на минимальном покрытии обучающей выборки эталонами.

Практические результаты исследования:

разработан метод решения задачи минимального покрытия с использованием эталонов выборки для формирования базы прецедентов, направленный на снижение сложности алгоритмов распознавания и повышение их способности к обобщению.

разработан метод, основанный на мере компактности, для выявления и удаления шумовых объектов, а также для отбора информативных признаков, что позволяет уменьшить избыточность данных в выборке, повысить точность, устойчивость и эффективность модели.

Достоверность результатов исследования Результаты обоснованы последовательным сопоставлением, использованием соответствующего математического аппарата и теоретических подходов, а также строгим применением математических моделей.

Научная и практическая значимость результатов исследования.

Научная значимость результатов исследования объясняется разработкой новых методов интеллектуального анализа данных, направленных на повышение способности алгоритмов распознавания к обобщению.

Практическая значимость результатов исследования заключается в разработке методологии определения эталонных объектов из выборки, выделения шумовых объектов и отбора информативных признаков на основе меры компактности.

Внедрение результатов исследования. На основе научных результатов, полученных в диссертационной работе, таких как формирование

эталонов минимального покрытия по мерам компактности, анализ структур взаимосвязей объектов и группировка по пересекающимся гиперсферам:

методология анализа взаимосвязей объектов по мерам компактности была использована в инновационном проекте IL-52421091471 “Разработка аппаратно-программных средств для мониторинга водных ресурсов” при анализе процессов сложных систем (справка Ташкентского университета информационных технологий 1027/05-2 от 24 марта 2025 года). Применение научных результатов позволило оценить эффективность за счёт снижения вычислительных ресурсов и повышения точности;

алгоритм группировки объектов выборки по системам пересекающихся гиперсфер был использован в инновационном проекте IL–7823051524 “Разработка лексической платформы для контекстуального перевода на основе параллельного корпуса PARATRANSLATOR” при группировке слов по их семантическому сходству (справка Национального университета Узбекистана имени Мирзо Улугбека 04/11-5787 от 3 мая 2025 года). Применение научных результатов показало, что по результатам кластерного анализа термины демонстрируют семантическую близость к предметным областям, что, в свою очередь, позволило выявить слова, близкие по смыслу.

Апробация работы. Результаты данного исследования были обсуждены на 6-научно практических конференциях, в том числе, на 2 международных и 4 республиканских.

Публикации. По теме исследования опубликовано всего 14 научных работ, из них 6 статей — в научных изданиях, рекомендованных Высшей аттестационной комиссией Республики Узбекистан для публикации основных результатов докторских диссертаций, в том числе 2 — в зарубежных изданиях, 1 — индексирована в базе Scopus и 4 — в республиканских журналах. Опубликовано также 8 тезисов на международных и республиканских научно-практических конференциях. Кроме того, получены 2 авторских свидетельства на созданные программные продукты.

Объем и структура работы. Диссертация состоит из введения, 3 глав, заключения и приложений. Полный объём диссертации составляет 107 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ ДИССЕРТАЦИИ

Во введении обоснована актуальность и востребованность темы диссертации, определено соответствие исследования приоритетным направлениям развития науки и технологий Республики Узбекистан, приведены обзор зарубежных научных исследований по теме диссертации и степень изученности проблемы, сформулированы цели и задачи, выявлены объект и предмет исследования, изложены научная новизна и практические результаты исследования, раскрыта теоретическая и практическая значимость полученных результатов, даны сведения о внедрении результатов исследования, об опубликованных работах и о структуре диссертации.

В главе I “**Метрические алгоритмы распознавания”** исследуется вопросы, связанные с обоснованием гипотезы о компактности объектов классов.

В §1.1 рассматриваются **алгоритмы метода ближайший сосед (NN) и *k* ближайших соседей (KNN).**

Проблемы связанные с NN заключается в поиске оптимального с точки зрения обобщающей способности числа шумовых объектов и формировании базы прецедентов по технологии минимального покрытия множеств. Относительно KNN исследования ведутся по выбору числа k ближайщих соседей, метрики, оценки и устойчивости объектов, отбора информативных признаков.

Отношение связанности объектов классов рассматривается в §1.2 для разбиения объектов классов на непересекающиеся группы. Объекты групп используются для вычислении мер компактности.

Постановка задачи

Рассматривается задача распознавания в стандартной постановке. Считается, что задано множество $E_0=\{S_1,\ldots,S_m\}$ объектов, разделённое на $l(l>2)$ непересекающихся подмножеств (классов) $K_1,\ldots,K_l, E_0=\bigcup_{i=1}^{l} K_i$. Описание объектов производится с помощью набора из n разнотипных признаков $X(n)=(x_1,\ldots,x_n)$, ξ из которых измеряются в интервальных шкалах, $(n\text{-}\xi)$ – в номинальной. На множестве объектов E_0 задана метрика $\rho(x,y)$.

Обозначим через $L(E_0,\rho)$ – подмножество граничных объектов классов, определяемое на E_0 по метрике $\rho(x,y)$. Объекты $S_i,S_j\in K_t,\ t=1,\ldots,l$ считаются связанными между собой $(S_i\leftrightarrow S_j)$, если $\{S\in L(E_0,\rho)|\rho(S,S_i)<r_i$ и $\rho(S,S_j)<r_j\}\neq\varnothing$, где $r_i(r_j)$ – расстояние до ближайшего от $S_i(S_j)$ объекта из CK_t ($CK_t=E_0\backslash K_t$) по метрике $\rho(x,y)$.

Множество $G_{tv}=\{S_{v_1},\ldots,S_{v_c}\},\ c\geq 2,\ G_{tv}\subset K_t,\ v<|K_t|$ представляет область (группу) со связанными объектами в классе K_t, если для любых $S_{v_i},S_{v_t}\in G_{tv}$ существует путь $S_{v_i}\leftrightarrow S_{v_k}\leftrightarrow\ldots S_{v_l}$. Объект $S_i\in K_t,\ t=1,\ldots,l$ принадлежит группе из одного элемента и считается несвязанным, если не существует пути $S_i\leftrightarrow S_j$ ни для одного объекта $S_j\neq S_i$ и $S_j\in K_t$. Требуется определить минимальное число непересекающихся групп из связанных и несвязанных объектов по каждому классу $K_t,\ t=1,\ldots,l$.

При определении минимального числа групп из связанных и несвязанных объектов классов используется $L(E_0,\rho)$ – подмножество граничных объектов (оболочка) классов по заданной метрике ρ и описание объектов в новом пространстве из бинарных признаков. Для выделения оболочки классов для каждого $S_i\in K_t,\ t=1,\ldots,l$ строится упорядоченная по $\rho(x,y)$ последовательность

$$S_{i_0},S_{i_1},\ldots,S_{i_{m-1}},S_i=S_{i_0}. \qquad (1)$$

Пусть $S_{i_\beta} \in CK_t$ ближайший к S_i объект из (1) не входящий в класс K_t. Обозначим через $O(S_i)$ окрестность радиуса $r_i = \rho(S_i, S_{i_\beta})$ с центром в S_i, включающую все объекты, для которых $\rho(S_i, S_{i_\tau}) < r_i$, $\tau = 1,\ldots,\beta\text{-}1$. В $O(S_i)$ всегда существует непустое подмножество объектов

$$\Delta_i = \left\{ S_{i_\alpha} \in O(S_i) \,|\, \rho(S_{i_\beta}, S_{i_\alpha}) = \min_{S_{i_\tau} \in O(S_i)} \rho(S_{i_\beta}, S_{i_\tau}) \right\} \qquad (2)$$

По (2) принадлежность объектов к оболочке классов определяется как

$$L(E_0, \rho) = \bigcup_{i=1}^{m} \Delta_i .$$

Свойствами отношения связанности объектов являются:

- выборка E_0 разделяется на единственное и фиксированное число непересекающихся групп объектов;
- между любыми двумя объектами S_i, S_j из одной группы всегда можно построить цепочку $S_i \leftrightarrow S_k \leftrightarrow \ldots \leftrightarrow S_c$.

Обозначим через $R(S_i,S_j)$ – многообразие цепочек, соединяющих объекты $S_i, S_j \in G \cup K_2$, $|G| \geq 2$. Выбор минимального кратчающего пути (КНП) будет определяется как оптимизация функционала

$$z(S_i, S_j) = \min_{\rho(S_i,S_j)} \sum_{S_u, S_v \in R(S_i,S_j)} \rho(S_u, S_v). \qquad (3)$$

Целями разбиения объектов классов на непересекающиеся группы являются:

– вычисление и анализ значений компактности объектов классов и выборки в целом;

– поиск минимального покрытия обучающей выборки объектами эталонами;

– оценка качества кластерного анализа.

Результаты вычислительного эксперимента на данных "*Australian*" по мере вклада в компактность и расстоянию в группе (3) по длине КНП приводятся в табл.1.

Таблица 1. Результаты анализа в исходном признаковом пространстве данных

Метрика	Число объектов в группе	Номер класса i =	Вклад в компактность $\frac{\lvert G^2 \rvert}{\lvert K_i \rvert^2}$	Длина КНП
Евклида	18	1	0,1258	883,18
	13	1	0,0656	1108,33
	32	2	0,1604	281,54
	10	2	0,0156	227,33
	15	1	0,0873	1384,88
Журавлёва	22	1	0,1724	441,82
	23	1	0,1884	98686,39

	145	2	0,8941	2582,58
	37	2	0,0582	652,24
	9	2	0,0034	682,09

По результатам анализа из табл. 1 можно сделать следующие выводы:
- значения вклада в компактность напрямую зависят от количества объектов в группе;
- нет прямой зависимости длины КНП от числа объектов в группах. Например, для двух групп из K_1 (см.табл.1) с числом объектов 18 и 13, значения вклада компактность были соответственно 0,1258 и 0,0656, длина в КНП соответственно 883,18 и 1108,33.

Меры компактности и методы их вычисления описаны в §1.3. Общность между предлагаемой мерой компактности и обобщающей способностью алгоритмов заключается в определении и использовании кластерной структуры на множестве объектов обучающей выборки. Единственность её (меры) значений гарантируется методом разбиения объектов классов на непересекающиеся группы. Потребность в использовании кластерной структуры с точки зрения качества обучения заключается в:
– обнаружении и удалении шумовых объектов;
– выделении объектов – эталонов минимального покрытия выборки без шумовых объектов, обеспечивающих корректное разделение ее на классы.

Пусть получено разбиение обучающей выборки по отношению связанности на непересекающиеся группы $G_{i1},...,G_{i\mu}$ по каждому классу $K_i \subset E_0$, из которого удалено множество шумовых объектов $D_i \subset K_i$. Целью удаления шумовых объектов является повышение обобщающей способности алгоритмов распознавания.

Обозначим через $m_{ij} = \left|G_{ij}\right|$, $j = 1,...,\mu$, $\sum_{j=1}^{\mu} m_{ij} = m_i$. Для анализа результатов разбиения класса K_i на непересекающиеся группы с учетом их числа, представительности (по количеству объектов) и удаления шумовых объектов предлагается использовать такую структурную характеристику как оценка компактности

$$\Theta_i = \frac{\sum_{j=1}^{\mu} m_{ij}^2}{m_i^2} \qquad (4)$$

Очевидно, что множество допустимых значений Θ_i по (4) лежат в интервале $\left[\frac{1}{m_i};1\right]$. Если группа G_{i1} содержит все объекты из $K_i \cap \left(E_0 \setminus \bigcup_{j=1}^{l} D_j\right)$, то $\Theta_i = 1$. Усредненная оценка компактности обучающей выборки в целом

производится с учетом доли $\left(\frac{\left| E_0 \setminus \bigcup_{i=1}^{l} D_i \right|}{m} \right)$ исключенных из рассмотрения шумовых объектов как

$$R\left(E_0 \setminus \bigcup_{i=1}^{l} D_i, \rho \right) = \left(\frac{\left| E_0 \setminus \bigcup_{i=1}^{l} D_i \right|}{m} \right) \frac{\sum_{i=1}^{l} m_i \Theta_i}{\left| E_0 \setminus \bigcup_{i=1}^{l} D_i \right|} = \frac{\sum_{i=1}^{l} m_i \Theta_i}{m} \qquad (5)$$

Значения (4) и (5) косвенно свидетельствуют об однородности (неоднородности) структуры обучающей выборки. Чем ближе сходство групп по числу входящих в них объектов класса, тем ближе значение (4) к $\frac{1}{m_i}$, а (5) – к $\frac{l}{m}$.При $R\left(E_0 \setminus \bigcup_{i=1}^{l} D_i, \rho \right) = 1$ число групп объектов на $E_0 \setminus \bigcup_{j=1}^{l} D_j$ равно числу классов. Множество значений по (4) и (5) соответственно в $\left[\frac{1}{m_i}; 1 \right]$ и $\left[\frac{l}{m}; 1 \right]$ предлагается рассматривать в качестве меры компактности классов и выборки в целом. Значения мер компактности в указанных выше интервалах можно использовать для обнаружения скрытых закономерностей по базам данных.

Зависимость между множеством допустимых классифицированных объектов и множеством алгоритмов распознавания постулируется в виде гипотезы о компактности. Неоднозначность интерпретации понятия компактности, как правило, связано с выбором мер близости между объектами и преобразования признакового пространства. Актуальным является вопрос выбора единиц (меры) измерения компактности, допускающих однозначную интерпретацию.

Понятие устойчивости объекта естественно вытекает из процедуры вычисления оптимальной окрестности из k ближайших соседей и в данном случае понимается как качественный показатель объекта, характеризующий его размещение среди объектов одного с ним класса. Устойчивость λ_i^j объекта $S_i \in K_j, j = \overline{1,l},\ i = \overline{1,m}$, в классе K_j определяется как

$$\lambda_i^j = \frac{d_i^j}{2 \min\limits_{1 \le j \le l} \left| K_j \right| - 3}, \qquad (6)$$

где d_i^j - число событий в E_0, когда среди k ближайших к S_i объектов, $k = 1,\ldots, 2 \min\limits_{1 \le j \le l} \left| K_j \right| - 3$ большинство составляют объекты из $K_j \cap E_0$. Очевидно, что $\lambda_i^j \subset [0,1]$.

Устойчивость объектов может быть использована для отбора информативных наборов признаков, мер близости между объектами и преобразований признакового пространства. Одним из возможных подходов при решении проблемы существования данных с разными структурами является адаптация их к заданному алгоритму путем преобразования данных. Очевидно, что требуется найти такое преобразование данных, после которого они как можно лучше адаптировались под требования модели алгоритмов распознавания.

Рассмотрим в табл.2 связь между показателем устойчивости и компактности на данных *Australian*. В скобках указано число непересекающихся групп объектов.

Таблица 2. Показатели устойчивости и компактности на сырых данных

Метрика	Устойчивость в классе		Компактность в классе	
	K_1	K_2	K_1	K_2
Евклида	0,3006	0,9690	0,0273(126)	0,0435(105)
Чебышева	0,2733	0,9676	0,0293(113)	0,0882(91)
Журавлёва	0,3163	0,9755	0,0297(119)	0,1603(87)

Анализ результатов из табл.2 показывает прямую коррелированность значений устойчивости и компактности по классам.

Формирование баз прецедентов для обучения алгоритмов рассматривается в главе II.

Цензурирование обучающих выборок – это процесс ограничения или удаления определённых данных из обучающих выборок, используемых для тренировки моделей машинного обучения.

Предложены две меры компактности для оценки:

1. структуры отношений объектов в интервале (0;1] по (5);
2. обобщающей способности алгоритмов распознавания.

Множество допустимых значений (5) в интервале (0;1] зависит от количества групп в каждом классе и их мощности. Мера 2 определяется как среднее число объектов выборки за вычетом шумовых, притягиваемое одним эталоном минимального покрытия. Эта мера предлагается для оценки обобщающей способности алгоритмов в качестве альтернативы известному методу кросс-валидации. Число и состав шумовых объектов после их удаления меняет конфигурацию граничных и как следствие мощность множества эталонов покрытия. Число эталонов покрытия служат показателем представительности обучающей выборки.

В § 2.1 обсуждается **Постановка задачи. Регуляризация на основе меры компактности.**

Рассмотрим задачу распознавания в стандартной постановке. Будем считать, что задано множество из m объектов $E_0=\{S_1,\ldots,S_m\}$, разделённое на l непересекающихся классов $K_1,...,K_l$. Описание объектов производится с помощью n разнотипных признаков $X(n)=(x_1,...,x_n)$, ξ из которых измеряются

в интервальных шкалах, n - ξ – в номинальной. На множестве объектов E_0 задана метрика $\rho(x,y)$.

Введем обозначения множеств:

$B(E,\rho)=\left\{ S\in E \middle| \rho(S_i,S)=\min\limits_{S_i\in K_j, S_d\in CK_j} \rho(S_i,S_d) \right\}$ – граничных объектов классов;

$T\subset B(E_0,\rho)$ – шумовых объектов, определяемых на E_0 по метрике $\rho(x,y)$;

$E=E_0\setminus T$.

По E формируется структура отношений связанности объектов, которое определено в § 1.2.

Считается, что на множестве E определен жадный алгоритм формирования множества эталонов минимального покрытия E_{ob} и вычисления меры компактности

$$\mu(E,\rho)=\left|E\right|/\left|E_{ob}\right| \qquad (7)$$

Близость к эталону $S\in E_{ob}\bigcap K_t$ вычислим по локальной метрике $\rho_s(x,y)=\alpha_s\rho(x,y)$, где α_s – параметр, определяемый по граничным объектам из $E\bigcap CK_t$. Мера компактности (7) рассматривается как среднее число объектов из E притягивается одним эталоном минимального покрытия из E_{ob} .

Требуется определить мощность множества шумовых объектов T и его состав, при котором

$$\mu(E,\rho)=\max_{T\subset E_0}\mu(E_0\setminus T,\rho) \qquad (8)$$

Процесс формирования минимального покрытия обучающей выборки эталонами реализуется путем последовательного выполнения следующих этапов:

- выделение множества граничных объектов классов $B(E_0,\rho)$ по заданной метрике $\rho(x,y)$;
- поиск и удаление шумовых объектов $T\subset B(E_0,\rho)$ из множества граничных;
- разбиение объектов классов на непересекающиеся группы по отношению связанности по множеству граничных на $E=E_0\setminus T$;
- формирование минимального покрытия из эталонов по каждой группе.

В **§ 2.2 О выборе параметра относительного отступа между классами** описывается пути повышения точности распознавания**.**

На множестве граничных объектов B=$B(E_0,\rho)$ сформируем множество пар BG={(S_i,S_j)}, $S_i\in K_t\cap B$, $t\geq 2$, $S_j\in CK_t\cap B$, $\rho(S_i,S_j)=\min\limits_{S_v\in B\cap CK_t}\rho(S_i,S_v)$. Для ($S_i$,$S_j$)$\in BG$ введём обозначения

$$r(S_i)=\rho(S_i,S_j),\ d(S_i)=\rho(S_i,S_v),$$

где $S_\nu = \arg \min_{S_a \in E_0 \cap K_t \setminus \{S_i\}} \rho(S_j, S_a)$. Аналогично для $S_j \in CK_t \cap B$ определим

$$r(S_j) = \rho(S_k, S_j) = \min_{S_\nu \in B \cap K_t} \rho(S_\nu, S_j), d(S_j) = \rho(S_j, S_\mu),$$

где $S_\mu = \arg \min_{S_a \in E_0 \cap CK_t \setminus \{S_j\}} \rho(S_k, S_a)$.

Отношение $\frac{r(S_i)}{d(S_i)} < \lambda, 0 < \lambda < 1$ рассматривается как необходимое условие отнесение объекта $S_i \in K_t \cap B$ к множеству шумовых. Достаточным условием является

$$\frac{r(S_i)}{d(S_i)} < \lambda \quad and \quad \frac{r(S_j)}{d(S_j)} \geq \lambda . \tag{9}$$

Иллюстрация определения принадлежности граничного объекта $S_i \in K_1$ к множеству шумовых на выборке $E_0 = K_1 \cup K_2$ по отношениям расстояний $r(S_i)$, $d(S_j)$, $d(S_i)$ показана на рис.1.

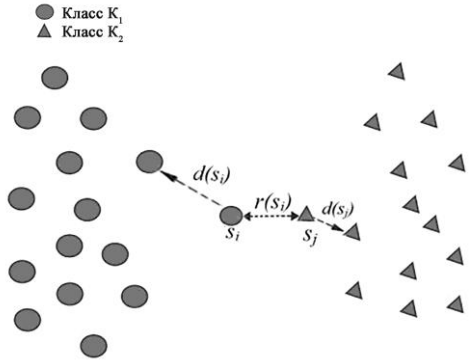


Рис.1. Отнесение граничного объекта $S_i \in K_1$ к множеству шумовых по отношениям расстояний $r(S_i)$, $d(S_j)$, $d(S_i)$

Значение λ, определяемое по (9) в качестве параметра (коэффициента) регуляризатора, применяется для поиска экстремального значения меры компактности (8) при фиксированных факторах. Решение об эффективности выбора факторов (мера расстояния между объектами, способ нормирования, состав набора признаков и т. д.), изменяющих структуру отношений объектов, как правило, принимается по результатам вычислительного эксперимента.

Таблица 3. Отбор шумовых и эталонных объектов на данных *German* в зависимости от значений коэффициентов регуляризации по метрике Журавлёва

Коэффициент Регуляризации	Число объектов		Среднее по эталону (8)
	шумовых	эталонов	
0,5	42	267 (126, 141)	3,4373
0,6	60	259 (120, 139)	3,4116
0,7	54	259 (112, 147)	3,4553
0,8	42	260 (114, 146)	3,5299
0,9	27	277 (127, 150)	3,4178

В качестве оптимального решения для данных German (см. табл. 3) рекомендуется удаление 42 шумовых объектов и отбор 260 эталонов при коэффициенте регуляризации 0,8.

Многообразие наборов шумовых объектов (см. табл. 3) связано с выбором коэффициента регуляризации. Для демонстрации этого утверждения в табл. 4 приведем данные о наличии общих объектов в пересечении наборов, полученных при разных значениях коэффициента.

Таблица 4. Количество общих объектов в пересечении наборов

Коэффициент регуляризации	0,5	0,6	0,7	0,8	0,9
0,5	0	26	12	7	7
0,6	26	0	35	22	17
0,7	12	35	0	33	22
0,8	7	22	33	0	26
0,9	7	17	22	26	0

При эксперименте на выборке данных *Spambase* количество объектов с исходных 4601 было уменьшено до 4204, из них 2528 представителей 1-го класса и 1676 – 2-го. Удалены пересекающиеся объекты из двух классов и из сходных по описанию объектов в каждом классе оставлено по одному представителю.

Для проверки эффективности отбора эталонных объектов в качестве прецедентов для обучения было произведено разбиение 4204 объектов Spambase на две равные по мощности выборки. При этом использован порядок следования четных и нечетных номеров индексов объектов в каждом классе. Каждая выборка (Chet и Nechet) применялась для обучения и контроля. Результаты отбора прецедентов по двум выборкам представлены в табл. 5 Прецедентами считаются эталоны минимального покрытия, при формировании которого использовались локальные метрики для вычисления расстояния по описаниям данных в [0;1].

Таблица 5. Результаты отбора прецедентов по выборкам Chet и Nechet

Метрики		Евклида		Чебышева	
Выборки		Chet	Nechet	Chet	Nechet
Коэффициент регуляризации		0,7	0,5	0,9	0,8
Число объектов	шумовых	41	15	55	65
	эталонов	223 (113, 110)	246 (122, 124)	176 (159, 17)	210 (137, 73)
Среднее по эталону (8)		9,0619	8,4232	11,3264	9,4000

Таблица 6. Точность распознавания по метрике Евклида

Прецеденты по выборке	Контрольная выборка	
	Chet	Nechet
Chet	–	88,20 (87,01)
Nechet	88,73(88,63)	–

В первой строке табл. 6 база прецедентов из 246 эталонов (табл. 5) для выборки Chet используется для тестирования 2102 объектов из – Nechet. Аналогично во второй строке база прецедентов из 223 эталонов (табл. 5) для выборки Nechet применена для тестирования 2102 объектов из – Chet. Точность 88,63 и 87,01, указанные в скобках, ниже результатов распознавания по базам прецедентов с применением локальных метрик. В разы снижена сложность вычислений алгоритмом по эталонам минимального покрытия.

Смысл предлагаемой в работе технологии отбора шумовых объектов для кластерного анализа заключается в:

- определении шумовых объектов для корректировки состава классов;
- разбиении на группы классов на основе описанных выше отношений связанности объектов;
- поиске скрытых закономерностей в изначально не размеченных данных.

Решение о включении (не включении) граничного объекта в множество шумовых принимается по отношению двух ближайших от него расстояний до объектов из своего класса и его дополнения.

Для анализа требуется определить условия, на основе которых принимается такое решение. В кластерном анализе граничный объект рассматривается как кандидат на переход из одной группы в другую. Для объяснения используются те же обозначения что и для (9). Отношение $\frac{r(S_i)}{d(S_i)}<1$ рассматривается как необходимое условие отнесение объекта $S_i \in K_t \cap B$ к множеству шумовых. Достаточным условием является

$$\frac{r(S_i)}{d(S_i)}<1 \ \ and \ \ \frac{r(S_j)}{d(S_j)}\geq 1. \qquad (10)$$

Иллюстрация определения принадлежности граничного объекта $S_i \in K_1$ к множеству шумовых на выборке $E_0 = K_1 \cup K_2$ по отношениям расстояний $r(S_i)$, $d(S_j)$, $d(S_i)$ показана на рис.2. Согласно условию (10) объект S_i из класса K_1 должен перейти в класс K_2.

Пусть $J_{11},\ldots, J_{1u}$ ($J_{11} \cup \ldots \cup J_{1u} = K_1$) и $J_{21}, \ldots, J_{2v}$ ($J_{21} \cup \ldots \cup J_{2v} = K_2$) – множества непересекающихся групп, полученных по отношению связанности объектов по классам K_1 и K_2. Каждая группа определяется по непересекающемуся подмножеству граничных объектов $B(E_0,\rho)$, принадлежащих одному из классов: K_1 или K_2. Мера компактности кластеров по классам и выборки в целом используется для оценки качества кластеризации и определяется как

$$\Omega(K_1) = \frac{|J_{11}|^2 + \cdots + |J_{1u}|^2}{|K_1|^2} \qquad (11)$$

$$\Omega(K_2) = \frac{|J_{21}|^2 + \cdots + |J_{2v}|^2}{|K_2|^2} \qquad (12)$$

Одним из упомянутых вариантов разделения объектов на классы является использование значений плотности распределения. Значение плотности можно вычислить по k ближайшим соседям или через локальную область признакового пространства в форме гипершара радиуса ε. Многообразие значений плотности зависит в том числе от параметров k и ε.

Уточнения состава классов будет производится через поиск шумовых объектов по (10). С изменением структуры отношений внутри групп и между группами связан процесс выделения шумовых объектов и перехода из состава одной группы в другую. Результаты отнесения объекта к множеству шумовых зависят от значения параметра ε для вычисления плотности распределения ε = ε(k).

Исследовать структуру отношений внутри группы можно через КНП между эталонами минимального покрытия и расстояния объектов до них.Значения КНП вычисляются при числе эталонов больше или равном 2. Источниками информации для обнаружения скрытых закономерностей по конфигурации группы являются:

– пересечение множества эталонов и граничных объектов;

– соотношение между числом эталонов и числом граничных объектов;

– среднее число объектов группы, притягиваемое одним эталоном минимального покрытия или компактность группы.

В скобках (см. табл.7) указано число шумовых объектов, включенных в состав класса на данных *Australian*.

Таблица 7. Мера компактности классов по метрике Евклида в исходном признаковом пространстве

Радиус ε(k)	Класс K_1		Класс K_2	
	Число объектов	Компактность по (11)	Число объектов	Компактность по (12)
ε(3)	345(2)	0,9942	345(2)	0,9318
ε(5)	345(5)	0,9884	345(5)	0,8521
ε(7)	345(5)	0,9884	345(5)	0,7918
ε(10)	345(6)	0,9827	345(6)	0,7867

В **§2.3 Информационные модели на основе метода ближайший сосед** рекомендуются следующие варианты применения NN:

1. Распознавание объектов с минимальными затратами вычислительных ресурсов;
2. Анализ кластерной структуры объектов и свойств эталонов минимального покрытия;
3. Исследование причин появления шумовых объектов.

Эталонные объекты являются предметом отдельного исследования как типичные представители групп. Шумовые (нетипичные) объекты (вариант 3) рассматриваются как выбросы или отклонения от эмпирических закономерностей. Например, по социологическим данным уровень потребления респондента существенно различается от уровня заявленных им доходов.

В **глава III. Модели распознавания на основе *k* ближайших соседей** рассматривается решение задачи распознавания с двумя классами с применением метрических алгоритмов на базе *k* ближайших соседей. Вариантами решения предлагаются: классический алгоритм KNN на основе непараметрической оценки плотности распределения вероятностей; метод Надарая - Ватсона (NW) вычисления регрессионной зависимости по определяемым значениям целевого признака; вычисления оценок в локальной окрестности объектов и разбиения их значений на непересекающиеся интервалы с доминированием в них представителей одного из двух классов. Критериями оптимальности служили правило мажоритарности для KNN, точность разделение значений целевого признака на числовой оси для NW и устойчивости оценок объектов в границах интервалов. При поиске оптимальных значений *k* для перечисленных вариантов применялся метод скользящего контроля.

В **§ 3.1 Выбор параметра ближайших соседей *k* методом скользящего контроля** является критическим шагом для достижения высокой точности классификации или регрессии. Метод скользящего контроля (cross-validation) является стандартным способом для подбора оптимального значения *k*.

Относительно метода распознавания по *k* ближайщим соседям (KNN) исследований по формированию базы прецедентов не проводилось. Для реализации KNN не рассматривались понятия шумовой, эталонный и граничный объекты, параметры локальных метрик, отступы между классами.

Выбор и обоснование методов вычисления значений целевого признака является отдельной задачей. Граница между классами определяется без каких – либо предположений о природе среды данных. Как правило, параметры гипотетический функции плотности распределения неизвестны. Функция плотности изначально полагается многоэкстремальной. Для решения задач регрессии существует различные методы, в том числе и нелинейные. К нелинейным относятся метод Надарая – Ватсона, нейронные сети KAN, реализующие восстановление зависимостей на основе вывода их из теоремы Колмогорова и Арнольда.

§ 3.2 Рассматривается вычисления оценок в локальной окрестности из *k* ближайших соседей.

Локальная окрестность из *k* ближайших соседей используется для вычисления оценок по количеству представителей классов. Исследуется предположения о наличии объектов с близкими значениями оценок, из которых можно формировать непересекающиеся группы. Значения функции принадлежности к группам предлагается применять в решающих правилах для определения принадлежности объектов к классам. Показана зависимость выбора оптимального значения *k* от функций принадлежности к классам**.**

В качестве целевого признака предложено использовать латентный, сформированный нелинейным алгоритмам вычисления обобщённых оценок.

Таблица 9. Латентные признаки, сформированные на данных *German*

Номер латентного признака	Порядок формирования латентного признака	Мера компактности	Границы интервалов	Точность в %
1	(((x1,x2),x4),x20)	0,3766	[-0,2796;0,0] (0,0;0,0971]	75,9
2	((((x3, x5),x6,x8),x10),x16)	0,3427	[-0,0734;0,0] (0,0;0,2694]	72,3
3	((((x7,x12),x13),x14),x9,x18)	0,2930	[-0,1330;0,0] (0,0;0,1601]	65,9
4	(x15,x17)	0,2749	[-0,0024;0,0] (0,0;0,2726]	63,6
5	(x11, x19)	0,2523	[-0,1678;0.0] (0,0;0,085]	59,5

Нелинейная комбинация (((x1, x2), x4), x20) с точностью 75,9 % из табл. 9 по выборке использовалось в качестве целевого признака для метода Надарая – Ватсона. Классическим вариантом в задачах регресии является выбор значений целевого признака из {-1,1}.

Предлагается выбор критерия для оптимизации $g(S_i,E_0\backslash\{S_i\})$ – оценки объекта S_i методом скользящего экзамена. Определен предикат

$$\beta\left(S_i\right)=[S_i\in K_1\ and\ g\left(S_i;E_0\setminus\{S_i\}\right)>\Gamma\ or\ S_i\in K_2\ and\ g\left(S_i;E_0\setminus\{S_i\}\right)<\Gamma],$$

где Γ - значение границы в окрестности нуля (см. табл.9). Критерий запишется таким образом

$$LOO\left(k,E_0\right)=\sum_{i=1}^{m}\beta\left(S_i\right)\to\max_k.$$

Таблица 10. Результаты вычисления оптимального *k* с помощью скользящего экзамена

Метод	Целевой признак	Значение *k*	Точность в %
KNN	-	11	74,7
Надарая - Ватсона	(((x1, x2), x4), x20)	10	75,0
	из { -1,1}	15	73,1

Относительно высокая точность 75% показана (см. табл.10) по латентному признаку в качестве целевого.

Задача о формировании множество шумовых объектов по устойчивости их оценок в классах исследуется в § 3.3.

Значениями латентного признака c(*k*) далее будем считать множества оценок $\{z(S_i,k)\}$ объектов E_0 в локальной окрестности из *k* ближайших соседей. К упорядоченному множеству (последовательности) оценок c(*k*) применим разбиение на непересекающиеся интервалы по критерию

$$\left|\frac{d_{tc}(u,v)}{\left|K_t\right|}-\frac{d_{3-t,c}(u,v)}{\left|K_{3-t}\right|}\right|\to\max,\qquad(13)$$

где $d_{tc}(u,v)$, $d_{3\text{-}t,c}(u,v)$ – количество представителей классов K_t, $K_{3\text{-}t}$ в интервале $[r_u;r_v]^i$, $i\in\{1,...,p_c\}$, p_c – число интервалов для покрытия (15).

Обозначим через b_r значение функции принадлежности $f(\mu)$ к классу K_1 по (16), если значения признака c(k) объекта S_r согласно (15) принадлежит интервалу $[r_u; r_v]^{\mu}$. Устойчивость признака $c(k)$ вычисляется как

$$\varphi\left(c\left(k\right)\right)=\frac{1}{m}\sum_{r=1}^{m}\begin{cases} b_r, b_r > 0.5, \\ 1-b_r, b_r < 0.5. \end{cases} \tag{14}$$

Множество допустимых значений (14) принадлежат (0.5;1]. Оптимальное значение k выбирается согласно критерия

$$k_{opt}=\arg\max_{k}\varphi\left(c(k)\right).$$

В качестве примера рассмотрим два способа вычисления оценок в окрестности объекта $S \in K_i$, i=1,2 из k ближайших соседей

$$\frac{d_i\left(k\right)}{|K_i|}-\frac{d_{3-i}\left(k\right)}{|K_{3-i}|}, \tag{15}$$

$$\frac{d_i\left(k\right)}{|K_i|}\left(1-\frac{d_{3-i}\left(k\right)}{|K_{3-i}|}\right) \tag{16}$$

при условии $d_i(k)\geq 0$, $d_i(k)+d_{3-i}(k)=k$. Оценка (15) является аддитивной, (16) – мультипликативной.

Для демонстрации результатов эксперимента использованы данные German. Описание каждого из 1000 объектов выборки при соотношении классов $|K_1|:|K_2|$=700:300 производилось 7-ю количественными и 13-ю номинальными признаками.

Согласия (14) кондидатами на удаление были определены признаки num_dependents и resident_since.

Какую устойчивость имеют оценки объектов по (15) при разном числе ближайших соседей видно из табл. 12.

Таблица 11. Устойчивость по оценкам (16) в окрестности *k* ближайших соседей

Значения k	Число интервалов	Устойчивость (14)	Точность в %
2	3	0,9772	95,9
3	3	0,7538	89,1
4	5	0,9977	99,5
5	3	0,8135	85,5

В качестве ценной информации для формирования информационной модели по (15) и (16) можно извлечь принадлежность объектов к множеству шумовых. Как и ранее предлагалось исследовать зависимость мощности и состава таких множеств от структуры отношений между объектами классов.

Значение оценки по (16), равное 0, является условием отнесение объекта к множеству шумовых при заданном числе k ближайших соседей. Зависимость числа шумовых объектов от разных значений k показана в табл.13.

Таблица 12. Распределения шумовых объектов по оценкам (16)

k =	Принадлежность к классам		
	K_1	K_2	Итого
2	44	102	146

3	16	70	86
4	5	52	57
5	3	36	39

Принадлежность к множеству шумовых объектов по оценкам (15) определяется по равенству их одному из двух значений $\frac{k}{|K_1|}$ или $\frac{k}{|K_2|}$. Совпадение множество шумовых объектов при вычислении оценок по аддитивному и мультипликативному принципу прослеживается по табл. 14.

Таблица 13. Распределения шумовых объектов по оценкам (15)

k =	Принадлежность к классам		
	K_1	K_2	Итого
2	44	102	146
3	16	70	86
4	5	52	57
5	3	36	39
6	1	29	30
7	1	23	24
8	1	17	18
9	1	15	16
10	1	12	13

Анализ содержимого табл.13 и табл.14 показывает наличие сильной перемешанности объектов класса K_2 относительно объектов класса K_1.

Какие выводы можно сделать по результатам отбора шумовых объектов по величине относительного отступа из §2.2 и вычисления оценок по k ближайшим соседям.

1. Метод из §2.2 применим для числа классов $l \geq 2$ и мерам близости по локальным метрикам.
2. Вычисление оценок по k ближайшим соседям рассчитано на разделение объектов на два класса.

ЗАКЛЮЧЕНИЕ

Диссертационная работа посвящена разработке и обоснованию метрических алгоритмов классификации и кластерного анализа на основе применения мер компактности. При выполнении диссертационного исследования получены следующие результаты.

1. Разработана методика оценки кластерной структуры отношений объектов, разделённых по категориям . Предложен метод группировки через отношения связанности объектов по системе пересекающихся гипершаров. Исследованы свойства этого метода.

2. Доказана корректность оценки обобщающей способности алгоритмов распознавания через меру компактности. Оценки по мере компактности являются альтернативой методами кросс- валидации.

3. Сформулированы необходимое и достаточное условия принадлежности объектов к множеству шумовых. Эти условия использовались в критерии – регуляризаторе по максимуму меры компактности для формировании базы прецедентов из эталонов минимального покрытия .

4. Разработана два метода отбора шумовых объектов для алгоритма метода ближайший сосед. В первом методе использовалась оптимизация величины относительного отступа между граничными объектами классов по регуляризатору. В основе второго метода лежит вычисления оценок по k ближайшим соседям выбор оптимального значения k и состава шумовых объектов по значению устойчивости на числовой шкале.

SCIENTIFIC COUNCIL AWARDING SCIENTIFIC DEGREES DSc.03/30.12.2019.FM.01.02 NATIONAL UNIVERSITY OF UZBEKISTAN

NATIONAL UNIVERSITY OF UZBEKISTAN

TURSUNMUROTOV DAVRBEK XUDAYOROVICH

DIMENSIONALITY OF COMPACTNESS AND IN FEATURE SPACE ANALYSIS OF OBJECT RELATIONS

05.01.11 – Digital technologies and artificial intelligence

ABSTRACT OF DISSERTATION OF THE DOCTOR OF PHILOSOPHY (PhD) ON PHYSICAL AND MATHEMATICAL SCIENCES

Tashkent–2025

The theme of dissertation of doctor of philosophy (PhD) on physical and mathematical sciences was registered at the Supreme Attestation Commission at the Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan under number №B2025.1.PhD/FM1271.

Dissertation has been prepared at the National University of Uzbekistan named after Mirzo Ulugbek.

Abstract of the dissertation is posted in three languages (Uzbek, Russian, English (resume)) on the website (http://ik-fizmat.nuu.uz/) and the "Ziyonet" Information and educational portal (www.ziyonet.uz).

Scientific supervisor:	**Ignatyev Nikolay Aleksandrovich** doctor of physical and mathematical sciences, professor
Official opponents:	**Kabulov Anvar Vasilovich** Doctor of Technical Sciences, Professor
	Khudaybergenov Kabul Kadirbergenovich Candidate of Physical and Mathematical Sciences, Associate Professor
Leading organization:	**Research Institute for the Development of Digital Technologies and Artificial Intelligence**

Defense will take place «____» _____2023 at _____ at the meeting of Scientific Council number DSc.03/30.12.2019.FM.01.02 at National University of Uzbekistan. (Address: 100174, Uzbekistan, Tashkent city, Almazar district, University str. 4, Ph.: (+99871) 227-12-24, fax: (+99871) 246-53-21, e-mail: nauka@nuu.uz).

Dissertation is possible to review in Information-resource centre at National University of Uzbekistan (is registered №____) (Address: 100174, Uzbekistan, Tashkent city, Almazar district, University str. 4, Ph.: +99871) 227-12-24).

Abstract of dissertation sent out on «____» _____2025 year
(Mailing report № _____on «____» _____2025 year)

M.M.Aripov
Chairman of Scientific Council on award of scientific degrees, D.F.M.S., Professor

Z.R.Rakhmanov
Scientific secretary of Scientific Council on award of scientific degrees, D.F.M.S.professor

D.T.Muhamediyeva
Chairman of Scientific Seminar under Scientific Council on award of scientific degrees, D.T.S., professor

INTRODUCTION (abstract of PhD thesis)

The aim of research work is to develop and justify methods of the “nearest neighbors” type by calculating measures of compactness and the density of objects distribution in the training dataset.

The object of the research work is development and justification of data mining methods and their application for creating knowledge-based information models.

Scientific novelty of the research work is as follows:

- formation of case bases based on solving the problem of minimal coverage of training samples with prototypes.

- necessary and sufficient conditions for selecting noisy objects based on the magnitude of the relative margin between the boundary objects of classes have been determined.

- it has been proven that when categorizing objects based on the density of their distribution, the measure of compactness must be used to evaluate the quality of clustering, employing the connectivity relationship in the system of intersecting hyperspheres.

- a regularizer has been developed to identify and remove the optimal number of noisy objects as a subset of the boundary objects of classes according to a given metric. The maximum value of the compactness measure, determined based on the minimal coverage of the training sample with prototypes, was used as the extremum.

Implementation of the research results. Based on the scientific results obtained in the dissertation research — such as the formation of minimal coverage prototypes according to compactness measures, the analysis of object relationship structures, and grouping based on intersecting hyperspheres — the following applications have been implemented:

the methodology for analyzing object relationships based on compactness measures was applied in the innovative project IL-52421091471 “Development of hardware and software tools for water resource monitoring” to analyze complex system processes (reference of the Tashkent University of Information Technologies 1027/05-2 dated March 24, 2025). The application of the scientific results made it possible to evaluate efficiency by reducing computational resources and increasing accuracy;

the algorithm for grouping sample objects based on systems of intersecting hyperspheres was used in the innovative project IL–7823051524 “Development of a lexical platform for contextual translation based on the PARATRANSLATOR parallel corpus” to group words according to their semantics (reference of the National University of Uzbekistan named after Mirzo Ulugbek 04/11-5787 dated May 3, 2025). The application of the scientific results showed, according to the results of cluster analysis, the semantic proximity of terms to specific subject domains, which in turn made it possible to identify semantically similar words.

The structure and volume of the dissertation. The dissertation work consists of the introduction, three chapters, conclusion, bibliography. The volume of the thesis is 107 pages.

E’LON QILINGAN ISHLAR RO‘YXATI
СПИСОК ОПУБЛИКОВАННЫХ РАБОТ
LIST OF PUBLISHED WORKS

I bo‘lim (Часть I; Part I)

1. Игнатьев Н.А., Турсунмуротов Д.Х. Цензурирование обучающих выборок с использованием регуляризации отношений связанности объектов классов // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 2. С. 322–329. doi: 10.17586/2226-1494-2024-24-2-322-329 (3, Scopus, IF=0.165).
2. Игнатьев Н. А. Турсунмуротов Д. Х. Об регуляризаторах в метрических алгоритмах распознавания// Вестник НУУз, 2023, №1. С. 254-261. (01.00.00. № 8).
3. Игнатьев Н.А, Турсунмуротов Д.Х. Оценка структуры отношений объектов при нормировании данных// Проблемы вычислительной и прикладной математики, 2022, № 4 . С. 138-146. (01.00.00. № 9).
4. Турсунмуротов Д.Х. Отбор информативных признаков по мере компактности// Проблемы вычислительной и прикладной математики, 2023, № 5 . С. 133-139. (01.00.00. № 9).
5. Игнатьев Н.А., Турсунмуротов Д.Х. Об эффективности метрических алгоритмов распознавания на базе *k* ближайших соседей // Проблемы вычислительной и прикладной математики, 2024, №4. С. 118-127. (01.00.00. № 9).

II bo‘lim (Часть II; Part II)

6. Ignatev N. A., Tursunmurotov D. X. Формирование баз прецедентов с применением регуляризации // Amaliy matematikaning zamonaviy muammolari va istiqbollari mavzusidagi Respublika ilmiy-amaliy konferensiya tezislar to‘plami. Qarshi, 2024-yil 24-25- may, 149 - 152 betlar.
7. Турсунмуротов Д.Х. Цензурирование обучающих выборок с использованием регуляризации // “Hisoblash modellari va texnologiyalari” (CMT2024). professor M.I. Isroilov tavalludining 90 yilligiga bag‘ishlangan uchinchi xalqaro seminar. 2024 – yil 27-april, 188-190 betlar.
8. Игнатьев Н.А., Турсунмуратов Д.Х. регуляризация отношений связанности объектов классов // “Актуальные задачи математического моделирования и информационных технологий” международной научно-практической конференции Нукус – 2023.С. 111-113.
9. Tursunmurotov D.X Select informative features, on the measure of compactness of class objects. // “Actual problems of applied mathematics and information technologies” - Al-khwarizmi 2023 of the 8th international conference C. 290
10. Турсунмуротов Д.Х., Акбаров Б.Х. Отношение связанности объектов классов.// “Математик моделлаштириш,алгоритмлаш ва дастурлашнинг долзарб муаммолари” Республика илмий-техник анжумани материаллари тўплами Тошкент, 2023. С. 136-139

11. Турсунмуротов Д.Х Оценка структуры отношений объектов с учётом плотности распределения данных // “Современное состояние и перспективы развития цифровых технологий и искусственного интеллекта” Сборник докладов республиканской научно-технической конференции Самарканд, 26-27 октября 2022 г. С 83-88.

12. Ignatev N. A., Tursunmurotov D. X. On the regularization of recognition algorithm estimates by the k-nearest neighbors method // Of the ix international scientific conference “actual problems ofapplied mathematics and information technologies al-khwarizmi 2024”. Dedicated to the 630 th anniversary of the birth of Mirzo Ulugbek. 22-23 October, 2024, Tashkent, 207 – 208 p.

13. Турсунмуротов Даврбек Худоёрович, Акбаров Бахриддин Хусниддин угли. Регуляризация в алгоритмах ближайших соседей методы и влияние на обобщающую способность // Актуальные вопросы науки и образования 2025 сборник статей II Международной научно-практической конференции, Состоявшейся 23 мая 2025 г. в г. Пенза, С. 11-13.

14. Tursunmurotov D.X. The significance and application of comtactness measures in machine learning // Journal of Multidisciplinary Sciences and Innovations, 2025, №6. P. 125 – 130.