15-§. Ma’lumotlarni intellectual qidirishda klasterlash

Reja:

1. *Ma’lumotlarni intellectual qidirishda klasterlash usullari.*
2. *Klasterizatsiya va uning turlari.*

Klasterlashtirish ob’yektlar to’plamini bir hil guruhlarga (klasterlar yoki sinflar) ajratish uchun mo’ljallangan. Agar namunadagi ma’lumotlar xususiyatlar makonida nuqta sifatida taqdim etilsa, unda klasterlash vazifasi "nuqta kontsentratsiyasi" ta’rifiga tushiriladi.

Klasterlash vazifasi tasniflash vazifasiga o’xshaydi, uning mantiqiy davomi, ammo farqi shundaki, o’rganilayotgan ma’lumotlar to’plamining sinflari oldindan aniqlanmagan.

Klasterlash avtomatik tasniflash, nazoratsiz o’rganish va taksonomiya bilan

sinonimdir.

Klasterlashning maqsadi mavjud tuzilmalarni topishdir. Klasterlash bu tavsiflash protsedurasi bo’lib, u hech qanday statistik xulosalar chiqarmaydi, ammo

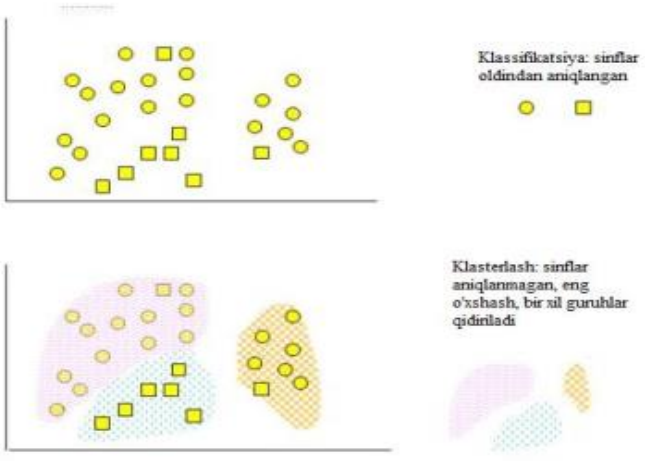
tahlil ma’lumotlarini va "ma’lumotlar tuzilishini" o’rganishga imkon beradi. Klasterni umumiy xususiyatlarga ega ob’yektlar guruhi sifatida tavsiflash mumkin.

Klaster ikki xususiyatga ega:

- ichki bir xillik;
- tashqi izolyatsiya.

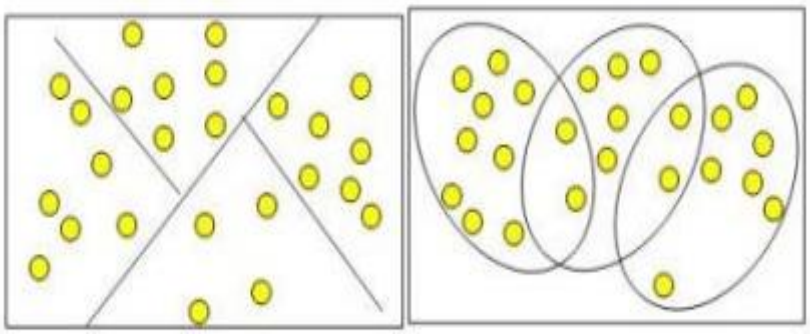
Ko’pgina muammolarni hal qilishda tahlilchilar savol berishadi - bu ma’lumotlarni vizual tuzilmalarga qanday tashkil qilish, ya’ni. taksonomiyalarni kengaytirish. Dastlab, klasterlash biologiya, antropologiya, psixologiya kabi fanlarda eng katta qo’llanmani oldi. Iqtisodiy muammolarni hal qilish uchun uzoq

vaqt davomida iqtisodiy ma’lumotlar va hodisalarning o’ziga xos xususiyatlari tufayli klasterlash kam ishlatilgan.



3.6-rasm. Klassifikatsiyalash va klasterlash muammolarini taqqoslash.

Klasterlar bir-biriga zid bo’lmagan yoki eksklyuziv (bir-biriga mos kelmaydigan, eksklyuziv) va bir-biri bilan qoplangan (bir-biri bilan) bo’lishi mumkin. Parchalanish va kesishgan klasterlarning sxematik ko’rinishi shakl. 3.7



3.7-rasm. Bir-biri bilan kesishmaydigan va bir-biri bilan kesishadigan klasterlar

Shuni ta’kidlash kerakki, klaster tahlilining turli usullarini qo’llash natijasida

turli shakllardagi klasterlarni olish mumkin. Masalan, "zanjir" tipidagi klasterlar mumkin, agar klasterlar uzun "zanjirlar", cho’zilgan shakldagi klasterlar va

boshqalar bilan ta’minlangan bo’lsa va ba’zi usullar o’zboshimchalik shaklidagi klasterlarni yaratishi mumkin. Turli xil usullar ma’lum o’lchamdagi klasterlarni yaratishga moyil bo’lishi mumkin (masalan, kichik yoki katta), yoki ma’lumotlar to’plamida turli o’lchamdagi klasterlar mavjudligini taxmin qilish mumkin.

Klasterlash jarayoni. Klaster jarayoni tanlangan usulga bog’liq va deyarli har doim iterativdir. Bu kulgili bo’lishi mumkin va turli xil parametrlarni tanlashda

tajribani o’z ichiga olishi mumkin, masalan, masofa o’lchovlari, masalan, standartlashtirish o’zgaruvchilari, klasterlar soni va boshqalar. Biroq, tajribalar o’zo’zidan tugamasligi kerak - axir, klasterlashning asosiy maqsadi o’rganilayotgan ma’lumotlarning tuzilishi haqida mazmunli ma’lumot olishdir. Olingan natijalar hosil bo’lgan klasterlarni aniq tavsiflash uchun ob’yektlarning xususiyatlari va xususiyatlarini keyingi izohlashni, tadqiq qilishni va o’rganishni talab qiladi.

Klassifikatsiya - o'rganilayotgan predmetlar, hodisalar, jarayonlarning jinsi, turlari, turlari bo'yicha, ularni o'rganish qulayligi uchun har qanday muhim belgilar bo'yicha tizimli ravishda taqsimlash; asl tushunchalarni guruhlash va ularni o'xshashlik darajasini aks ettiruvchi ma'lum bir tartibda joylashtirish. Klassifikatsiya - bu ba'zi bir printsipga muvofiq buyurtma qilingan ob’yektlar to'plami bo'lib, ular o'xshash tasniflash xususiyatlariga ega (bir yoki bir nechta xususiyatlar), ushbu ob’yektlar o'rtasidagi o'xshashlik yoki farqni aniqlash uchun tanlangan. Tasniflash quyidagi qoidalarga rioya qilishni talab qiladi:

- har bir bo'linish aktida faqat bitta bazani qo'llash kerak;
- bo'linish mutanosib bo'lishi kerak, ya'ni. aniq kontseptsiyalarning umumiy hajmi bo'linadigan umumiy kontseptsiya hajmiga teng bo'lishi kerak;
- bo'linma a'zolari o'zaro mutlaqo bo'lishi kerak, ularning hajmi bir-biriga zid bo'lmasligi kerak;
- bo'linish ketma-ket bo'lishi kerak.

Farqlanadi:

- tashqi xususiyatga ko'ra tuzilgan va ob'yektlar (jarayonlar, hodisalar) to'plamini kerakli tartibda berishga xizmat qiladigan yordamchi (sun'iy) tasnif;
- ob'yektlar va hodisalarning ichki hamjamiyatini tavsiflovchi muhim xususiyatlarga ko'ra tuzilgan tabiiy tasnif.

Tanlangan xususiyatlarga, ularning kombinatsiyasiga va tushunchalarni bo'lish tartibiga qarab tasniflash quyidagicha bo'lishi mumkin:

- oddiy
- umumiy

tushunchani faqat barcha turlar ochilgunga qadar va faqat bir marta bo'lish. Bunday tasniflashning misoli dixotomiya bo'lib, unda faqat ikkita tushuncha bo'linish a'zolari bo'lib, ularning har biri boshqasiga ziddir (ya'ni, printsipga rioya qilinadi: "A va A emas");

- murakkab - bitta kontseptsiyani turli asoslar bo'yicha ajratish va bunday sodda bo'linishlarning yaxlitligini birlashtirish. Bunday tasniflashga misol kimyoviy elementlarning davriy jadvali.

Klassifikatsiya - bu ma'lum bir guruhning xususiyatlarini aniqlash to'g'risida xulosa chiqarish imkonini beradigan naqsh. Shunday qilib, tasniflash uchun, u yoki bu hodisa yoki ob’yekt tegishli bo'lgan guruhni tavsiflovchi belgilar bo'lishi kerak (odatda, ba'zi qoidalar allaqachon tasniflangan hodisalarni tahlil qilish asosida tuzilgan). Klassifikatsiya boshqariladigan yoki boshqariladigan ta'lim deb ham yuritiladigan, nazorat qilinadigan o'quv strategiyasini anglatadi.

Klassifikatsiya vazifasi odatda doimiy va/yoki kategoriyali o'zgaruvchilar namunasiga asoslangan kategoriyaga bog'liq o'zgaruvchini (ya'ni, toifaga bog'liq bo'lgan o'zgaruvchini) bashorat qilish deb nomlanadi. Masalan, firmaning mijozlaridan qaysi biri ma'lum bir mahsulotni potentsial xaridor ekanligini va kim bo'lmaganligini, kompaniyaning xizmatlaridan kim foydalanishini va kim xohlamasligini va hokazolarni taxmin qilishingiz mumkin. Muammoning bu turi

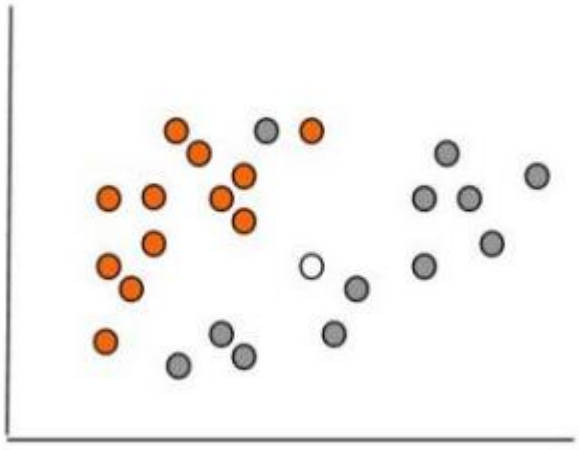
ikkilik tasniflash muammolariga tegishli bo'lib, bunda bog'liq o'zgaruvchi faqat ikkita qiymatni olishi mumkin (masalan, ha yoki yo'q, 0 yoki 1). Boshqa klassifikatsiya optsiyasi, agar bog'liq o'zgaruvchi oldindan belgilangan sinflar to'plamidan qiymatlarni olishi mumkin bo'lsa, paydo bo'ladi. Masalan, mijoz qaysi markadagi avtomobilni sotib olishni xohlashini oldindan aytib berish kerak bo'lganda. Ushbu holatlarda ko'plab o'zgaruvchilar bog'liq bo'lgan o'zgaruvchi uchun ko'rib chiqiladi. Klassifikatsiya bir o'lchovli (bitta atribut) va ko'p o'lchovli (ikki yoki undan ortiq atribut) bo'lishi mumkin. Ko'p o'lchovli klassifikatsiya biologlar tomonidan organizmlarni tasniflash uchun diskriminatsiya masalalarini hal qilish uchun ishlab chiqilgan. Ushbu yo'nalishga bag'ishlangan dastlabki ishlardan biri R. Fisher (1930) ning ishi bo'lib, unda organizmlar fizik parametrlarini o'lchash natijalariga qarab kichik turlarga bo'lingan. Biologiya ko'p o'lchovli tasnif usullarini ishlab chiqish uchun eng mashhur va qulay muhit bo'lib kelgan va shunday bo'lib qolmoqda.

Klassifikatsiya muammosini oddiy misol yordamida ko'rib chiqamiz. Aytaylik, sizning yoshingiz va oylik daromadingiz to'g'risida ma'lumotlarga ega bo'lgan sayyohlik agentligining mijozlari ma'lumotlar bazasi mavjud. Reklama materiallarining ikki turi mavjud: qimmatroq va qulay dam olish va arzonroq, yoshlar ta'tili. Shunga ko'ra, mijozlarning ikkita klassi aniqlanadi:
1-sinf va 2-sinf. Ma'lumotlar bazasi 3.1-jadvalda keltirilgan.

Mijoz ID	Yoshi	Foyda	Sinf
1	18	25	1
2	22	100	1
3	30	70	1
4	32	120	1
5	24	15	2
6	25	22	1
7	32	50	2
8	19	45	2
9	22	75	1
10	40	90	2

3.1-jadval. Sayyohlik agentligining mijozlar bazasi

Vazifa. Yangi mijoz qaysi sinfga tegishli ekanligini va qaysi turdagi reklama materiallarini yuborishi kerakligini aniqlang. Aniqlik uchun, keling, bizning ma'lumotlar bazamizni 1-sinfga (apelsin yorlig'i) va 2-sinfga (kul rang yorliq) tegishli ob’yektlar to'plami sifatida ikki o'lchovli o'lchovda (yosh va daromad) taqdim etamiz. Shaklda 3.1-da ikkita sinfdagi ob’yektlar ko'rsatilgan.



3.8-rasm. 2D ichida bir nechta ma'lumotlar bazasi ob’yektlari

Bizning muammoni hal qilish oq yorlig'i bilan rasmda ko'rsatilgan yangi mijoz qaysi sinfga tegishli ekanligini aniqlash bo'ladi. Klassifikatsiya jarayoni. Klassifikatsiya jarayonining maqsadi predmetli atributlarni kirish sifatida ishlatadigan va bog'liq bo'lgan atributning qiymatini oladigan modelni yaratishdir. Tasniflash jarayoni muayyan mezon bo'yicha ob’yektlar to'plamini sinflarga bo'lishdan iborat.

Tasniflagich(Klassifikator) - bu atributlar vektoriga tegishli predmetning qaysi sinfiga tegishli ekanligini aniqlaydigan muayyan ob’yekt. Klassifikatsiyani matematik usullardan foydalangan holda amalga oshirish uchun tasniflashning matematik apparati yordamida boshqarilishi mumkin bo'lgan ob’yektning rasmiy tavsifi bo'lishi kerak. Bizning holatda, bunday tavsif ma'lumotlar bazasi. Har bir ob’yekt (ma'lumotlar bazasi yozuvi) ob’yektning ba'zi mulki haqida ma'lumotni olib yuradi. Dastlabki ma'lumotlar to'plami (yoki ma'lumotlar namunasi) ikkita to'plamga bo'lingan: o'quv va sinov.

O'quv to'plami - bu modelni tayyorlash (qurish) uchun ishlatiladigan ma'lumotlarni o'z ichiga olgan to'plam. Ushbu to'plam misollar kirish va chiqish (maqsad) qiymatlarini o'z ichiga oladi. Chiqish qiymatlari modelni o'qitish uchun mo'ljallangan. Sinov to'plamida, shuningdek, misollar kirish va chiqish qiymatlari mavjud. Bu erda, chiqish qiymatlari modelning sog'lig'ini sinash uchun ishlatiladi. Tasniflash jarayoni ikki bosqichdan iborat: modelni qurish va undan foydalanish.

Klasterlash vazifasi tasniflash vazifasiga o'xshaydi, uning mantiqiy davomi, ammo farqi shundaki, o'rganilayotgan ma'lumotlar to'plamining sinflari oldindan aniqlanmagan. Klasterlash avtomatik tasniflash, nazoratsiz o'rganish va taksonomiya bilan sinonimdir.

Klasterlashtirish ob’yektlar to'plamini bir hil guruhlarga (klasterlar yoki sinflar) ajratish uchun mo'ljallangan. Agar namunadagi ma'lumotlar xususiyatlar makonida nuqta sifatida taqdim etilsa, unda klasterlash vazifasi "nuqta kontsentratsiyasi" ta'rifiga tushiriladi. Klasterlashning maqsadi mavjud

tuzilmalarni topishdir. Klasterlash bu tavsiflash protsedurasi bo'lib, u hech qanday statistik xulosalar chiqarmaydi, ammo tahlil ma'lumotlarini va "ma'lumotlar tuzilishini" o'rganishga imkon beradi. Klaster" tushunchasi noaniq ravishda aniqlanadi: har bir tadqiqotda o'ziga xos "klasterlar" mavjud. Klaster tushunchasi "klaster", "to'plam" deb tarjima qilinadi. Klasterni umumiy xususiyatlarga ega ob’yektlar guruhi sifatida tavsiflash mumkin.

Klaster ikki xususiyatga ega:

- ichki bir xillik;
- tashqi izolyatsiya.

Ko'pgina muammolarni hal qilishda tahlilchilar savol berishadi - bu ma'lumotlarni vizual tuzilmalarga qanday tashkil qilish, ya'ni. taksonomiyalarni kengaytirish. Dastlab, klasterlash biologiya, antropologiya, psixologiya kabi fanlarda eng katta qo'llanmani oldi. Iqtisodiy muammolarni hal qilish uchun uzoq vaqt davomida iqtisodiy ma'lumotlar va hodisalarning o'ziga xos xususiyatlari tufayli klasterlash kam ishlatilgan.

Klaster tahlilini qo’llash. Klaster tahlili turli sohalarda qo’llaniladi. Bu katta miqdordagi ma’lumotlarni tasniflash zarur bo’lganda foydalidir. Tibbiyotda kasalliklarni klasterlash, kasalliklarni davolash yoki ularning alomatlarini davolash, shuningdek, bemorlarning taksonomiyasi, dorilar va boshqalar qo’llaniladi. Arxeologiyada tosh konstruktsiyalari va qadimiy buyumlar va boshqalarning taksonomiyalari o’rnatiladi. Marketingda bu raqobatchilar va iste’molchilarni segmentlashtirish vazifasi bo’lishi mumkin. Menejmentda klasterlash vazifasi misolida xodimlarni turli guruhlarga bo’lish, iste’molchilar va yetkazib beruvchilarni tasniflash, nikoh ro’y beradigan shunga o’xshash ishlab chiqarish holatini aniqlash mumkin. Tibbiyotda alomatlar tasnifi. Sotsiologiyada klasterlashning vazifasi respondentlarni bir hil guruhlarga bo’lishdir.

Marketing tadqiqotlaridagi klaster tahlili. Marketing tadqiqotlarida klasterli tahlil juda keng qo’llaniladi - ham nazariy tadqiqotlarda, ham turli ob’yektlarni guruhlash muammolarini hal qiluvchi marketologlar tomonidan. Bu

mijozlar guruhlari, mahsulotlar va boshqalar haqida savollarni hal qiladi. Shunday
qilib, marketing tadqiqotlarida klasterli tahlilni qo’llashda eng muhim vazifalardan
biri iste’molchilar xatti-harakatlarini tahlil qilishdir, ya’ni har bir guruhdan mijozning xulq-atvori va uning xatti-harakatlariga ta’sir qiluvchi omillar to’g’risida xaridorlarni bir hil sinflarga guruhlash.

Klaster tahlili hal qila oladigan muhim vazifa - joylashishni aniqlash, ya’ni. bozorda yangi mahsulotni joylashtiradigan joyni aniqlash. Klaster tahlilini qo’llash natijasida xarita tuzilib, unga binoan bozorning turli segmentlarida raqobat darajasini va ushbu segmentga kirish imkoniyati uchun tovarlarning tegishli xususiyatlarini aniqlash mumkin.

Bunday xaritani tahlil qilish orqali bozorda yangi, bo’sh joylarni topish mumkin, ularda mavjud mahsulotlarni taklif qilishingiz yoki yangilarini ishlab chiqishingiz mumkin. Klaster tahlili, masalan, kompaniya mijozlarini tahlil qilish uchun ham foydali bo’lishi mumkin. Buning uchun barcha mijozlar klasterlarga guruhlangan va har bir klaster uchun individual siyosat ishlab chiqilgan. Ushbu yondashuv sizga tahlil qilish ob’yektlarini sezilarli darajada kamaytirishga imkon beradi va shu bilan birga mijozlarning har bir guruhiga individual yondoshadi.

Marketing tadqiqotlarida klaster tahlilidan foydalanish amaliyoti. Marketing tadqiqotlarida klaster tahlilidan foydalanish bo’yicha ba’zi taniqli maqolalar. 1971 yilda mijozlarning xohish-istaklarini tavsiflovchi ma’lumotlar asosida mijozlarni qiziqish doirasi bo’yicha segmentatsiya qilish to’g’risida maqola e’lon qilindi. 1974 yilda Sextonning maqolasi nashr etildi, uning maqsadi mahsulot iste’molchilari bo’lgan oilalar guruhlarini aniqlash edi, natijada brendni aniqlash strategiyalari ishlab chiqilgan. Tadqiqot respondentlarning mahsulot va brendlarga bergan reytinglariga asoslandi. 1981 yilda bir qator o’zgaruvchilardan olingan omil yuklamalari asosida yangi avtomobil sotib oluvchilarning xatti-harakatlarini tahlil qiluvchi maqola e’lon qilindi.

Klassifikatsiya va klasterlash muammolarini o’xshashligi ko’rinishiga qaramay, ular turli yo’llar bilan va turli usullardan foydalangan holda hal qilinadi. Vazifalardagi farq birinchi navbatda dastlabki ma’lumotlarda. Data Mining-ning eng oddiy vazifasi bo’lgan tasniflash "boshqariladigan o’rganish" strategiyasiga tegishli, chunki uni echish uchun o’quv namunasi kirish va chiqish (maqsadli) o’zgaruvchilarning qiymatlarini o’z ichiga olishi kerak.

O’z navbatida, klasterlash - bu ma’lumotni ishlab chiqarishni nazorat qilinmaydigan o’rganish strategiyasi bilan bog’liq, ya’ni. o’quv namunasida maqsadli o’zgaruvchilar qiymatining mavjudligini talab qilmaydi.

Klassifikatsiya muammosi turli usullar yordamida hal qilinadi, eng sodda – chiziqli regressiya. Usulni tanlash dastlabki ma’lumotlar to’plamini o’rganishga asoslangan bo’lishi kerak. Klaster muammosini hal qilishda eng keng tarqalgan usullar: kvositalar usuli (faqat raqamli atributlar bilan ishlaydi), ierarxik klaster tahlili (shuningdek, ramziy atributlar bilan ishlaydi), SOM usuli. Klasterlashning
murakkabligi uni baholash zarurati hisoblanadi.

Nazorat savollari

Ma’lumotlarni intellectual qidirishda klasterlash usullarini izohlang.

Klasterlashning asosiy vazifasi nimadan iborat?

. 1. 2.

Mavzuni mustahkamlash uchun savollar.

1.K-o’rtacha algoritmi qanday muammoni hal qilish uchun mo’ljallangan:

a) Klasterlash
b) Tasniflar
c) Bashoratlar
d) O’lchamlarni qisqartirish

2.Qaysi ma’lumotlar hajmidan boshlab ma’lumotlarni saqlash uchun Hadoop klasteridan foydalanish maqsadga muvofiq?

a) 100TB
b) 100Pb
c) 100 GB
d) 1TB

3.Xususiy klinika klinikaning daromadiga hissasi bo’yicha o’z mijozlari tarkibini o’rnatmoqchi. Ushbu ma’lumotlarni tahlil qilish vazifasi qandayturga tegishli?

a) klasterlash
b) bashorat qilish
c) regressiya
d) tasnifi