

KLASTERIZATSIYA

Tojiamatov Israiljon Nurmamatovich

Farg'ona Davlat Universiteti Amaliy matematika va informatika kafedrası katta o'qituvchi

israiltojiamatov@gmail.com

Abdulahobova Ozoda Begali qizi

Farg'ona Davlat Universiteti 3-kurs talabasi

ozodaabdulahobova7@gmail.com

Anotatsiya: Ushbu maqola klasterizatsiya usullarining amaliy jihatlariga bag'ishlangan bo'lib, ularning turli sohalardagi muhim qo'llanilishi haqida batafsil ma'lumot beradi. Klasterizatsiya algoritmlari qo'llaniladigan bir nechta sohalar ko'rib chiqiladi: sun'iy intellektda ma'lumotlarni guruhlash, marketingda mijozlarni segmentlash, tibbiyotda diagnostika ma'lumotlarini tahlil qilish va transport tizimlarida marshrutlarni optimallashtirish. Bundan tashqari, maqolada klasterizatsiya texnikasining rivojlanish istiqbollari va uning yirik ma'lumotlar bazalari bilan ishlashdagi potentsiali muhokama qilinadi. Xulosa qilib aytganda, klasterizatsiya algoritmlarining ilmiy va amaliy qiymati tobora ortib borayotganini va ularni turli sohalarda samarali qo'llash insoniyatning kelajakdagi hayot tarzini sezilarli darajada o'zgartirishi mumkinligini ko'rishimiz mumkin.

Kalit so'zlar: Klasterizatsiya, Ma'lumotlarni guruhlash, K-means algoritmi, Hierarchical Clustering, DBSCAN, Ma'lumotlar tahlili, Sun'iy intellekt, Marketing segmentatsiyasi, Tibbiyot diagnostikasi, Transport optimallashtirish, Dinamik tizimlar.

Annotation: This article is devoted to the practical aspects of clustering methods and provides detailed information on their important applications in various fields. Several areas where clustering algorithms are used are considered: data grouping in artificial intelligence, customer segmentation in marketing, analysis of diagnostic data in medicine, and route optimization in transport systems. In addition, the article discusses the development prospects of clustering techniques and their potential for working with large databases. In conclusion, we can see that the scientific and practical value of clustering algorithms is increasing, and their effective application in various fields can significantly change the future lifestyle of mankind.

Keywords: Clustering, Data Grouping, K-means Algorithm, Hierarchical Clustering, DBSCAN, Data Analysis, Artificial Intelligence, Marketing Segmentation, Medical Diagnostics, Transportation Optimization, Dynamic Systems.

Аннотация: В этой статье основное внимание уделяется практическим аспектам методов кластеризации и подробному описанию их важных применений в различных областях. Рассмотрены несколько областей

применения алгоритмов кластеризации: группировка данных в искусственном интеллекте, сегментация клиентов в маркетинге, анализ диагностических данных в медицине и оптимизация маршрутов в транспортных системах. Кроме того, в статье рассматриваются перспективы развития методики кластеризации и ее возможности при работе с большими базами данных. В заключение мы видим, что научная и практическая ценность алгоритмов кластеризации возрастает, а их эффективное применение в различных областях может существенно изменить образ жизни человечества в будущем.

Ключевые слова: Кластеризация, Группировка данных, Алгоритм K-средних, Иерархическая кластеризация, DBSCAN, Анализ данных, Искусственный интеллект, Маркетинговая сегментация, Медицинская диагностика, Транспортная оптимизация, Динамические системы.

Kirish:

Klasterizatsiya – bu statistik tahlil va ma'lumotlarni qayta ishlash jarayonida keng qo'llaniladigan texnikalardan biri bo'lib, u ma'lumotlarni guruhlash orqali yashirin tuzilmalarni aniqlashga imkon beradi. Ushbu texnika oldindan belgilangan yorliqlarsiz ma'lumotlarni tahlil qilishga mo'ljallangan bo'lib, sun'iy intellekt, tibbiyot, marketing, biologiya, transport, va boshqa sohalarda keng qo'llaniladi. Klasterizatsiyaning asosiy vazifasi o'xshash xususiyatlarga ega ma'lumotlarni bitta guruhga birlashtirish va farqli xususiyatlarga ega bo'lganlarni boshqa guruhlarga ajratishdan iboratdir. Klasterizatsiya bu oldindan belgilangan o'quv majmuasi yoki ushbu sinflarning tabiati haqida ma'lumotga ega bo'lmasdan, ob'ektlarni guruhlarga bo'lish imkonini beradigan noyob usul. Model ba'zi elementlarning o'xshashligini mustaqil ravishda aniqlaydi va ularni bir sektorga birlashtiradi. Klasterlashning afzalliklaridan biri shundaki, u qanday sinflar tashkil etilishi va ularning soni qancha bo'lishi haqida bilishni talab qilmaydi. Klasterlashning ilmiy nomi nazoratsiz tasnifdir - muammo bayonining o'xshashligi tufayli. Klasterlash usullari o'quv namunasini yig'ish qiyin yoki qimmat bo'lganda tasniflash muammolarini hal qilishning samarali vositasidir. Tasdiqlash namunasi jarayon natijalarini baholash uchun kamroq misollarni talab qiladi. Ammo shuni esda tutish kerakki, nazorat qilinadigan usullarning aniqligi sezilarli darajada yuqori. Va agar o'quv namunasini to'plash mumkin bo'lsa, uni tasniflash muammosini hal qilish uchun ishlatish yaxshiroqdir. Klasterlash usullaridan foydalanishning yaxshi misollaridan biri bu geoma'lumotlarni tahlil qilishdir. Mobil telefonlarda ilovalardan foydalanilganda, ko'pincha aniq manzilni aniqlash kerak bo'ladi. GPS ma'lumotlaridagi xatolik foydalanuvchi harakatidan kelib chiqadi, ko'pincha aniq pozitsiya o'rniga ko'plab nuqtalarni kuzatishni talab qiladi. Bu, ayniqsa, ma'lum bir joyda minglab odamlarning xatti-

harakatlarini tahlil qilishda, masalan, aeroportda foydalanuvchilar taksiga o'tiradigan eng mashhur joylarni aniqlashda to'g'ri keladi.

Klasterlash muammolari

Ushbu yondashuv turli xil xususiyatlarga ega ma'lumotlar to'plami mavjud bo'lganda qo'llaniladi. Biroq, ular qandaydir birlikka ega bo'lishi kerak - klasterlashni boshqa yo'l bilan amalga oshirish mumkin bo'lmaydi. Guruhlarni quyidagilarga bo'lish mumkin:

- Mijozlarning o'ziga xos birlashmalarining xatti-harakatlarini tahlil qilish.
- Biznes raqiblari - bozorni o'rganishda.
- Kasalliklar - tiklanish statistikasini o'rganish.
- So'rov ishtirokchilari - turli guruhlardagi odamlarning fikrlarini tahlil qilish.
- SEO kalitlari - veb-sayt sahifalarida mavzular yaratish uchun.
- Olingan fayllar qulay ishlov berish uchun turli formatlarga ega.

Klasterlash turli sohalarida keng tarqalgan, chunki u bir tizimga birlashtirilishi va yagona tuzilma berilishi kerak bo'lgan deyarli barcha ma'lumotlarga nisbatan qo'llanilishi mumkin.

- **Tushunish**

Tahlilchi ma'lumotlar qanday asosda olinganligini aniqlay olishi uchun turli xil ma'lumotlarni guruhlariga ajratish kerak. Keyin qayta ishlash jarayonini amalga oshirish, masalan, turli klasterlarga ma'lum klasterlash tahlil usullarini qo'llash osonroq bo'ladi.

- **Anomaliyalarni aniqlash**

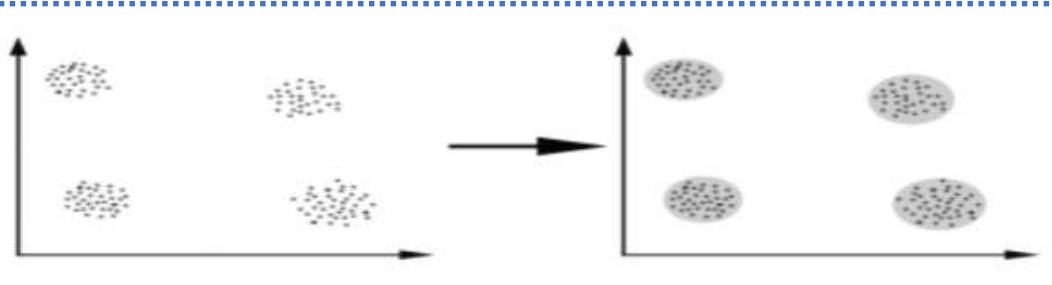
Klasterlashni amalga oshirish orqali hech qanday guruhga tegishli bo'lmagan individual ma'lumotlarni aniqlash mumkin. Xato yoki qiziqarli hodisa mavjudligini aniqlash uchun uni qayta ishlash kerak.

- **Kengaytma**

Ba'zan ma'lumot to'plashda ba'zi ma'lumotlar ko'proq xususiyatlarga ega, ba'zilari esa kamroq. O'rganilayotgan yondashuv bizga guruhning boshqa elementlarida mavjud bo'lmagan guruh xususiyatlari haqida taxminlar qilish imkonini beradi. Klasterlashtirishga misol keltiramiz. Ma'lumki, "erkaklar" guruhidagi ishtirokchilarning saytdagi o'rtacha vaqtlari 15 daqiqani tashkil qiladi. Agar klasterda saytda o'tkazgan noma'lum vaqt bilan yangi odam paydo bo'lsa, u uchun bu ham 15 daqiqani tashkil qiladi deb taxmin qilishimiz mumkin.

- **Siqish**

Katta hajmdagi ma'lumotlar guruhlariga bo'linishi mumkin, so'ngra o'rtacha hisoblab, har bir klaster uchun bitta ob'ekt bilan qoldirilishi mumkin. Bu kelajakdagi tahlillarda kamroq quvvatdan foydalanishni rag'batlantiradi.



Obyektlar va ularni klasterlarga ajratilishi

Bugungi kunda klasterizatsiya masalasini yechish uchun ko'plab uslublar va ular asosida bir nechta algoritmlar ishlab chiqilgan. Lekin bu algoritmlarni hech biri optimal hisoblanmaydi. Ba'zi algoritmlar bir xil masalalarda to'g'ri klasterlarga ajratsa, shu algoritm boshqa masala uchun to'g'ri yechim qabul qila olmasligi mumkin. Mavjud algoritmlarni ishlash uslubiga qarab quidagi sinflarga ajratish mumkin:

- Exclusive
- Ketma-ketlikka asoslangan(Overlapping)
- Daraxtsimon(Hierarchical)
- Extimollik bo'yicha(Probabilistic)

Eksklusiv klasterlash algoritmlariga misol qilib k-means algoritmini, ketma-ketlikka asoslangan fuzzy c-means, ierarxik uchun CobWeb, extimollik bo'yicha qidiruvchi algoritmga esa misol qilib EM algoritmini aytishimiz mumkin.

Weka API. Yuqorida takidlangan o'tkanimizdek hech qaysi klasterizatsiya algoritmi istalgan obyektlar to'plami uchun eng optimal bo'la olmaydi. Shu sababli biz katta obyektlar to'plamini ixtiyoriy tanlangan qismini ajratib olgan xolda ular ustida bir nechta eng ko'p qo'llaniladigan algoritmlar bilan tajriba o'tkazib ularni solishtirib ko'rishimiz kerak. Buni oson hal qilish uchun Weka API (Application Programming Interface) dan foydalanish ish jarayonini osonlashtiradi. Weka API Yangi Zelandiyaning Waikato Universiteti tomonidan Ma'lumotlarni intellektual taxlili masalalarini yechish uchun ishlab chiqilgan bo'lib, sinflarga ajratish, klasterizatsiya, bashoratlash, assotativ qoidalarni qurish va vizualizatsiya masalalarini yechish uchun bir nechta algoritmlarni o'z ichiga oladi. Weka API Java dasturlash tilida yaratilgan. Bu maqolada, algoritmlarni ishlash vaqti, egallaydigan hotira hajmi kabi ko'rsatkichlarini hisoblash maqsadida, undan qo'shimcha kutubxona sifatida foydalanamiz. Wekada klasterizatsiya masalalarini yechish uchun weka.clusterers paketi mavjud. Bu paket o'z ichiga quidagi klasterizatsiya algoritmlarini oladi:

Klasterizatsiya algoritmlari.

K-Means. Ma'lumotlarni intellektual tahlilida k-means klasterizatsiya algoritmi eng sodda, eng tushunarli va eng ko'p ishlatiladigan algoritmlardan biri hisoblanadi. K-means algoritmi berilgan n ta obyektidan iborat toplamni bir biriga o'xshash obyektlardan iborat k ta guruhga ajratadi. Bu algoritm uchun k -guruhlar soni aniq belgilangan bo'lishi kerak. Algoritmning asosiy g'oyasi k ta markazni ushlab olish va

obyektlarni shu markazlar atrofiga yig'ib chiqishdan iborat. Bunda obyektlar k ta markazdan qaysi biriga yaqin bo'lsa shu guruhga qo'shib olinadi. K-means algoritmidagi obyektlar orasidagi masofalarni hisoblash uchun Evklid masofasi, Manhattan masofasi kabilar ishlatiladi. Algoritmni asosiy abzalligi uni ishlash tezligida, k-means boshqa olgarmatlarga qaraganda tezroq ishlaydi. Lekin unga guruh(klaster)lar sonini oldindan ko'rsatish kerak. Bu k-means algoritmini eng katta kamchiligi hisoblanadi. EM(Expectation Maximization). EM algoritmi ham k-means algoritmi kabi iterativ usulda klasterlarga ajratishga mo'ljallangan. K-means yaxshi natija ko'rsatadigan barcha to'plamlar uchun EM ham yaxshi natija ko'rsata oladi. Bu algoritmi statik ma'lumotlar bazasi uchun mo'ljallangan. EM obyektlarni bir biriga o'xshashligini masofa bo'yicha emas, extimollik orqali hisoblaydi, va bu bazi holatlarda yaxshi natija berishi mumkin. Chiziqli bo'lmagan xolatlarda k-means guruhlariga ajratishda xatolikka yo'l qo'yadi, EM esa bu holatlarda ancha yaxshi natija beradi. EM real ma'lumotlar uchun boshqa algoritmlarga qaraganda yaxshi natija ko'rsatadi. Kamchiligi bir biriga yaqin joylashgan obyektlarni klasterlashda ko'pincha xatolikka yo'l qo'yadi, ishlash tezligi boshqa algoritmlarga nisbatan sekinroq.

CobWeb. CobWeb algoritmi 1980 yilda yaratilgan. Bu algoritmi ierarxik klasterlash algoritmlari qatoriga kiradi. Klassifikatsiya daraxti asosiga qurilgan. Berilgan obyektlarni extimollik bo'yicha qaysi sinfga tegishli ekanligini aniqlab, daraxtga barg sifatida qo'shib qo'yadi. Algoritmi k-means, EM algoritmlarida uchramaydigan ko'plab imkoniyatlarga ega. Masalan CobWeb algoritmi dinamik ma'lumotlarni klasterlashda ham ishlatiladi. Yangi obyektini kiritish va uni qaysi sinfga tegishli ekanligini aniqlash uchun update funksiyasi ishlatiladi, va bu amal $\log(n)$ vaqtda bajariladi. Klasterlar sonini avtomatik tarzda aniqlaydi. Bundan tashqari so'ngi qo'shilgan obyektlarni o'chirib tashlash ham mumkin. Bir so'z bilan aytganda online tarzda klasterlash uchun CobWeb algoritmi juda samarali. Lekin bu algoritmi k-means kabi barcha xolatlar uchun yaxshi yechim topa olmaydi, ko'pincha klasterlarga ajratishda xatoliklarga yo'l qo'yadi. Ma'lumotlarni berilish taribi, klasterlash natijasiga ta'sir ko'rsatadi.

DBScan. DBScan algoritmi Martin Ester, Hans-Peter, Jorge Sander va Xiaowei Xu tomonidan 1996 yilda yaratilgan. Bu algoritmi zichlikka asoslangan. Klasterlar soni o'zi aniqlab oladi. Obyektlarni berilish tartibini axamiyati yo'q, har qanday tartibda berilganda ham bir xil natija chiqaradi. Bu algoritmi bugungi kundagi eng optimal algoritmlardan biri hisoblanadi. Bu algoritmi asosiga OPTICS algoritmi ham qurilgan. Tezkor, klasterlash uchun k-means kabi samarali algoritmi. Kamchiligi sifatida bu algoritmi masofani topish bo'yicha ishlashda deb aytish mumkin. Chunki bazi holatlarda masofani Evklid masofasi bo'yicha olgan foydali bo'lsa, bazan Manhattan va shunga o'xshagan masofa formulalari yaxshi samara beradi. Iris obyektlar to'plamidan olingan natijalar shuni ko'rsatadiki, vaqt va xotira bo'yicha

eng yaxshi algoritmi bu k-means. EM algoritmi klasterlarga ajratishda biroz xatolikka yo'l qo'ygan. DBScan va CobWeb algoritmlari klasterlarga to'g'ri ajratgan, k-meansga qaraganda ko'p vaqt va xotira ishlatgan bo'lsada, undagi klasterlar sonini avtomatik aniqlagani uchun bu algoritmlarni ham optimal deb hisoblashimiz mumkin.

Xulosa

Klasterizatsiya — ma'lumotlarni ma'lum bir mezonlarga asoslangan holda guruhlariga ajratish usuli bo'lib, u katta hajmdagi ma'lumotlar bilan ishlashda samarali vosita hisoblanadi. Bu usul orqali bir-biriga o'xshash obyektlar bir klasterga jamlanib, o'xshamagan obyektlar esa boshqa klasterlarga ajratiladi. Klasterizatsiya turli sohalarda, jumladan, marketing, tibbiyot, biologiya, ijtimoiy tarmoqlar va boshqa ko'plab yo'nalishlarda qo'llaniladi. Maqolada klasterizatsiyaning asosiy usullari, jumladan, K-means, iyerarxik klasterizatsiya va DBSCAN usullari tahlil qilindi. Har bir usulning afzalliklari va kamchiliklari ko'rib chiqilib, ularning qaysi sharoitda samaraliroq ekani haqida tushuncha berildi. Shuningdek, klasterizatsiya jarayonida optimal klaster sonini aniqlash muhimligini, bu esa natijaga katta ta'sir ko'rsatishini ta'kidladik. Klasterizatsiya algoritmlari rivojlanishi bilan katta hajmdagi ma'lumotlarni samarali tahlil qilish va qaror qabul qilish jarayonlarini soddalashtirish imkoniyati oshmoqda. Shu sababli, zamonaviy axborot texnologiyalari davrida klasterizatsiya usullari nafaqat ilmiy izlanishlarda, balki kundalik hayotda ham keng qo'llanilmoqda.

Foydalanilgan adabiyotlar

1. Tan, P.-N., Steinbach, M., Kumar, V. Introduction to Data Mining
2. Jain, A. K., Murty, M. N., Flynn, P. J. Data Clustering: A Review
3. Bishop, C. M. Pattern Recognition and Machine Learning
4. Han, J., Kamber, M., Pei, J. Data Mining: Concepts and Techniques
5. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction
6. Xu, R., Wunsch, D. Clustering
7. MacQueen, J. B. Some Methods for Classification and Analysis of Multivariate Observations
8. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise
9. Kaufman, L., Rousseeuw, P. J. Finding Groups in Data: An Introduction to Cluster Analysis
10. Aggarwal, C. C., Reddy, C. K. Data Clustering: Algorithms and Applications