KLASTERLASH (CLUSTERING) USULLARI VA ULARNING AVZALLIKLARI

Mo‘ydinova Madinaxon Dilshodjon qizi
FarDU Axborot tizimlari va texnologiyalari yo‘nalishi 3-kurs talabasi
madinichkamadi1055@gmail.com
Sobirjonov Behzod Qahramonovich
FarDU Axborot texnologiyalari kafedrasi Katta o‘qituvchisi
bekzodbekqahromonovich@gmail.com
https://doi.org/10.5281/zenodo.20355123

Annotatsiya: Ushbu maqola klasterlash metodlari va ularning turli sohalarda qo‘llanilishini tahlil etadi. Klasterlash -bu ma’lumotlar to‘plamini o‘xshashlik yoki masofa asosida bir-biriga yaqin guruhlarga ajratish jarayoni. Bizning tadqiqotimiz klasterlashning o‘rtacha algoritmi, maksimum masofa algoritmi, ISODATA algoritmi va maksimallashtirishni kutish algoritmi kabi asosiy usullariga qaratilgan. Har bir algoritmning matematik asoslari, afzalliklari va cheklovlari ko‘rib chiqiladi, shuningdek ularning amaliy dasturlari misollar orqali tushuntiriladi. Bunda klasterlash (clustering) modellarining ahamiyati, ularning ishlash prinsiplari, afzallik va kamchiliklari tahlil qilinadi. Tadqiqotda K-means, Hierarchical, DBSCAN, Gaussian Mixture Model (GMM) va BIRCH kabi asosiy modellar ko‘rib chiqilgan. Shuningdek, real amaliyotlarda – marketing, elektron tijorat va moliya sohalarida – klasterlash modellarining qo‘llanilishi misollar yordamida yoritilgan. Maqola natijalari mijozlar xatti-harakatlarini chuqur tahlil qilish va maqsadli marketing strategiyalarini shakllantirishda samarali vosita sifatida klasterlash usullarining dolzarbligini ko‘rsatadi.

Kalit so‘zlar: Klaster yondashuvi ,klasterlash, K-means, DBSCAN, GMM, marketing tahlili. Tibbiyot diagnostikasi, Transport optimallashtirish, Dinamik tizimlar

CLUSTERING METHODS AND THEIR ADVANTAGES

Mo‘ydinova Madinaxon Dilshodjon qizi
3rd-year student of the Information Systems and Technologies program, FarDU
madinichkamadi1055@gmail.com
Sobirjonov Behzod Qahramonovich
Senior Lecturer, Department of Information Technologies, FarDU
bekzodbekqahromonovich@gmail.com

Abstract: This article analyzes clustering methods and their applications across various fields. Clustering is the process of dividing a dataset into groups that are close to each other based on similarity or distance. Our research focuses on the main methods of clustering, including the K-means algorithm, maximum distance algorithm, ISODATA algorithm, and Expectation-Maximization algorithm. The mathematical foundations, advantages, and limitations of each algorithm are examined, and their practical applications are explained through examples. In addition, the importance of clustering models, their working principles, advantages, and disadvantages are analyzed. The study considers major models such as K-means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Model (GMM), and BIRCH. Furthermore, the application of clustering models in real-world practices — including marketing, e-commerce, and finance — is illustrated with examples. The results of the article demonstrate the relevance of clustering methods as an effective tool for in-depth analysis of customer behavior and the development of targeted marketing strategies.

МЕТОДЫ КЛАСТЕРИЗАЦИИ И ИХ ПРЕИМУЩЕСТВА

Мўйдинова Мадинахон Дилшоджон қизи
студентка 3 курса направления «Информационные системы и технологии», ФерГУ
madinichkamadi1055@gmail.com
Собиржонов Бехзод Кахрамонович
старший преподаватель кафедры информационных технологий ФерГУ
bekzodbekqahromonovich@gmail.com

Keywords: Cluster approach, clustering, K-means, DBSCAN, GMM, marketing analytics, medical diagnostics, transport optimization, dynamic systems.

Аннотация: В данной статье анализируются методы кластеризации и их применение в различных областях. Кластеризация — это процесс разделения набора данных на группы, близкие друг к другу на основе сходства или расстояния. Наше исследование сосредоточено на основных методах кластеризации, таких как алгоритм K-means, алгоритм максимального расстояния, алгоритм ISODATA и алгоритм ожидания-максимизации. Рассматриваются математические основы, преимущества и ограничения каждого алгоритма, а также их практическое применение на примерах. Кроме того, анализируются значение моделей кластеризации, принципы их работы, преимущества и недостатки. В исследовании рассмотрены основные модели, такие как K-means, иерархическая кластеризация, DBSCAN, Gaussian Mixture Model (GMM) и BIRCH. Также с помощью примеров освещается применение моделей кластеризации в реальной практике — в маркетинге, электронной коммерции и финансовой сфере. Результаты статьи показывают актуальность методов кластеризации как эффективного инструмента для глубокого анализа поведения клиентов и формирования целевых маркетинговых стратегий.

Ключевые слова: кластерный подход, кластеризация, K-means, DBSCAN, GMM, маркетинговая аналитика, медицинская диагностика, оптимизация транспорта, динамические системы.

Kirish. Klasterlash – 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. Klasterlashning asosiy vazifasi o‘xshash xususiyatlarga ega ma’lumotlarni bitta guruhga birlashtirish va farqli xususiyatlarga ega bo‘lganlarni boshqa guruhlarga ajratishdan iboratdir. Klasterlash 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:

1.Mijozlarning o‘ziga xos birlashmalarining xatti-harakatlarini tahlil qilish.

2.Biznes raqiblari - bozorni o‘rganishda.

3.Kasalliklar - tiklanish statistikasini o‘rganish.

4.So‘rov ishtirokchilari - turli guruhlardagi odamlarning fikrlarini tahlil qilish.

5.SEO kalitlari - veb-sayt sahifalarida mavzular yaratish uchun.

6.Olingan fayllar qulay ishlov berish uchun turli formatlarga ega.

Klasterlash turli sohalarda 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 guruhlarga 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 guruhlarga 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.

Natija va muhokama

Bugungi kunda klasterizatsiya masalasini yechish uchun ko‘plab uslublar va ular asosida birnechta 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, ketmaketlikka asoslangan fuzzy c-means, ierarxik uchun CobWeb, extimollik bo‘yicha qidiruvchi algorimlarga esa misol qilib EM algoritmini aytishimiz mumkin. Weka API.Yuqorida takidlangab o‘tkanimizdek hech qaysi klasterizatsiya algoritmi istalgan obyektlar to‘plami uchun eng optimal bo‘la olmaydi. Shu sababli biz katta obyektlar to‘plamiz ixtiyoriy tanlangan qismini ajratib olgan xolda ular ustida birnechta eng ko‘p qo‘llaniladigan algoritmlar bilan tajriba o‘tkazib ularni solishtrib ko‘rishimiz kerak. Buni oson hal qilish uchun Weka API (Application Programming Interface) dan foydalanish ish jarayonini osonlashtiradi. Weka API Yangi Zerlandiyaning Waikato Universiteti tomonidan Ma’lumotlarni intellectual taxlili masalalarini yechish uchun ishlab chiqilgan bo‘lib, sinflarga ajratish, klasterizatsiya, bashoratlash, assotativ qoidalarni qurish va vizualizatsiya masalalarini yechish uchun bir nechta algotirmlarni 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.Ma’lumotlarni intelektual tahlilida k-means klasterizatsiya algoritmi eng sodda, eng tushunarli va eng ko‘p ishlatiladigan algoritmlardan biri xisoblanadi. K-means algoritmi berilgan n ta obyektdan 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 algoritmida obyektlar orasidagi masofalarni hisoblash uchun Evklid masofasi, Manhetton masofasi kabilar ishlatiladi. Algoritmni asosiy abzalligi uni ishlash tezligida, k-means boshqa olgaritmlarga 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‘plamalar uchun EM ham yaxshi natija ko‘rsata oladi. Bu algoritm 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 guruhlarga 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 algoritm 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. Algoritm k-means, EM algoritmlarida uchramaydigan ko‘plab imkoniyatlarga ega. Masalan CobWeb algoritmi dinamik ma’lumotlarni klasterlashda ham ishlatiladi. Yangi obyektni 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 algoritm k-means kabi barcha xolatlar uchun yaxshi yechim topa olmaydi, ko‘pincha klasterlarga ajratishda xatoliklarga yo‘q qo‘yadi. Ma’lumotlarni berilish taribi, klasterlash natijasiga tasir ko‘rsatadi. DBScan. DBScan algoritmi Martin Ester, Hans-Peter, Jorge Sander va Xiaowei Xu tomonidan 1996 yilda yaratilgan. Bu algoritm zichlikka asoslangan. Klasterlar soni o‘zi aniqlab oladi. Obyektlarni berilish tartibini axamiyati yo‘q, har qanday tartibda berilganda ham bir xil natija chiqaradi. Bu

algoritm bugungi kundagi eng optimal algoritmlardan biri xisoblanadi. Bu algoritmni asosiga OPTICS algoritmi ham qurilgan. Tezkor, klasterlash uchun k-means kabi samarali algoritm. Kamchiligi sifatida bu algoritmni masofani topish bo‘yicha ishlashida deb aytish mumkin. Chunki bazi holatlarda masofani Evklid masofasi bo‘yicha olgan foydali bo‘lsa, bazan Manhetton 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 algoritm 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: Klasterlash — ma’lumotlarni ma’lum bir mezonlarga asoslangan holda guruhlarga 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.

Adabiyotlar, References, Литературы:

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