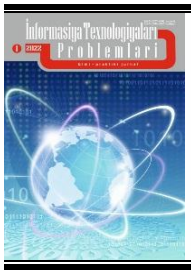


Available online at www.jpit.az

13 (1)
2022

Modification of the DBSCAN algorithm for big data clustering

Aygul F. Fakhraddingizi

Institute of Information Technology, Azerbaijan National Academy of Sciences, B. Vahabzade str., 9A, AZ1141 Baku, Azerbaijan

aygul.fexreddin@gmail.com

ARTICLE INFO

http://doi.org/10.25045/jpit.v13.i1.04

Article history:
Received 14 July 2021
Received in revised form 23 September 2021
Accepted 1 December 2021

Keywords:
Big data
Clustering algorithms
Density-based clustering
DBSCAN algorithm

ABSTRACT

The development of Information and Communication Technologies (ICT) has led to the rapid growth of digital information and the consequent emergence of the concept of large-scale data. Therefore, there is a need to delve into large-scale data and its essence, the possibilities and problems of analytical technologies. Clustering is one of the main methods of analyzing big data. The main purpose of clustering is to separate data into clusters according to certain characteristics. When clusters come in different sizes, densities, and shapes, the problem of detection arises. The article explores the density-based DBSCAN clustering algorithm for working with big data. One of the main features of this algorithm is to create an effective cluster by detecting the noise points in big data. During the implementation of the algorithm, real databases containing noise points were used. Metrics such as adjusted rand index, homogeneity, Davis-Boldin index were used to evaluate the results of the experiment. The proposed method was more effective than the traditional DBSCAN algorithm in detecting noise points.

Böyük həcmli məlumatların klasterləşdirilməsi üçün DBSCAN alqoritminin modifikasiyası

Açar sözlər:
Böyük ölçülü verilənlər
Klasterləşdirmə alqoritmləri
Sıxlığa əsaslanan klasterləşdirmə DBSCAN alqoritmi

İnformasiya-kommunikasiya texnologiyalarının (İKT) inkişafı rəqəmsal informasiyanın sürətli artımına və nəticədə böyük ölçülü verilənlər konsepsiyasının meydana gəlməsinə səbəb olmuşdur. Bu səbəbdən böyük ölçülü verilənlər və onun mahiyyətini, analiz texnologiyalarının imkanlarını, problemlərini hərtərəfli tədqiq etməyə ehtiyac yaranmışdır. Klasterləşdirmə böyük ölçülü verilənlərin əsas analiz üsullarından biridir. Klasterləşdirmənin əsas məqsədi verilənləri müəyyən xüsusiyyətlərə görə klasterlərə ayırmaqdır. Klasterlər fərqli ölçüdə, sıxlıqda və formada olduqda onların aşkarlanması problemi ortaya çıxır. Məqalədə böyük ölçülü verilənlərlə işləmək üçün sıxlığa əsaslanan DBSCAN klasterləşdirmə alqoritmi tədqiq edilmişdir. Bu alqoritmin əsas xüsusiyyətlərindən biri böyük ölçülü verilənlərdə küy nöqtələri aşkar etməklə effektiv klasterin yaradılmasından ibarətdir. Alqoritmin tətbiqi zamanı küy nöqtələri ehtiva edən real verilənlər bazalarından istifadə edilmişdir. Eksperimentin nəticələrinin qiymətləndirilməsi üçün nizamlanmış rand indeksi, bircinslik, Devis–Boldin (Davies-Bouldin) indeksi və s. kimi metrikalar istifadə edilmişdir. Təklif edilən metod ənənəvi DBSCAN alqoritminə nisbətən küy nöqtələri aşkar etməkdə daha effektiv nəticə göstərmişdir.

Модификация алгоритма DBSCAN для кластеризации больших данных

Ключевые слова:
Большие данные
Кластеризация
Алгоритмы кластеризации
Кластеризация на основе плотности
Алгоритм DBSCAN

Развитие информационных и коммуникационных технологий (ИКТ) привело к быстрому росту цифровой информации и последующему появлению концепции крупномасштабных данных. Поэтому возникает необходимость вникать в крупномасштабные данные и их сущность, возможности и проблемы аналитических технологий. Кластеризация—один из основных методов анализа больших данных. Когда кластеры бывают разного размера, плотности и формы, возникает проблема обнаружения. В статье исследуется алгоритм кластеризации DBSCAN на основе плотности для работы с большими данными. Одной из основных особенностей этого алгоритма является создание эффективного кластера путем обнаружения точек шума в больших данных. При реализации алгоритма использовались реальные базы данных, содержащие шумовые точки. Для оценки результатов эксперимента использовали такие показатели, как скорректированный ранд-индекс, однородность, индекс Дэвиса-Болдина. Предложенный метод оказался более эффективным, чем традиционный алгоритм DBSCAN, при обнаружении шумовых точек.

1. Giriş

XXI əsrin əvvəllərindən başlayaraq texnika və texnologiyalar – kompüterlər, mobil telefonlar, İnternet, sensor şəbəkələri, yerin süni peykləri, kosmik teleskoplar, bulud hesablamaları və s. vasitəsi ilə generasiya olunan rəqəmsal verilənlər hər il həndəsi silsilə ilə artmaqdadır. Nəticədə verilənlərin emalı, idarə olunması, saxlanması və istifadəsində yeni eranı əks etdirən "böyük ölçülü verilənlər" (big data) konsepsiyası meydana çıxmışdır (Əliquliyev, 2014; Əliquliyev, Hacırəhimova, & Əliyeva, 2016). Big data fenomenal hadisə olaraq cəmiyyətin iqtisadi inkişafında inqilabi dəyişikliklərlə yanaşı, elmi ictimaiyyəti bir sıra problemlərlə üz-üzə qoymuş, yeni tədqiqat paradiqması yaratmışdır. Bu verilənlərin emalı üçün yeni texnologiyalardan istifadə etmək zərurəti yaranmışdır. Başqa sözlə, big data termini həcm və müxtəliflik baxımından mürəkkəb olan məlumatları ifadə edir, lakin onları ənənəvi emal texnologiyaları vasitəsilə idarə etmək mümkün olmadığına görə real zaman rejimində yeni biliklərin əldə edilməsi çətinləşir.Big datanın analizi vasitəsilə qiymətli məlumatlar və faydalı biliklər əldə etmək olar. Ancaq Big Datanın inkişafı təhlükəsizlik və gizlilik risklərini də özü ilə bərabər gətirir (big data, 2008; Alguliyev, & Imamverdiyev, 2014). Həmçinin məlumatların toplanması, saxlanması və istifadəsi zamanı fərdi məlumatların asanlıqla sızmasına və məlumatların klassifikasiyası zamanı çətinliklərə səbəb olur. Böyük ölçülü verilənlərin təhlükəsizliyini təmin etmək və gizliliyini qorumaq cari araşdırmalar mərhələsində ən aktual problemlərdən birinə çevrilmişdir (Alguliyev, Aliguliyev, & Sukhostat, 2020). Klasterləşdirmə böyük ölçülü verilənlərin əsas analiz üsullarından biridir (Fəxrəddinqızı A., 2019). Məqalədə DBSCAN klasterləşdirmə alqoritminin tətbiqi ilə eksperiment həyata keçirilir. Müxtəlif qiymətləndirmə indeksləri vasitəsilə nəticələr qiymətləndirilir. Məqalənin sonrakı hissələri aşağıdakı şəkildə strukturlaşdırılmışdır: ikinci bölmədə tədqiq olunan problemlə əlaqəli işlərin qısa icmalı, üçüncü bölmədə klasterləşdirmə alqoritmlərinin təsnifatı verilmişdir. Yekun – təklif olunan metodun mərhələlərinin təsviri və eksperimentlərin analizi isə dördüncü və beşinci bölmədə verilmişdir.

2. Əlaqəli işlər

Sıxlığa əsaslanan bir sıra klasterləşdirmə alqoritmləri mövcuddur. Bu alqoritmlər bir-birindən bir sıra mənfi və ya müsbət cəhətlərinə görə fərqlənirlər. Məsələn, bu alqoritmlərdən biri Ankerst tərəfindən tədqiq edilmiş OPTİCS alqoritmidir (Zhou, Pan, Wang, & Vasilakos, 2017). Belə ki, bu alqoritm DBSCAN-ın bir sıra zəif cəhətlərini, məsələn, fərqli sıxlıqdakı verilənlərdə əhəmiyyət kəsb edən klasterlərin aşkarlanması problemini aradan qaldırır. Bu alqoritm verilənlər çoxluğundan geniş bir sıralama düzəldərək fəza verilənlərində sıxlığa əsaslanan klasteri tapmaq üçün istifadə olunan bir alqoritmdir. Sıxlığa əsaslanan digər alqortmlərdən biri Liu tərəfindən təklif olunmuş VDBSCAN (Liu, Zhou, Wu, 2007) alqoritmidir. Bu alqoritm DBSCAN-ın sıxlıq problemini həll etmək üçün fərqli sıxlıqlı verilənlər dəstinin təhlili məqsədi ilə yaradılmışdır. VDBSCAN-ın əsas ideyası ənənəvi DBSCAN alqoritmini qəbul etməzdən əvvəl epsilon parametrinin bir neçə qiymətini seçmək üçün bəzi metodlardan istifadə olunmasıdır. Eps-in fərqli qiymətləri ilə müxtəlif sıxlığa malik klasterləri təyin etmək mümkündür. Digər bir alqoritm LDBSCAN (Duan, Xu, Guo, Lee, Yan, 2007) lokal sıxlığa əsaslanır. Bu üsulda uyğun parametrlərin seçilməsi çətin deyil, eyni zamanda, sıxlığa əsaslanan digər klasterləşdirmə alqoritmləri ilə müqayisədə küy nöqtələri aşkar etmək üçün anomaliyaların aşkarlanmasında istifadə olunan lokal kənaraçıxma faktorunun (lokal outlier factor, LOF) üstünlüyündən istifadə edir. AUTOEPSDBSCAN alqoritmi isə giriş parametrlərini avtomatik olaraq seçən genişləndirilmiş bir alqoritmdir (Gaonkar, & Sawant, 2013). Eksperimental nəticələr göstərir ki, AUTOEPSDBSCAN alqoritmi küy və kənaraçıxma nöqtələrə sahib nöqtələrdən ibarət böyük ölçülü verilənlərdə müxtəlif forma və ölçüyə malik klasterləri aşkar edə bilir. Qeyd edilən bütün bu alqoritmlər böyük ölçülü verilənlər bazası ilə əlaqədar problemlərin həllində istifadə olunur. Bu alqoritmlər klasterləri müəyyən etmək üçün Eps parametrindən istifadə edir, buna görə də müxtəlif sıxlıqlara əsasən Eps parametrinin eyni qiymətində bir klaster digərlərinə görə daha sıx ola bilər.

3. Klasterləşdirmə alqoritmləri və onların təsnifatı

"Maşın təlimi (Machine Learning)" böyük ölçülü verilənlərdə istifadə olunan süni intellektin (Artificial İntelligence) vacib bir sahəsidir. Maşın təliminin əsas məqsədi biliyi aşkarlamaq, düzgün qərarlar vermək və verilənləri analiz etməkdən ibarətdir (Andrieu, De Freitas, Doucet, & Jordan, 2003; Berkhin, Kogan, Nicholas, 2006). Maşın təlimi alqoritmləri təlimsiz (supervised), təlimlə (unsupervised) və yarı təlimlə (semi-supervised) öyrənmə üsullarına görə kateqoriyaya bölünür. Maşın təlimi alqoritmlərini klassifikasiya, klasterləşdirmə, reqressiya, sıxlıq qiymətləndirməsi və s. kimi də təsnifatlandırmaq olar. Maşın təlimi alqoritmlərinə digər nümunə olaraq qərarlar ağacı, süni neyron şəbəkələr, SVM, Bayes şəbəkələri, genetik alqoritmləri və s. qeyd etmək olar. Təlimlə öyrənmə alqoritmlərinə Naive Bayes, SVM (Support Vector Machine) və maksimal entropiya metodu (MaxENT) və s. aiddir.

Klasterləşdirmə alqoritmləri təlimsiz öyrənmə metodlarına daxildir. Təlimsiz öyrənmə alqoritmləri qruplaşdırılmamış verilənləri onların xüsusiyyətlərinə görə müqayisə edir və uyğun qruplara ayıraraq, təsnifatlaşdırır. Yəni, təlimsiz öyrənmə alqoritmləri oxşar obyektləri eyni qrupda birləşdirir. Belə ki, verilənlər bazasındakı məlumatları şərh etməklə ümumi nöqtələri tapır və onları qruplaşdıraraq oxşar məlumatlar əldə edir. Təlimsiz öyrənmə alqoritmləri oxşarlıq/fərqlilik ölçüsü verildikdə klaster daxilindəki obyektlər arasında oxşarlığı artırır. Lakin klasterlərarası oxşarlıq bir-birindən ciddi şəkildə fərqlənir. Burada xüsusi bir obyektiv funksiyadan istifadə olunur. Təlimsiz öyrənmə alqoritmlərinə klasterləşdirmə (k-means, sıxlığa əsaslanan, iyerarxik və s.), özünü təşkil edən xəritələr (SOM, self-organizing map) və s. daxildir (Fahad et al., 2014). Klasterləşdirmə alqoritmləri yeni texnologiyaların tətbiqi nəticəsində böyük ölçülü verilənlərin həcmini dəqiq analiz etmək üçün alternativ, daha güclü bir meta öyrənmə vasitəsi olaraq meydana gəlmişdir. Qeyd etdiyimiz kimi bir sıra klasterləşdirmə alqoritmləri mövcuddur. Bir neçəsi ilə tanış olaq:

Bölünməyə əsaslanan klasterləşdirmə alqoritmlərində qruplar dərhal təyin olunur. İlkin qruplar müəyyən edilir və birliyə doğru yenidən paylanılır. Başqa sözlə, bölünməyə əsaslanan alqoritmlər məlumat obyektlərini bir neçə bölməyə bölür, burada hər bölmə çoxluq təşkil edir. Başqa sözlə, bölünməyə əsaslanan alqoritmlər, məlumat obyektlərinin bölüm sayına bölünməsi tapşırığını yerinə yetirir, hər bölmə "klaster" adlanır (Sajana, Rani, & Narayana, 2016; Zhao, Ma, & He, 2009).

Məsafə metrikasının rolu bütün alqoritm növlərində fərqlidir. Bölünməyə əsaslanan klasterləşdirmə metodlarında məsafə metrikası fərqli iterasiyalarda seçilmiş şablon nöqtələr klasterin sentroidi (verilənlər mövcud olmadıqda) kimi faktiki nöqtələr ola bilər.

Bölünməyə əsaslanan klasterləşdirmə metodları ilə əlaqəli bəzi çatışmazlıqları aradan qaldırmaq üçün iyerarxik klasterləşdirmə alqoritmləri hazırlanmışdır. Bildiyimiz kimi bölünməyə əsaslanan klasterləşdirmə alqoritmləri, ümumiyyətlə, keyfiyyətli klasterlər əldə etmək üçün əvvəlcədən istifadəçinin təyin etmiş K parametrini saxlayır və bu alqoritm təbiətinə görə qeyri-müəyyəndir. Məlumat obyektlərinin klasterləşdirilməsi zamanı müəyyənləşdirici və əlçatan bir mexanizm yaratmaq üçün iyerarxik alqoritmlər hazırlanmışdır (Chana, & Arora, 2014; Chandra, & Anuradha, 2011). Bu alqoritmdə məlumatlar yaxınlıq dərəcəsindən asılı olaraq iyerarxik şəkildə təşkil olunur. Yaxınlıq aralıq qovşaqlarla əldə edilir. İyerarxiya davam etdikcə ilkin klaster tədricən bir neçə qrupa bölünür. İyerarxik alqoritmləri aqlomerasiya (agglomerative - aşağıdan yuxarıya doğru bir yanaşmadır: hər müşahidə öz klasterində başlayır və iyerarxiyaya doğru irəlilədikcə cüt-cüt birləşir) və ya bölünmə (divisive - yuxarıdan aşağıya doğru bir yanaşmadır: bütün müşahidələr bir qrupda başlayır və iyerarxiya aşağı düşdükcə parçalanmalar rekursiv şəkildə aparılır) üsullarına görə təsnif etmək olar. Aqlomerativ metod zamanı alt səthdə (klasterdə cəmi bir məlumat obyekti olan) bir klaster götürərək başlayır və klasterlərin aşağıdan yuxarıya iyerarxiyasını qurmaq üçün hər iterasiya zamanı iki klasteri birləşdirməyə davam edir. Belə ki, bu üsulla klasterləşməni həyata keçirmək üçün hər klaster bir obyektdən başlayır və iki və ya daha çox uyğun klasteri rekursiv şəkildə birləşdirir. Digər tərəfdən ayrılma üsulu nəhəng bir makro-klasterdəki bütün məlumat obyektləri ilə başlayır

və davamlı olaraq yuxarıdan aşağıya doğru klaster iyerarxiyası təşkil edərək iki qrupa bölünür. Yəni, bu üsulda klasterlər vahid klaster kimi bir verilənlər bazasından başlayır və ən uyğun klasteri rekursiv şəkildə bölür. Proses dayanma meyarına çatana qədər davam edir. (Karypis, Han, & Kumar, 1999) Bununla yanaşı, iyerarxik metodun əhəmiyyətli bir çatışmazlığı var ki, bu da bir addım icra edildikdən sonra (birləşmə və ya bölünmə) geri qaytarıla bilməməsi ilə əlaqədardır.

4. Sıxlığa əsaslanan klasterləşdirmə alqoritmləri

Bir çox məşhur klasterləşdirmə alqoritmlərinin verilənlərin müəyyən bir növünün ehtimal paylanmasından yarandığı ehtimal edilir. Xüsusilə EM (Expectation Maximization) və k-means klasterləşdirmə alqoritmləri üçün bu hal daha uyğundur. Bu fərziyyəyə görə bu alqoritmlər sferik çoxluqlar əmələ gətirir və faktiki çoxluqların qabarıq formalara malik olan verilənlər bazasında yaxşı işləyə bilmir. Qabarıq çoxluqlar təbii olaraq fəza verilənlərində, yəni real dünyadan fərqli iki və ya üçölçülü fəzalarda meydana gəlir. Fəza nöqtələri dağlar və çaylar kimi coğrafi obyektlər tərəfindən qoyulan məhdudiyyətlər səbəbilə ixtiyari formaya sahib ola bilər. Bu vəziyyətdə k-means kimi alqoritmlər real qrupları parçalayaraq və ya birləşdirərək yanlış nəticələrə səbəb olacaqdır. Bu müşahidə ixtiyari formalı klasterlərin tapılmasına gətirib çıxarır. Getdikcə artan real böyük verilənlər bazalarında səmərəliliyə ehtiyac duyulur. Böyük ölçülü verilənlər bazalarının klasterləşdirilməsi zamanı küyün və kənaraçıxmaların aşkarlanması və aradan qaldırılması tələb olunur (Alguliyev, Aliguliyev, & Abdullayeva, 2019). Bütün bu tələbləri yerinə yetirmək üçün sıxlığa əsaslanan klasterləşdirmə alqoritmlərinin paradiqması təklif edilmişdir. Sıxlığa əsaslanan klasterləşdirmə qeyri-parametrik metod hesab edilə bilər, çünki klasterlərin sayı və ya onların paylanması ilə bağlı heç bir fərziyyə yoxdur. Sıxlığa əsaslanan klasterlər bir-birindən seyrək (sparser) sahələr ilə ayrılmış məlumat sahəsindəki sıx sahələrdir. Bundan əlavə, küy sahələri arasındakı sıxlığın hər hansı bir klasterdəki sıxlıqdan daha aşağı olduğu qəbul edilir. Təbiətinə görə məlumat sahəsindəki sıx sahələr ixtiyari formaya malik ola bilər (Dharni, & Bnasal, 2013; El-Sonbaty Y, Ismail, Farouk, 2004). Sahə sorğularını dəstəkləyən bir indeks quruluşunu nəzərə alsaq, sıxlığa əsaslanan klasterlər verilənlər bazası obyekt başına ən çox sahə sorğusunu yerinə yetirməklə səmərəli hesablanır. Verilənlər sahəsindəki seyrək sahələr küy kimi qəbul edilir və heç bir klasterə aid edilmir. Ədəbiyyatlarda sıxlığa əsaslanan klasterləşdirmə alqoritmləri haqqında bəzi fikirlərin olduğunu qeyd etmək lazımdır. İlk növbədə r məsafəsində yerləşən k-dan az qonşusu olan heç bir klasterə aid olmayan (nondense) nöqtələr çıxarılır. İkincisi, qalan nöqtələri klasterləşdirmək üçün tək əlaqə üsulundan istifadə olunur. Nəhayət, bəzi meyarlara görə klasterlərdən birində nondense nöqtələr təyin edilir. Sıxlığa əsaslanan klasterləşdirmə və orta növbəli (mean-shift) klasterləşdirmə paradiqmalarının əlaqəsini də qeyd etmək olar. Sıxlığa əsaslanan klasterləşdirmə alqoritmi yaradılarkən bir neçə əsas suallara cavab verilməlidir (Parimala, Lopez, & Senthilkumar, 2011):

- Sıxlıq necə qiymətləndirilir?
- Bağlantı necə qurulur?
- Hansı məlumat strukturları alqoritmin səmərəli həyata keçirilməsini dəstəkləyir?

Növbəti bölmədə sıxlığa əsaslanan klasterləşdirmə alqoritmləri təqdim olunacaq və bu suallara cavab verməyin yolları müzakirə olunacaq.

4.1. DBSCAN alqoritmi

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 1996-cı ildə Martin Ester, Hans-Peter Kriegel, Jörg Sander və Xiaovei Xu tərəfindən irəli sürülmüş bir klasterləşdirmə alqoritmidir (Ester, Kriegel, Sander, & Xu, 1996). Bu alqoritm sıxlığa əsaslanan klasteri təşkil edən qeyri-parametrik bir alqoritmdir: bəzi fəza nöqtələr toplusu verilərək, bir-birinə yaxın yerləşmiş nöqtələri (yaxın qonşuları olan nöqtələri) bir yerə toplayır, ancaq aşağı sıxlığa (seyrək) malik sahələrdəki tək qalan nöqtələri (ən yaxın qonşuları çox uzaq məsafədə yerləşən) kənaraçıxma (outlier) nöqtələr kimi qeyd edir. DBSCAN elmi ədəbiyyatda da ən çox istinad edilən alqoritmlərdən biridir (Cassisi et al., 2013).

DBSCAN alqoritmi sabit radiuslu qonşuluqda yerləşən nöqtələrin sayını hesablayaraq sıxlığı qiymətləndirir və hər hansı iki nöqtə bir-birinin qonşuluğunda yerləşirsə, bu nöqtələri bir-birilə bağlı hesab edir. DBSCAN

alqoritminin iki əsas parametri mövcuddur: Eps (Epsilon) – qonşuluqları təyin edən məsafə. İki nöqtə arasındakı məsafə Eps-dən az və ya bərabər olduqda qonşu hesab olunur (Rahmah, & Sitanggang, 2016).

$MinPts$ (Minimum Points) - klasteri təyin etmək üçün minimum məlumat nöqtələrinin sayı. Bu iki parametrə əsasən sıxlığa əsaslanan klasterləşdirmə alqoritmi nöqtələri üç fərqli nöqtələr tipinə ayrır:

- əsas nöqtələr (core points), yəni sıx qonşuluqda yerləşən nöqtələr ($|NEps(p)| \geq MinPts$); qonşuluq radiusundakı Eps ən azı $MinPts$ nöqtəsindən ibarətdirsə, yəni qonşuluqdakı sıxlıq bəzi həddi keçməlidirsə, bu nöqtə *əsas nöqtə* adlanır.
- sərhəd nöqtələr (border points), yəni hər hansı klasterə aid olan, ancaq sıx qonşuluqda yerləşməyən nöqtələr; bir nöqtə əsas nöqtədən əldə edilə biləndirsə və ətrafındakı sahədə MinPts nöqtəsindən az nöqtə varsa, bu nöqtə *sərhəd nöqtə* adlanır.
- küy nöqtələr (noise points), yəni heç bir klasterə aid olmayan nöqtələr; əgər bir nöqtə əsas nöqtə deyilsə və hər hansı əsas nöqtədən əldə edilə bilən nöqtə deyilsə, bu nöqtə *küy nöqtə* kimi qiymətləndirilir (Moreira, Santos, & Carneiro, 2005).

4.2. DBSCAN alqoritminin analizi

Tutaq ki, D məlumat nöqtələrindən ibarət verilənlər bazası verilmişdir. Cüt nöqtələr üçün $dist(p,q)$ məsafə funksiyasının olduğunu fərz edək. $NEps(p)$ ilə işarələnmiş p nöqtəsinin Eps qonşuluğu $NEps(p) = \{ q \in D | dist(p,q) \leq Eps\}$ kimi müəyyən edilir [30_].

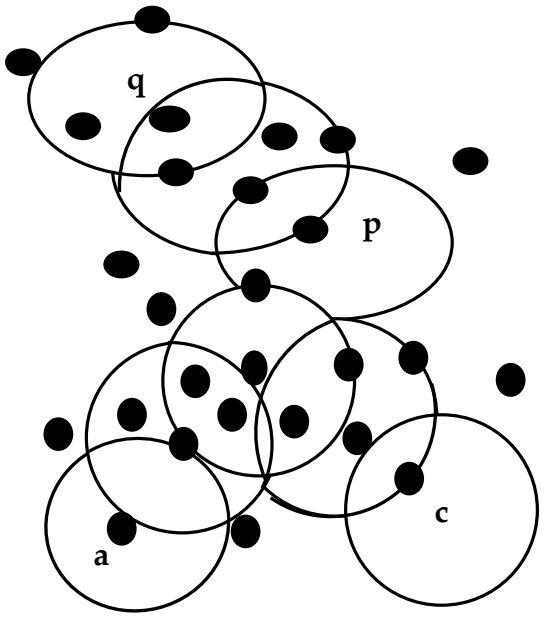
Tərif 1. Əgər (1) $p \in NEps(q)$ və (2) $|NEps(q)| \geq MinPts$ kimidirsə, Eps və $MinPts$ parametrlərinə görə p nöqtəsi q nöqtəsindən *birbaşa sıxlıq əldə edilə bilən* (*directly density-reachable*) nöqtədir.

Tərif 2. $p_1, \dots, p_n,\ p_1 = q,\ p_n = p$ nöqtələr ardıcıllığında p_{i+1} nöqtəsi p_i nöqtəsindən birbaşa sıxlıq əldə edilə bilən olduğundan Eps və $MinPts$ parametrlərinə görə p nöqtəsi q nöqtəsindən *sıxlıq əldə edilə bilən* (*density - reachability*) nöqtədir. Sıxlıq əldə edilə bilən birbaşa sıxlıq əldə edilə bilən nöqtənin kanonik uzantısıdır. Bu əlaqə keçici olmadığından başqa bir əlaqə tətbiq olunur.

Tərif 3. Əgər Eps və $MinPts$ parametrlərinə görə o nöqtəsi hər iki p və q nöqtəsindən sıxlıq əldə edilə bilən nöqtədirsə, onda həmin parametrlərə görə p nöqtəsi q nöqtəsindən *sıxlığa bağlı (density-connected)* nöqtədir. Şəkil 1-də bu bağlılıq göstərilmişdir. Belə ki, p nöqtəsi q nöqtəsindən sıxlıq əldə edilə bilən nöqtədirsə, ancaq q nöqtəsi p nöqtəsindən sıxlıq əldə edilə bilən deyilsə, onda a və c nöqtələri b nöqtəsindən sıxlığa bağlı nöqtələrdir. İntuitiv olaraq, sıxlığa bağlı nöqtələr çoxluğu sıxlıq əldə edilə bilən nöqtələr çoxluğunun maksimalıdır (Sharma, Sharma, & Soni, 2017).

Formal olaraq, Eps və $MinPts$ parametrlərinə görə C klasteri aşağıdakı iki şərti ödəyən D-nin boş olmayan alt çoxluğudur.

1. $\forall\, p, q$ üçün, əgər $p \in C$ və Eps və $MinPts$ parametrlərinə görə q nöqtəsi p nöqtəsindən sıxlıq əldə edilə biləndirsə, onda $q \in C$ (maksimum) olacaq.
2. $\forall\, p, q \in \mathrm{C}$ üçün q nöqtəsi Eps və $MinPts$ parametrlərinə görə p nöqtəsinə sıxlığa bağlıdır.



Şəkil 1. Sıxlıq əldə edilə bilən və sıxlığa bağlı nöqtələr

Tutaq ki, Eps və MinPts parametrlərinə görə $\mathrm{C}_1, \dots, \mathrm{C}_k$ D verilənlər bazasının klasterləridir. D verilənlər bazasında hər hansı C_i klasterinə aid olmayan küy nöqtələr $noise = \{p \in D | \, p \notin C_i \forall i\}$ kimi təyin olunur.

Şəkil 1-də, məsələn, q və b nöqtələri mərkəz nöqtələr, p, a və c nöqtələri isə sərhəd nöqtələrdir.

Sıxlığa əsaslanan klasterləşdirmə alqoritmləri səmərəli hesablanmalara imkan verən iki vacib xüsusiyyətə malikdir. Tutaq ki, p nöqtəsi D-nin mərkəz nöqtəsidir və D-dən çəkilən O çoxluğuna daxil olan bütün p nöqtələri Eps və $MinPts$ parametrlərinə görə sıxlıq əldə edilə biləndir. O çoxluğu Eps və $MinPts$ parametrlərinə görə klasterdir. Tutaq ki, C D-yə daxil olan klasterdir. C-nin bütün nöqtələri bu klasterin mərkəz nöqtələrindən sıxlıq əldə edilə bilən nöqtələrdir. Buna görə də C klasteri bu klasterin ixtiyari mərkəz nöqtəsindən sıxlıq əldə edilə bilən bütün nöqtələrdən ibarətdir. Beləliklə, Eps və $MinPts$ parametrlərinə əsasən C klasteri hər hansı mərkəz nöqtələrindən biri ilə unikal şəkildə müəyyənləşir.Bu da DBSCAN alqoritminin əsasını təşkil edir (Shah, 2012).

Klasteri tapmaq üçün DBSCAN alqoritmi ixtiyari verilənlər bazasına daxil olan p nöqtəsi ilə başlayır. Eps və $MinPts$ parametrlərinə əsasən p-dən sıxlıq əldə edilə bilən bütün nöqtələr alınır, ehtiyac varsa p-nin birbaşa və ya dolayı yolla qonşularını tapmaq üçün ilk p üçün sahə sorğularını yerinə yetirir. Əgər p mərkəz nöqtədirsə, bu proses Eps və $MinPts$ parametrləri əsasında bir klaster yaradır. Əgər p mərkəz nöqtə deyilsə, p-dən sıxlıq əldə edilə bilən nöqtə yoxdursa, onda DBSCAN p nöqtəsini küy nöqtə kimi təyin edir və eyni proses verilənlər bazasındakı növbəti nöqtələr üçün tətbiq olunur. Əgər p, əslində hər hansı C klasterinin sərhəd nöqtəsidirsə, onda C-nin hər hansı mərkəz nöqtəsindən sıxlıq əldə edilə bilən bütün nöqtələr bir yerə yığılır və daha sonra C klasterinə təyin olunur. Alqoritm bütün nöqtələri klasterləşdirdikdən və küy nöqtələri aşkar etdikdən sonra sona çatır (Xiong, Chen, Zhang, & Zhang, 2012).

Standart DBSCAN tətbiqetmələri müəyyən bir nöqtənin Eps qonşuluğunda yerləşən sahə sorğularını səmərəli şəkildə dəstəkləyən $R-tree$ və ya $X-tree$ kimi fəza indeksində tətbiq edir. Ən pis halda DBSCAN verilənlər bazası nöqtəsinə görə bir sahə sorğusunu yerinə yetirir. Bu DBSCAN üçün $O(nlogn)$ işləmə mürəkkəbliyinə səbəb olur, burada $n-$ verilənlər bazası nöqtələrinin sayını göstərir. Təəssüf ki, fəza indeksləri böyük ölçülü verilənlər üçün yaxşı nəticələr vermir, yəni sahə sorğularının performansı $O(nlogn)$-dən $O(n)$-ə qədər pozulur və DBSCAN-ın işləmə mürəkkəbliyi bu verilənlər üçün $O(n^2)$ kimi olur. Digər tərəfdən, $O(1)$ sahə sorğularını dəstəkləyən bir şəbəkəyə əsaslanan (grid-based) verilənlər strukturu mövcuddursa, DBSCAN-ın işləmə mürəkkəbliyi $O(n)$-ə qədər azalır. Qeyd edək ki, $O(nlogn)$ işləmə mürəkkəbliyi böyük ölçülü verilənlər çoxluğu üçün genişlənə bilər. (Kroger, Kriegel, & Kailing, 2004).

Sıxlığa əsaslanan klasterləşdirmənin əsas ideyası bir neçə şəkildə ümumiləşdirilə bilər. Birincisi, qonşuluq təyini simmetrik və yenidən mövcud olan predikat $NPred(p,q)$ əsasında qurulduğu müddətcə məsafəyə əsaslanan Eps-qonşuluğu əvəzinə hər hansı bir qonşuluq anlayışı işlədilə bilər. p N-nin qonşuluq nöqtəsidirsə, onda bütün q nöqtələr çoxluğu $NPred(p,q)$ kimi təyin olunur. İkincisi, sadəcə bir qonşuluqdakı elementləri saymaq əvəzinə, N qonşuluğunda sıx olub-olmadığını müəyyən etmək üçün, əgər $MinWeight$ N-də monotondursa, ümumi $MinWeight(N)$ predikatından istifadə edə bilərik. Nəhayət, yalnız nöqtə şəklində olan cisimlər deyil, fəzada yayılan çoxölçülü (polygons) cisimlər də klasterlənə bilir. Ümumiləşdirilmiş sıxlığa əsaslanan klasterləri tapmaq üçün GDBSCAN alqoritmi DBSCAN alqoritminin sadə modifikasiyasıdır (Kaufman, & Rousseeuw, 1990).

DBSCAN alqoritminin mərhələləri: DBSCAN alqoritminin işləmə prosesini aşağıda kimi şərh edə bilərik:

Alqoritm ixtiyari bir nöqtədən başlayır və qonşuluq məlumatları $\varepsilon(Eps)$ parametrindən alınır. Bu nöqtə $MinPts$ parametrinin ε qonşuluğunda yerləşirsə, klaster əmələ gətirir. Əks halda nöqtə küy nöqtə kimi işarələnir. Bu nöqtə sonradan fərqli bir nöqtənin ε qonşuluğunda yerləşə bilər və bununla da klasterin bir hissəsi ola bilər. Burada sıxlıq əldə edilə bilən və sıxlığa bağlı nöqtələr anlayışı vacibdir. Bir nöqtənin əsas nöqtə olduğu aşkar edilərsə, ε qonşuluğundakı nöqtələr də klaster hesab oluna bilər. Beləliklə, ε qonşuluğu içərisində tapılan bütün nöqtələr, əsas nöqtələrdirsə, onda bu nöqtələr öz qonşuluqda yerləşən nöqtələr ilə birlikdə əlavə olunurlar. Yuxarıdakı proses sıxlığa bağlı klaster tamamilə tapılana qədər davam edir. Proses yeni bir klasterin bir hissəsi ola bilən və ya küy nöqtələr

kimi işarələnən yeni bir nöqtə ilə yenidən başlayır.

5. Eksperimentlər

DBSCAN alqoritminin analizi üçün ilkin olaraq aşağıdakı şəkildə göstərildiyi kimi giriş parametrləri daxil edilir.

a) Verilənlər çoxluğunun daxil edilməsi

Burada fərqli verilənlər bazasından istifadə edilmişdir. Verilənlər bazasının əksəriyyəti siniflərdirilmiş atributlara aiddir. Onu da qeyd etmək lazımdır ki, verilənlər bazasının əksəriyyəti UCI Machine Learning Repository və Kaggle saytından əldə edilmişdir (https://archive.ics.uci.edu/ml/datasets.php, https://www.kaggle.com/).

b) İstifadə olunan vasitə (Python)

Bazanı oxuduqdan sonra növbəti addım bütün bu verilənlər bazasını işlətmək üçün istifadə olunan Python proqramlaşdırma dilidir. (Python – şərh edilən, obyektyönümlü, dinamik semantikaya sahib yüksək səviyyəli proqramlaşdırma dilidir.)

c) DBSCAN alqoritminin tətbiqi

Bu alqoritmin vacib mərhələlərindən biri Python proqramlaşdırma dilində bütün verilənlər bazasını daxil etdikdən sonra DBSCAN alqoritminin həmin verilənlər bazasına tətbiq edilməsidir.

d) Parametrlərin hesablanması

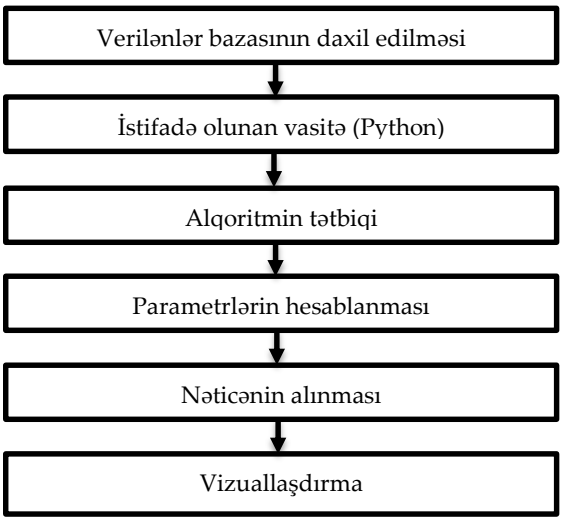
DBSCAN alqoritmini verilənlər bazasına tətbiq etməzdən əvvəl istifadə olunan parametrin standart qiymətindən nisbətən fərqli nəticə təmin etmək üçün parametrlərin qiyməti konfiqurasiya edilməlidir. Parametrlərin qiymətlərinin düzgün seçilməsi DBSCAN alqoritmin vacib məsələsidir.

e) Nəticənin alınması

Bu mərhələdə alqoritmin effektivliyi qiymətləndirilir və Eps və $MinPts$ parametrlərinin köməyi ilə qurulmuş klasterlər, səhv klasterləşdirilmiş hallar, vaxt ölçmə və küy nöqtələri analiz edilir.

f) Vizuallaşdırma

Nəticə alındıqdan sonra müxtəlif verilənlər bazası qrafik təsviri ilə vizuallaşdırılır.



Şəkil 2. Tətbiq olunan alqoritmin strukturu

Təkmilləşdirilmiş DBSCAN alqoritminin effektivliyini qiymətləndirmək üçün müxtəlif növ verilənlər bazasından istifadə edirik. Bütün verilənlər bazasına DBSCAN alqoritmi tətbiq olunur və Python proqramlaşdırma dilində işlənir. İstifadə olunmuş verilənlər bazalarının təsviri cədvəl 1-də göstərildiyi kimidir.

Cədvəl 1. Verilənlər bazasının xarakteristikaları

Adı	Nöqtələrin sayı	Atributların sayı	Tətbiq sahəsi
Mall Customer	200	5	Biznes
Wholesela	440	36	Biznes
Loan	614	16	Sosial
Live	7050	16	Biznes
Churn	10000	14	Sosial
Online Shoppers	12330	18	Biznes
Adult	48842	15	Siyahıyaalma Bürosu
Bank – marketing	45211	17	Maliyyə
Diabetic	101766	50	Tibb

Qeyd etdiyimiz kimi DBSCAN alqoritminin "keyfiyyətli" nəticə verməsi üçün Eps və $MinPts$ parametrləri düzgün seçilməlidir. Əvvəlcə verilənlər bazasındakı atributlar arasında korrelyasiya əmsalı hesablanır. İki ən güclü korrelyasiyaya sahib atributlar arasında qrafik qurulur. Eps parametrinin optimal qiymətini təyin etmək üçün $k-NN$ ($k-$ nearest neighbors) metodundan istifadə edilir. $k-NN$ şablonların tanınmasında $k-$ ən yaxın qonşuluq alqoritmi

klassifikasiya, reqressiya və klasterləşdirmə üçün istifadə olunan qeyri-parametrik bir metoddur (şəkil.3). $MinPts$ parametri isə susmaya (default) görə 5 qəbul edilir və sonra bu qiymət ətrafındakı bütün nöqtələr yoxlanılaraq eksperimentin nəticələrinin analizini reallaşdırmaq məqsədilə bir neçə qiymətləndirmə indekslərindən istifadə olunur.

Eksperimentin nəticələrinin qiymətləndirilməsi üçün Silhouette əmsalı (Silhouette score), tənzimlənən Rand indeksi (the adjusted Rand index), Davies-Bouldin indeksi (3), Purity indeksi (4) və Homogenlik indeksi istifadə edilmişdir. Eyni zamanda tədqiqatda $k-NN$ vasitəsilə optimal Eps qiyməti tapılsa da yaxşı klaster tapmaq üçün təyin olunan Eps parametri ətrafında da bütün nöqtələr yoxlanılır.

Qeyd edək ki, silhouette əmsalı klasterdaxili nöqtələr və ən yaxın klasterlərarası məsafəni hesablayır. Məsələn, bir-birinə yaxın çoxlu məlumat nöqtələrinə (yüksək sıxlıqlı) malik klaster bir-birindən uzaq məsafədə yerləşən klasterlərə görə daha yüksək silhouette əmsalına malik olacaqdır. Silhouette əmsalı -1 ilə 1 intervalında dəyişir. -1 mümkün olan ən pis qiymət, 1 isə ən yaxşı maksimum qiymət hesab olunur. Silhouette əmsalı 0-a bərabər olduqda, üst-üstə düşən klasterlər təklif edilir.

Davies-Bouldin indeksi hər klasterin ən oxşar klasterlə ortalama ölçüsü kimi müəyyən edilir. Burada oxşarlıq klaster daxilindəki məsafələrin klasterlərarası məsafələrə nisbətidir. Beləliklə, daha uzaq məsafələrdə yerləşmiş klasterlərdə daha yaxşı nəticə əldə ediləcəkdir. Bu indeksin minimum qiyməti sıfırdır və indekslərin aşağı qiymətləri klasterləşdirmə üçün daha yaxşı hesab edilir.

Davies-Bouldin indeksi aşağıdakı kimi hesablanır:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \underbrace{max}_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right), \qquad (1)$$

burada $n-$ klasterlərin sayı, c_i-i-ci klasterinin mərkəzi, σ_i- mərkəzi c_i olan i klasterinin bütün elementlərinin orta məsafəsi və $d(c_i,c_j)$ c_i və c_j mərkəzləri (centroid) arasındakı məsafədir.

Purity indeksinin hesablanması üçün (2) düsturundan istifadə edilir:

$$Purity = \frac{1}{n}\sum_{i=1}^{k} max_j\left|C_i \cap A_j\right|, \qquad (2)$$

burada $n-$ obyektlərin sayı, $k-$ klasterlərin sayı, C_i-i-ci klaster, A_j-j-ci sinifdir. Purity indeksinin qiyməti nə qədər böyük olarsa, alqoritm bir o qədər effektiv hesab edilir.

$S=\{O_1,O_2,\dots,O_n\}$ şəkildə n obyektlər çoxluğunu nəzərdən keçirək. Fərz edək ki, $U=\{u_1,u_2,\dots,u_R\}$ və $V=\{v_1,v_2,\dots,v_C\}$ S çoxluğunun iki müxtəlif altçoxluqlarıdır. Onda, $1\le i\neq i'\le R$ və $1\le j\neq j'\le C$ üçün $\cup_{i=1}^{R} u_i = S = \cup_{j=1}^{C} v_j$; $u_i\cap u_i=\emptyset=v_j\cap v_j$ kimidir. Tutaq ki, n_{ij}, u_i və v_j siniflərində olan obyektlərin ümumi sayını göstərir. Rand İndeksi (Hubert, & Arabie, 1985;), U və V arasında uyğunluq ölçülərininin obyekt cütlərinin $R\times C$ ölçülü təsadüfi verilənlər cədvəlində necə təsnif edildiyinə əsaslanır.

Cədvəl 2. Eksperimentlərin nəticələri

Verilən lər bazası	Eps	Min Pts	ARI	Silhouette score	Purity	Homo-Geneity	Davies-Bouldin index
Mall Customer (200 x 5)	0.04	5	0.046	0.046	---	---	4.54
	0.05	2	0.049	0.146	---	---	2.90
	0.06	5	0.048	0.181	---	---	3.19
	0.04	4	0.052	0.062	---	---	2.95
	0.04	**3**	**0.047**	**0.088**	**---**	**---**	**1.85**
Wholesela (440 x 36)	0.04	5	0.022	0.441	---	---	0.77
	0.05	**5**	**0.002**	**0.745**	**---**	**---**	**0.72**
	0.03	5	0.003	0.670	---	---	0.87
	0.04	4	0.019	0.724	---	---	0.78
	0.04	6	0.006	0.702	---	---	0.85
Loan (614 x 16)	**0.07**	**5**	**0.001**	**0.699**	**0.68**	**0.84**	**1.47**
	0.03	5	0.001	0.554	0.68	0.74	1.47
	0.05	4	-0.002	0.473	0.68	0.78	0.70
	0.05	5	-0.003	0.452	0.68	0.84	0.72
	0.05	6	-0.001	0.725	0.68	0.68	1.00
Live (7050 x 16)	0.02	5	0.021	0.719	0.62	0.90	2.36
	0.02	9	0.022	0.720	0.62	0.80	1.96
	0.02	7	0.021	0.719	0.62	0.80	1.78
	0.03	**7**	**0.021**	**0.671**	**0.62**	**0.92**	**1.74**
	0.01	7	0.036	0.697	0.62	0.72	0.93

Churn (10000 x14)	0.01	10	0.019	-0.238	0.79	0.74	2.34
	0.02	10	-0.007	-0.311	0.79	0.86	5.75
	0.03	10	-0.013	0.162	0.79	0.76	28.50
	0.02	8	-0.006	-0.268	0.79	0.82	7.68
	0.02	12	-0.005	-0.278	0.79	0.81	6.75
Onlin Shoppers (12330 x 18)	0.01	10	-0.068	0.578	0.85	0.961	1.043
	0.01	14	-0.073	0.705	0.85	0.969	1.058
	0.01	8	-0.068	0.583	0.85	0.965	1.412
	0.02	**10**	**-0.056**	**0.787**	**0.85**	**0.967**	**1.045**
	0.009	10	-0.072	0.574	0.85	0.964	1.099
Adult (48842 x15)	0.02	10	0.0	0.631	1.0	1.0	0.988
	0.03	**10**	**0.0**	**0.607**	**1.0**	**1.0**	**0.804**
	0.01	10	0.0	0.608	1.0	1.0	1.336
	0.01	7	0.0	0.580	1.0	1.0	1.411
	0.01	12	0.0	0.624	1.0	1.0	1.434
Bank marketing (45211 x 17)	0.02	12	0.011	0.84	0.88	0.88	1.05
	0.02	10	0.009	0.84	0.88	0.88	1.06
	0.03	10	0.005	0.83	0.88	0.94	1.20
	0.01	**10**	**0.025**	**0.70**	**0.88**	**0.94**	**1.07**
	0.02	8	0.008	0.70	0.88	0.81	1.90
Diabetic (101766 x 50	0.02	12	0.0	0.10	1.0	1.0	1.63
	0.03	**12**	**0.0**	**0.31**	**1.0**	**1.0**	**0.99**
	0.02	14	0.0	0.01	1.0	1.0	1.88
	0.01	12	0.0	-0.50	1.0	1.0	1.58
	0.02	10	0.0	0.02	1.0	1.0	1.05

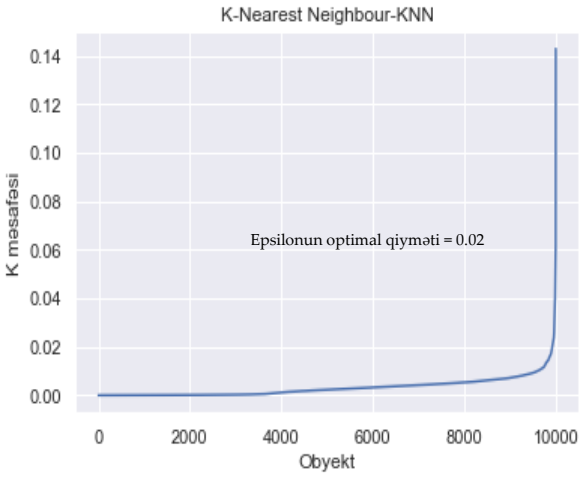
Xüsusilə də, $\binom{n}{2}$ cütləri arasında təyin edilən dörd fərqli növü vardır:

1. cütdəki obyektlər U-da və V-də eyni sinifdə yerləşdirilir;
2. cütdəki obyektlər U-da və V-də fərqli siniflərdə yerləşdirilir;
3. cütdəki obyektlər U-da fərqli siniflərdə və V-də eyni sinifdə yerləşdirilir;
4. cütdəki obyektlər U-da eyni sinifdə və V-də fərqli siniflərdə yerləşdirilir.

Birinci və ikinci tiplər səciyyəvi olaraq bir cütdəki obyektlər arasında uyğunluğu, üçüncü və dördüncü tiplər isə bir cütdəki obyektlər arsındakı uyğunsuzluğu ifadə edir. Aydındır ki, əgər A uyğunluqların ümumi sayını, D isə uyğunsuzluqların ümumi sayını göstərirsə, onda $A + D = \binom{n}{2}$. Beləliklə,

$$A = \binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right) = \binom{n}{2} + 2\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n^{ij}}{2} - \left(\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right). \quad (3)$$

Bütün bu qeyd edilənlər müxtəlif verilənlər bazasına tətbiq olunduqdan sonra məlumatların və parametrlərin ətraflı təsviri aşağıdakı cədvəl 2-də göstərilmişdir. Aşağıdakı eksperimentlərdə fərqli verilənlər bazası üçün ilkin klasterləşdirmə, $k-NN$ metodu və DBSCAN alqoritminin tətbiqi ilə yekun qiymətləndirmə təsvir olunmuşdur.



Şəkil 3. k-NN vasitəsilə epsilonun optimal qiymətinin tapılması

Şəkilə əsasən qeyd edə bilərik ki, k-NN metodunun tətbiqi ilə Eps parametrinin optimal qiyməti tapılır. MinPts parametrinin qiyməti isə verilənlər bazası kiçik ölçülü olduqda 6-dan aşağı, böyük ölçülü olduqda isə 7 və daha yüksək qiymətlər olaraq götürülür. Tapılan parametrlərin qiymətlərinə əsasən metrikalar vasitəsilə qiymətləndirmə aparılır. Daha effektiv klasterin əldə edilməsi üçün bu parametrlərin qiymətləri təyin etdiyimiz qiymətləri bir neçə vahid artırıb və ya azaltmaqla dəyişdirilir. Qeyd etmək lazımdır ki, qiymətləndirmə üçün istifadə edilmiş Davies-

Bouldin indeksinin minimal qiyməti, Bircinslik (Homogeneity) və Purity indeksinin isə maksimal qiyməti daha yaxşı nəticəni ifadə edir. Cədvəl 2-də sinifləndirilmiş verilənlər üçün əsasən Homogeneity indeksinə görə, sinifləndirilməmiş (ilk iki verilənlər bazası) verilənlərdə isə Davies-Bouldin indeksinə görə klasterlər yaradılır. Əgər sinifləndirilmiş verilənlərdə Homogeneity və ya Purity indeksi bütün hallarda eynidirsə, onda burada da Davies-Bouldin indeksi nəzərə alınaraq, klaster yaradılır.

Nəticə

Tədqiqat işində böyük ölçülü verilənlər konsepsiyası araşdırılmış, böyük ölçülü verilənlərin klasterləşdirilməsi üçün klasterləşdirmə alqoritmlərinin nəzəri cəhətdən müqayisəli analizi aparılmışdır. Böyük ölçülü verilənlərin analizi və küy nöqtələrin dəqiq aşkar edilməsi üçün sıxlığa əsaslanan DBSCAN alqoritmi tətbiq edilmiş və böyük ölçülü verilənlər üçün praktiki əhəmiyyəti müəyyən edilmişdir. DBSCAN alqoritminin qiymətləndirilməsində səmərəliliyin artırılması üçün müxtəlif metrikalardan (Silhouette score, Adjusted Rand index, Purity index və Homogeneity index və s.) istifadə olunmuşdur. Eksperimentlərin nəticələrinə əsasən DBSCAN alqoritmi müxtəlif indekslərə görə yüksək nəticə göstərmişdir. Qeyd etdiyimiz kimi alqoritm əsas iki parametrin ($Eps, MinPts$) seçilməsində həssas olduğuna baxmayaraq tədqiqat nəticəsində keyfiyyətli klasterlər əldə edilmişdir. Nəticədə təklif edilən metod ənənəvi DBSCAN alqoritminə nisbətən küy nöqtələri aşkar etməkdə daha effektiv nəticə göstərmişdir.

Ədəbiyyat

Alguliyev, R. M., Aliguliyev, R. M., & Sukhostat, L. V. (2019). Efficient algorithm for big data clustering on single machine. CAAI Transactions on Intelligence Technology, *5*(1), 9-14. https://doi.org/10.1049/trit.2019.0048

Alguliyev R.M., Aliguliyev R.M., Abdullayeva F.J. (2019). Privacy-preserving deep learning algorithm for big personal data analysis. Journal of Industrial Information Integration, 15, 1-14. https://doi.org/10.1016/j.jii.2019.07.002

Alguliyev, R., Aliguliyev, R., & Sukhostat, L. (2017). Anomaly detection in Big data based on clustering. *Statistics, Optimization & Information Computing*, *5*(4), 325-340. https://doi.org/10.19139/soic.v5i4.365

Alguliyev, R., & Imamverdiyev, Y. (2014). Big data: Big promises for information security. In *2014 IEEE* 8th International Conference on Application of Information and Communication Technologies (AICT) (pp. 1-4). IEEE. 10.1109/ICAICT.2014.7035946

Alguliyev R., & Hajirahimova M. (2014). "BIG DATA" PHENOMENON: CHALLENGES AND OPPORTUNITIES. Problems of information technology, *5*(2), 3-16. https://jpit.az/en/journals/120

Aliguliyev R., Hajirahimova M., & Aliyeva A. (2016). Current scientific and theoretical problems of Big data. Problems of information society, (2), 37-49 (Əliquliyev, R. M., Hacırəhimova, M. Ş., & Əliyeva, A. S. (2016). Big data-nin aktual elmi-nəzəri problemləri. İnformasiya cəmiyyəti problemləri, (2), 37-49). https://jpis.az/az/journals/138

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning*, *50*(1), 5-43. https://doi.org/10.1023/A:1020281327116

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-28349-8_2

Big data (2008). http://www.nature.com/news/specials/bigdata/index.html

Cassisi, C., Ferro, A., Giugno, R., Pigola, G., & Pulvirenti, A. (2013). Enhancing density-based clustering: Parameter reduction and outlier detection. *Information Systems*, *38*(3), 317-330. https://doi.org/10.1016/j.is.2012.09.001

Chana I.A., & Arora S. (2014). Survey of clustering techniques for big data analysis. 5th International Conference - Confluence the Next Generation Information Technology Summit, 59-65. 10.3233/JIFS-202503

Chandra, E., & Anuradha, V. P. (2011). A survey on clustering algorithms for data in spatial database management systems. International Journal of Computer Applications, *24*(9), 19-26.

Dharni, C., & Bnasal, M. (2013). An improvement of DBSCAN Algorithm to analyze cluster for large datasets. In 2013 IEEE international conference in MOOC, innovation and technology in education (MITE) (pp. 42-46). IEEE. 10.1109/MITE.2013.6756302

Duan, L., Xu, L., Guo, F., Lee, J., & Yan, B. (2007). A local-density based spatial clustering algorithm with noise. Information systems, 32(7), 978-986. https://doi.org/10.1016/j.is.2006.10.006

El-Sonbaty, Y., Ismail, M. A., & Farouk, M. (2004). An efficient density based clustering algorithm for large databases. In 16th IEEE international conference on tools with artificial intelligence (pp. 673-677). IEEE. 10.1109/ICTAI.2004.27

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, *2*(3), 267-279. 10.1109/TETC.2014.2330519

Fakhraddingizi A. (2019). Fundamental issues of data security in big data technologies. Actual multidisciplinary scientific-practical problems of information security, V republic conference, 226-228. (Fəxrəddinqızı A. (2019). Big data texnologiyalarında verilənlərin təhlükəsizliyinin əsas məsələləri. İnformasiya təhlükəsizliyinin aktual multidissiplinar elmi-praktiki problemləri V respublika konfransı, 226-228).

Gaonkar, M. N., & Sawant, K. (2013). AutoEpsDBSCAN: DBSCAN with Eps automatic for large dataset. International Journal on Advanced Computer Theory and Engineering, 2(2), 11-16.
https://archive.ics.uci.edu/ml/datasets.php
https://www.kaggle.com/
https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of classification, 2(1), 193-218.
https://doi.org/10.1007/BF01908075

Karypis, G., Han, E. H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer, 32(8), 68-75.
10.1109/2.781637

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.
https://books.google.az/books

Kailing, K., Kriegel, H. P., & Kröger, P. (2004, April). Density-connected subspace clustering for high-dimensional data. In Proceedings of the 2004 SIAM international conference on data mining (pp. 246-256). Society for Industrial and Applied Mathematics.
10.1137/1.9781611972740.23

Liu, P., Zhou, D., & Wu, N. (2007). VDBSCAN: varied density based spatial clustering of applications with noise. In 2007 International conference on service systems and service management (pp. 1-4). IEEE.
10.1109/ICSSSM.2007.4280175

Moreira, A., Santos, M. Y., & Carneiro, S. (2005). Density-based clustering algorithms–DBSCAN and SNN. University of Minho-Portugal, 1-18.
http://get.dsi.uminho.pt/local/download/SNN&DBSCAN.pdf

Parimala, M., Lopez, D., & Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. International Journal of Advanced Science and Technology, 31(1), 59-66.
10.1.1.643.6121

Rahmah, N., & Sitanggang, I. S. (2016). Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra. In IOP conference series: earth and environmental science (Vol. 31, No. 1, p. 012012). IOP Publishing.
10.1088/1755-1315/31/1/012012

Sajana, T., Rani, C. S., & Narayana, K. V. (2016). A survey on clustering techniques for big data mining. Indian journal of Science and Technology, 9(3), 1-12.
10.17485/ijst/2016/v9i3/75971

Shah, G. H. (2012). An improved DBSCAN, a density based clustering algorithm with parameter selection for high dimensional data sets. In 2012 Nirma university international conference on engineering (NUiCONE) (pp. 1-6). IEEE.
10.1109/NUICONE.2012.6493211

Sharma, S., Sharma, A. K., & Soni, D. (2017). Enhancing DBSCAN algorithm for data mining. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1634-1638). IEEE.
10.1109/ICECDS.2017.8389724

Uncu, O., Gruver, W. A., Kotak, D. B., Sabaz, D., Alibhai, Z., & Ng, C. (2006, October). GRIDBSCAN: GRId density-based spatial clustering of applications with noise. In 2006 IEEE International Conference on Systems, Man and Cybernetics (Vol. 4, pp. 2976-2981). IEEE.
10.1109/ICSMC.2006.384634

Xiong, Z., Chen, R., Zhang, Y., & Zhang, X. (2012). Multi-density dbscan algorithm based on density levels partitioning. Journal of Information and Computational Science, 9(10), 2739-2749.

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, 350-361.
https://doi.org/10.1016/j.neucom.2017.01.026

Zhao, W., Ma, H., & He, Q. (2009). Parallel k-means clustering based on mapreduce. In IEEE international conference on cloud computing (pp. 674-679). Springer, Berlin, Heidelberg.
https://doi.org/10.1007/978-3-642-10665-1_71.

Aygül F. Fəxrəddinqızı

AMEA İnformasiya Texnologiyaları İnstitutu.
Azərbaycan, Bakı ş., AZ1141, B.Vahabzadə küç., 9A.

Айгюль Ф. Фахраддингизи

Институт Информационных Технологий НАН Азербайджана.
Азербайджан, г. Баку, AZ1141, ул. Б.Вахабзаде, 9A.