**UOT 004.056** DOI: 10.25045/jpit.v12.i2.11

***Əhmədov E.Y.***
AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
eltunehmedov95@gmail.com

# PYTHON MÜHİTİNDƏ K-MEANS, K-MEANS++ VƏ MİNİ BATCH K-MEANS ALQORİTMLƏRİNİN MÜQAYİSƏLİ ANALİZİ

Daxil olmuşdur: 08.05.2021 Düzəliş olunmuşdur: 17.05.2021 Qəbul olunmuşdur: 24.05.2021

*Məqalədə k-means alqortitmi və onun modifikasiyalarının Python mühitində müxtəlif ölçülü verilənlərə tətbiqi məsələlərinə baxılır. Eyni zamanda ənənəvi k-means klasterləşdirmə alqoritmi və onun modifikasıyalarının mövcud vəziyyəti, imkanları, çatışmazlıqları, meydana çıxan problemlər tədqiq edilmiş və onların həlli üçün təkliflər verilmişdir. k-means++ alqoritmi vasitəsilə ənənəvi k-means metodunun başlanğıc mərkəzlərin təsadüfi seçilməsi çatışmazlığı aradan qaldırılmışdır. Mini batch k-means alqoritmi vasitəsilə böyük verilənlər paketlərə bölünməklə analiz edilmişdir ki, bu da böyük həcmli, kompleks verilənlərin analiz prosesini sürətləndirmişdir. Verilənlərin klasterləşdirilməsi zamanı ölçünün azaldılması və klasterlərin optimal sayının tapılması üçün hibrid PCA və elbow metodu təklif edilmişdir. Bu yanaşmanın effektivliyinin qiymətləndirilməsi üçün alqoritmlər müxtəlif ölçülü bir neçə verilənlər çoxluğu üzərində sınaqdan keçirilmişdir. Alqoritmlərin effektivliyinin qiymətləndirilməsi üçün siluet və Devis-Boldin indekslərindən istifadə edilmişdir. Eksperimentin nəticəsi göstərmişdir ki, təklif olunan yanaşma böyük ölçülü verilənlərin klasterləşdirilməsində daha effektivdir. Təklif edilən hibrid PCA və elbow metodu irihəcmli, çoxölçülü verilənlərin emal prosesində böyük hesablama resursları tələb edən məsələlərin həllinə yeni imkanlar yaradır.*

***Açar sözlər:*** *data mining, klasterləşdirmə, k-means, k-means++, mini batch k-means, elbow, PCA.*

## Giriş

İnsan varlığının yarandığı ilk gündən bəşəriyyətin tərəqqisi üçün əsas amil informasiya idi. Vəziyyət bu gün də eynidir, lakin çox böyük informasiya ilə qarşı-qarşıya olduğumuz üçün verilənlərin analizində bir sıra çətinliklərin yaranması qaçılmaz olmuşdur. Əsl problem bu məlumat qarışıqlığından nəyin dəyərli və əhəmiyyətli olduğunu çıxarmaqdır. Bu kimi çətinliklər yeni texnoloji həllərin meydana çıxmasına şərait yaratmışdır. Aydındır ki, informasiya-kommunikasiya texnologiyaları sahəsinin sürətli inkişafı ilə hər bir sahədə strukturlaşdırılmamış böyük həcmli verilənlər (*big data*) istehsal olunur. Bu xam verilənlər müxtəlif metodlar ilə analiz edilir. Hal-hazırda *data mining* verilənlərin analizində kompüter elmləri sahəsində ən çox tələbat olunan texnologiyalardan biridir. Son illər verilənlərin həcminin sürətlə artması ilə bu texnologiyaya tələbat daha da artmışdır.

Verilənlər çoxluğu həcmindən, tipindən və istifadə sahəsindən asılı olaraq müxtəlif metod və alqoritmlər vasitəsilə analiz olunur. *Data mining* sahəsində klasterləşdirmə metodları ən çox istifadə olunan və effektiv texnologiyalardan hesab olunur. Klasterləşdirmə metodları arasında sadəliyinə və effektivliyinə görə geniş istifadə olunan metodlardan biri *k-means* alqoritmidir. Bu məqsədlə məqalədə müxtəlif həcmdə olan verilənlər üzərində *k-means* alqoritmi və onun modifikasiyalarından istifadə etməklə eksperimentlər aparılmışdır. Eksperimentin nəticələrinə əsasən, *k-means* alqoritmi və onun modifikasiyaları müxtəlif qiymətləndirmə indeksləri vasitəsilə müqayisə edilir.

## *Data Mining* konsepsiyası

*Data Mining* xam (emal olunmamış) böyük verilənlər (BV) çoxluğundan əvvəlcədən məlum olmayan, qeyri-trivial, faydalı informasiyanın aşkarlanması prosesidir [1]. Bu məlumatlar insan fəaliyyətinin müxtəlif sahələrində generasiya olunur. *Data mining* böyük həcmli verilənlər

çoxluğundan faydalı informasiya əldə olunması probleminin öhdəsindən gəlmək üçün maşın təlimi, obrazların tanınması, statistika, verilənlər bazası və vizuallaşdırma kimi sahələri özündə birləşdirən fənlərarası bir elm sahəsidir.

Məlumdur ki, ticarət və bank əməliyyatları, elmi məlumatlar, sosial şəbəkələr, şəkillər və videolar kimi müxtəlif tipli verilənlərin həcmi böyük sürətlə artmaqda davam edir. Beləliklə, mövcud məlumatların mahiyyətini çıxara biləcək və daha yaxşı qərar qəbul etmək üçün avtomatik olaraq hesabat, baxış və ya məlumatların xülasəsini yarada biləcək bir sistemə ehtiyacımız var. Verilənlər bazalarında biliklərin aşkarlanması (*ing. Knowledge Discovery in Databeses – KDD*) olaraq da bilinən *data mining* məlumat bazalarındakı gizli, əvvəllər məlum olmayan və potensial faydalı biliklərin çıxarılmasında istifadə olunur. *KDD* xam verilənlərdən faydalı biliklərin aşkarlanmasının ümumi prosesinə deyilir, *data mining* isə bu prosesin addımlarından biridir. Əsas *KDD* tətbiqetmə sahələrinə marketinq, saxtakarlığın aşkarlanması, telekommunikasiya və istehsal sahələri daxildir. *KDD* prosesinin addımları aşağıdakı kimidir:

1. Verilənlərin təmizlənməsi (*ing. data cleaning*) – bu mərhələdə küy məlumatları və əhəmiyyətsiz məlumatlar verilənlərdən çıxarılır.
2. Verilənlərin inteqrasiyası (*ing. data integration*) – bir çox mənbələrdən alınan məlumatların çox vaxt verilənlər anbarı (*ing. data warehousing*) adlanan mərkəzləşdirilmiş bazada yerləşdiyi bir prosesdir.
3. Verilənlərin seçilməsi (*ing. data selecting*) – işlənməsi lazım olan verilənlər dəsti müəyyənləşdirilir və seçilir.
4. Verilənlərin çevrilməsi (*ing. data transformation*) – verilənlərin transformasiyası məlumatların müəyyən bir formatdan digərinə çevrilməsidir.
5. Verilənlərin intellektual analizi (*ing. data mining*) – potensial faydalı nümunələri çıxarmaq üçün "ağıllı" texnologiyaların tətbiq olunduğu həlledici addımdır.
6. Model qiymətləndirmə (*ing. Pattern Evaluation*) – bu mərhələ aşkarlanan biliklərin istifadəçinin baxış bucağından maraqlılığını qiymətləndirmək üçün istifadə olunur.
7. Biliklərin təsviri (*ing. Knowledge Representation*) – biliklərin təsviri, vizuallaşdırma vasitələrindən istifadə etməklə *data mining* nəticələrini əks etdirən son mərhələ hesab olunur.

Müştərilərin əldə edilməsi və saxlanılması şirkətlərin uzunmüddətli iş qabiliyyəti və gəlirliliyi üçün çox vacibdir. Məsələn, *data mining* texnologiyası keçmiş müştəri davranışlarına əsaslanaraq hansı potensial müştərilərin şirkətin xidmətləri ilə maraqlandığını görməyə imkan verir. Bu texnologiya vasitəsilə müştəriləri müəyyən qruplara bölmək və hər bir qrupa uyğun təkliflər etməklə gəliri artırmaq, xərcləri isə azaltmaq mümkündür. *Data mining* alətləri gələcək tendensiyaları və davranışları proqnozlaşdırır, təşkilatların biliyə əsaslanan qərarlar qəbul etməsinə kömək edir.

### *Data Mining* texnologiyasının tətbiq sahələri

Hal-hazırda *data mining* ilk növbədə güclü istehlakçı mərkəzləri olan şirkətlər – pərakəndə satış, maliyyə, rabitə və marketinq qurumları tərəfindən əməliyyat məlumatlarını "araşdırmaq" və qiymətləri müəyyən etmək, müştərilərin üstünlüklərini aşkarlamaq və məhsulların yerləşdirilməsini təmin etmək, həmçinin satış, müştəri məmnuniyyəti və korporativ mənfəət üçün istifadə edilir. Bu texnologiya bir çox müəssisələrdə populyardır, çünki müştərilər haqqında daha çox məlumat əldə etməyə və ağıllı marketinq qərarları qəbul etməyə imkan verir. *Data Mining* texnologiyasının geniş istifadə olunduğu sahələr aşağıdakılardır:

1. Bazar səbəti analizi – müəyyən bir qrup məhsulu almaqla başqa bir qrup məhsulun alınma ehtimalı olduğuna dair bir nəzəriyyəyə əsaslanan modelləşdirmə üsuludur. Bu məlumatlar pərakəndə satıcıya müştərinin ehtiyaclarını bilməkdə və mağazanın tərtibatını buna görə dəyişdirməkdə kömək edə bilər. Müxtəlif mağazalar və ya fərqli

demoqrafik qruplardakı müştərilər arasında nəticələrin diferensial təhlilindən istifadə etməklə müqayisələr aparıla bilər.

2. Təhsildə *data mining* – təhsil mühitində olan mövcud məlumatlardan bilinməyən, faydalı bilikləri aşkar edən metodların inkişaf etdirilməsinə deyilir. Bu sistemin iş prinsipi tələbələrin gələcəkdə öyrənmə üsulunu proqnozlaşdırmaq, təhsil dəstəyinin təsirlərini öyrənmək və öyrənmə üsulları haqqında elmi təcrübələri inkişaf etdirmək kimi müəyyən edilir. *Data mining* bir təşkilat tərəfindən dəqiq qərarlar qəbul etmək və tələbənin nəticələrini proqnozlaşdırmaq üçün istifadə edilə bilər. Alınan nəticələr ilə təhsil müəssisəsi "nəyi öyrətmək?" və "necə öyrətmək?" barədə düşünə bilər.
3. Müdaxilələrin aşkarlanması – bir mənbənin bütövlüyünə və məxfiliyinə xələl gətirəcək hər hansı bir hərəkət müdaxilə adlanır. Müdaxilənin qarşısını almaq üçün müdafiə tədbirlərinə istifadəçi identifikasiyası, proqramlaşdırma səhvlərindən yayınmaq və məxfiliyin qorunması daxildir. *Data mining* anomaliyanın aşkarlanmasına fokus səviyyəsini əlavə edərək müdaxilənin aşkarlanmasını yaxşılaşdırmağa kömək edə bilər.
4. Yalanların aşkarlanması – cinayətkarı tutmaq asandır, lakin həqiqəti ondan almaq çətindir. *Data mining* texnologiyası hüquq-mühafizə orqanları tərəfindən cinayətləri araşdırmaq, şübhəli terrorçuların ünsiyyətini izləmək üçün istifadə edilir. Qeyd edilən proses strukturlaşdırılmamış mətndən ibarət verilənlərdə mənalı əlaqələri tapmağa çalışır. Bu texnologiyaya mətnlərin analiz olunması (*ing. text mining*) deyilir.
5. Maliyyə və bankçılıq sahəsi – kompüterləşdirilmiş bankçılıqda yeni əməliyyatlarla birgə böyük ölçüdə məlumatların yaranması nəzərdə tutulur. *Data mining*, məlumat həcmi çox böyük olduğundan, menecerlər üçün görünməyən iş məlumatları və bazar qiymətlərindəki nümunələri, səbəbləri və əlaqələri taparaq, bank və maliyyə sahəsindəki iş problemlərinin həllinə kömək edə bilər.
6. Terrorizmlə mübarizə – *data mining* bir çox daxili təhlükəsizlik təşəbbüsünün əsas xüsusiyyətinə çevrilmişdir. Mürəkkəb riyazi alqoritmlər terrorizmlə mübarizə fəaliyyətində hansı kəşfiyyat qurumunun əsas rol oynamalı olduğunu göstərə bilər. *Data mining*, həmçinin işçi qüvvəsinin yerləşdiriləcəyi yerin təyin edilməsinə və sərhəd-keçid məntəqələrində axtarış yerlərinin müəyyən edilməsinə kömək edə bilər.
7. Enerji sənayesi – *big data* günümüzdə enerji sektorunda da mövcuddur ki, bu da müvafiq *data mining* üsullarına ehtiyac olduğunu göstərir. *Data mining* texnologiyası vasitəsilə elektrik enerjisi çıxışlarını və elektrik enerjisinin təmizlənmə qiymətini proqnozlaşdırmaq və bununla məhsuldar qazanc əldə etmək mümkündür.

**BV-də klasterləşdirmə alqoritmləri**

Son illər İKT sahəsinin sürətli inkişafı ilə bir çox sahələrdə böyuk həcmli verilənlər generasiya edilir. Bu verilənlər insan fəaliyyətinin bütün sahələrində (biznes, tibb, sosial şəbəkə, bank, sığorta, neft və qaz sənayesi və s.) istehsal olunur [2] və onların həcmi çox böyük sürətlə artır. Verilənlərin həcminin eksponensial qanunla artımı yeni terminin – BV-nin meydana gəlməsinə səbəb olmuşdur. Elmi ədəbiyyatda ən çox istinad olunan tərif Chen və Zhang tərəfindən verilmişdir: "Mövcud texnologiyaların, metodların və alqoritmlərin köməyilə toplanılması, saxlanılması, analizi və vizuallaşdırılması çətin və ya mümkün olmayan verilənlər çoxluğuna big data deyilir" [3].

BV-ninin əsas problemlərini özündə əks etdirən bir sıra xarakteristikaları mövcuddur: həcm (*ing. Volume*), sürət (*ing. Velocity*), müxtəliflik (*ing. Variety*), həqiqilik (*ing. Veracity*), dəyər (*ing. Value*). Buna, həmçinin «5*V*» də deyilir. Son illər BV-nin analizi zamanı ortaya çıxan yeni problemlərlə birlikdə "*V*"-lərin sayı artmağa davam edir. Son illərdə verilənlərin həcminin çox sürətli artması ilə onların saxlanması, işlənməsi və ötürülməsi ilə bağlı problemlər yaranır [4]. BV-nin xarakterik xüsusiyyətlərindən (böyük həcm, yüksək sürət, müxtəliflik) irəli gələn bir sıra problemlər (hesablama, əhatəlilik, saxlanma, düzgün olmayan korrelyasiyalar və s.) mövcuddur

ki, onların həllində yeni elmi baxışlar, yanaşmalar, modelləşmə, riyazi metodlar, optimallaşma üsulları və s. tələb olunur [5]. Klasterləşdirmə alqoritmləri BV-nin analizi üçün geniş istifadə olunan metodlardan biridir.

Klasterləşdirmə maşın təliminin nəzarətsiz öyrənmə metodu olub, bir-birinə mümkün dərəcədə ən çox oxşar verilənlərin bir klasterdə, oxşar olmayan verilənlərin isə digər klasterdə toplanması metodudur. Bu metodun əsas məqsədi eyni klasterdəki obyektlər arasındakı məsafələri minimallaşdırmaq və müxtəlif klasterlərdə olan obyektlər arasındakı məsafəni maksimallaşdırmaqdır. Qrupların əvvləcədən təyin olunmaması klasterləşdirmə metodunun əsas cəhətlərindən biridir. Klasterləşdirmə metodları 5 sinfə bölünür: bölünməyə əsaslanan (*ing. partition-based*), iyerarxik (*ing. hierarchical*), sıxlığa əsaslanan (*ing. density-based*), modelə əsaslanan (*ing. model-based*), şəbəkəyə əsaslanan (*ing. grid-based*) [1].

Klasterləşdirmə metodlarının tətbiq sahələri çox genişdir. Klasterləşdimə alqoritmləri müxtəlif xidmətlər göstərmək üçün müştərilərində ümumi cəhətləri aşkarlamaq istəyən şirkətlərdə tətbiq olunur. Beləliklə, müştərilərin əhəmiyyətli bir hissəsinin ümumi cəhətləri varsa, şirkət müəyyən bir kampaniya, xidmət və ya məhsulun istifadəsinə razılıq verə bilər.

Ümumiyyətlə, bir sıra tələblər var ki, klasterləşdirmə alqoritmləri bunları təmin etməlidir:

- Klasterləşmə alqoritmləri miqyaslana bilməlidir, yəni həcmi böyüyən verilənlərdə istifadəsi mümkün olmalıdır;
- Tətbiqetmə sahəsi fərqli ola biləcəyi üçün alqoritm fərqli atribut növləri ilə işləməyi bacarmalıdır;
- Küylü məlumatların və anomaliyaların aşkarlanması bacarığı;
- Alqoritm çoxölçülü verilənləri klasterləşdirmə bacarığına malik olmalıdır [6].

### *k-means* alqoritmi və onun modifikasiyaları

*k-means* alqoritmi bölünməyə əsaslanan klasterləşmə metodlarından biri olub, verilənlər dəstinin klasterlər arasında təkrarlanan yerdəyişməsinə əsaslanır. *k-means* metodu verilənlər çoxluğunun dəyişənlərini ortaya çıxan xüsusiyyətlərinə görə oxşar olmayan qruplara və ya klasterlərə bölmək üçün istifadə edilir. Metodun başlıca məqsədi qrup içində mümkün olan ən yüksək dərəcədə oxşarlığı, qruplar arasında mümkün olan ən aşağı dərəcədə oxşarlığı təmin etməkdir [7].

*k-means* alqoritmi strukturlaşmamış verilənlər dəstini giriş olaraq götürür, həmin verilənləri *k* sayda klasterlərə bölür və ən yaxşı qrupları tapana qədər prosesi təkrarlayır. *k*-nın qiyməti bu alqoritmdə əvvəlcədən təyin edilməlidir. *k-means* alqoritminin addımları aşağıdakı kimidir:

1. Klasterlərin sayı müəyyən edilir;
2. *k* sayda təsadüfi mərkəz nöqtələr seçilir;
3. Bütün nöqtələr çoxluğu ən yaxın klaster mərkəzinə təyin edilir;
4. Yeni yaradılmış qrupların mərkəzləri yenidən hesablanır;
5. Mərkəz nöqtələr eyni qalanadək 3-cü və 4-cü addımlar təkrarlanır.

*k-means* metodunun ən böyük üstünlüklərindən biri tətbiqinin asan olması və daha da əhəmiyyətlisi alqoritm kodunun bir çox proqramlaşdırma dillərində mövcud olmasıdır. *k-means* alqoritmi kifayət qədər sürətli işləyir və kiçik verilənlər dəsti üçün böyük hesablama resursu tələb etmir [8]. *k-means* alqoritminin əsas üstünlükləri aşağıda qeyd olunmuşdur:

- Başqa nümunələrə asanlıqla uyğunlaşa bilir;
- Riyazi düsturlarla hesablanması mümkündür;
- Elliptik qruplar kimi müxtəlif forma və ölçülərdə qruplarda ümumiləşdirir;
- Alqoritm BV-də belə, işini sürətli yerinə yetirir;
- Verilənlər çoxluğu aydın və qabarıq formadadırsa, çox yaxşı nəticə verir.

*k-means* alqoritminin ən böyük çatışmazlığı onun başlanğıc klaster mərkəzlərinin seçilməsindən asılı olmasıdır. Belə ki, ilkin mərkəzin seçilməsindən asılı olaraq, nəticələrin keyfiyyəti də müxtəlif ola bilər. Klasterlərin keyfiyyətinin ilkin mərkəzin seçilməsindən asılılığı,

böyük həcmli verilənlərin analizində çətinliklər, emal prosesində texniki təminata qoyulan yüksək tələblər kimi bir sıra məsələləri aradan qaldırmaq məqsədi ilə *k-means* alqoritminin müxtəlif modifikasiya edilmiş versiyalarına nəzər yetirilir [9].

Klaster sayının düzgün təyin olunmasının mürəkkəbliyi *k-means* alqoritminin digər çatışmazlıqlarından biridir. Verilənlərin analizi zamanı optimal sayda klasterlərin seçilməsi üçün bir neçə metod mövcuddur:

- **"*elbow*" (**dirsək) metodu;
- Siluet metodu;
- Klasterlərarası məsafə xəritəsi.

*k-means* alqoritmində hər bir obyektin hansı qrupa aid olduğunu bilmək üçün həmin obyekt ilə bütün mərkəzlər arasındakı məsafə hesablanılır. Hesablanan məsafələrə əsasən, hər bir verilənlər çoxluğu ən yaxın mərkəzə təyin edilir. Beləliklə, obyektlər və mərkəzlər arasındakı məsafələrin hesablanması klasterləşdirmə alqoritmində mühüm rol oynayır. Bildiyimiz kimi, iki obyekt arasındakı məsafənin hesablanması müxtəlif metrika üsulları vasitəsilə mümkündür. Əsas məqsəd mövcud metrikalardan verilənlər çoxluğuna uyğun olanını seçməkdir. Lakin bu cür metrikaları seçərkən verilənlərin xassəsi və ölçüsünə diqqət yetirmək vacib məqamlardır. *k-means* alqoritmində obyektlər arasındakı məsafələrin ölçülməsində Evklid metrikasından istifadə olunub. $A = (a_1, \dots, a_n)$ və $B = (b_1, \dots, b_n)$ nöqtələri arasındakı Evklid məsafəsi aşağıdakı şəkildə hesablanılır:

$$dist(A, B) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

*k-means* alqoritmindən bir çox sahədə geniş istifadə olunur, lakin onun tətbiqi ilə bağlı bir sıra problemlər movcuddur. *k-means* alqoritmi bütün verilənlərin yaddaşda saxlanmasını tələb edir. Bundan əlavə verilənlər çoxluğu və ya klasterlərin sayı çox olduqda, hesablama baxımından baha başa gələ bilər. Qeyd olunan çatışmazlıqları aradan qaldırmaq məqsədilə *k-means* alqoritminin modifikasiya olunmuş versiyaları olan *mini batch k-means* (kiçik paketli *k-means*) və *k-means*++ alqoritmlərindən istifadə edilir.

### A. *Mini batch k-means* alqoritmi

*Mini batch k-means* alqoritminin əsas ideyası verilənlərin yaddaşda saxlanıla bilməsi üçün sabit ölçülü kiçik təsadüfi paketlərdən istifadə etməkdir. Hər bir iterasiyada verilənlər dəstindən yeni təsadüfi paketlər əldə edilir və klasterləri yeniləmək üçün istifadə olunur. Bu proses həqiqi həllə yaxınlaşana (*ing. convergence*) qədər təkrarlanır. Hər bir kiçik paket nümunələrin qiymətlərinin qabarıq birləşməsindən istifadə edərək klasterləri yeniləyir.

### B. *k-means*++ alqoritmi

Ənənəvi *k-means* alqoritmi klaster mərkəzlərinin təsadüfi seçilməsi ilə başlayır, *k-means*++ alqoritmi isə bir klaster mərkəzinin təsadüfi seçilməsi ilə başlayır və daha sonra birinci klaster mərkəzinə əsasən digər mərkəzlər müəyyən edilir. Bu alqoritm klaster mərkəzlərinin daha düzgün seçilməsinə zəmanət verir və klasterləşmənin keyfiyyətini yaxşılaşdırır. *k-means*++ alqoritmi *k-means* metoduna nisbətən daha sürətli və effektiv hesab olunur.

### Eksperiment

*k-means*, *k-means*++, *mini batch k-means* alqoritmlərini təklif edilən yanaşma ilə qiymətləndirmək üçün Intel core (i7), 2.20 GHz, 8 Gb RAM xarakteristikalarına malik kompüterdə Python proqramlaşdırma dilindən (Python 3.8) istifadə edilmişdir. Eksperiment üçün bir neçə müxtəlifölçülü verilənlər üzərində ənənəvi *k-means* metodu və onun modifikasiyaları tətbiq edilmişdir. Eksperimentlər "Sales Transactions Dataset Weekly", "Facebook Live Sellers in Thailand", "Anuran Calls", "KEGG Metabolic Relation Network", "Gender by Name", "Query

Analytics Workloads Dataset" verilənləri üzərində tətbiq edilmişdir. İstifadə edilmiş verilənlərin xarakteristikaları cədvəl 1-də verilmişdir.

Cədvəl 1.

Verilənlər çoxluğunun xarakteristikası

| | **Verilənlər çoxluğu** | | | | | |
|---|---|---|---|---|---|---|
| | **Sales Transactions Dataset Weekly** | **Facebook Live Sellers in Thailand** | **Anuran Calls** | **KEGG Metabolic Relation Network** | **Gender by Name** | **Query Analytics Workloads Dataset** |
| Obyektlərin sayı | 811 | 7051 | 7195 | 53414 | 147270 | 260000 |
| Atributların sayı | 53 | 12 | 22 | 24 | 4 | 8 |

Eksperiment zamanı *k-means* və onun modifikasiyalarının daha effektiv nəticə verməsi üçün *PCA* (*ing. Principal Component Analysis*) və *elbow* ("dirsək") metodu təklif edilir. *PCA* verilənlərin ölçüsünün azaldılması, eyni zamanda məlumat itkisinin minimuma endirilməsi üçün bir metoddur. Bu metod vasitəsilə alqoritmlər daha sürətli işləyir və verilənlərin vizuallaşdırılması prosesi çox asan başa gəlir. *PCA* metodu vasitəsilə hər bir veriləndə atributların sayı 2-yədək azaldılmışdır. Klaster sayının təsadüfi seçilməsi və ya düzgün təyin olunmaması *k-means* alqoritminin keyfiyyətinə mənfi təsir göstərir. Eksperiment zamanı optimal klaster sayının təyin olunması üçün *elbow* metodu tətbiq olunmuşdur. *Elbow* metodunun iş prinsipi klasterləşmə alqoritminin məqsəd funksiyasının *k*-dan asılılıq qrafikinin qurulmasına əsaslanır. Qrafikin əyilmə nöqtəsinə *elbow* deyilir və həmin nöqtədəki *k*-nın qiyməti klasterlərin optimal sayı hesab olunur. Qrafiklərə əsasən, "*Facebook Live Sellers in Thailand*" və "*Gender by Name*" verilənlər çoxluğunun klaster sayının 4, digər verilənlərin klaster sayının isə 3 olduğu müəyyən edilmişdir. Eksperimentin nəticələrinin müqayisəli analizi siluet və Devis-Boldin (*ing. Davies-Bouldin*) qiymətləndirmə indeksləri, həmçinin alqoritmlərin yerinə yetirilmə vaxtına görə aparılmışdır.

A. Siluet indeksi. **İxtiyari $X_i \in C_p$ nöqtəsi üçün**

$$s(i) = \begin{cases} \dfrac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}, & \text{əgər } |C_p| > 1 \\ 0, & \text{əgər } |C_p| = 1 \end{cases}$$

burada $a(i)$ – $i$-ci verilən ilə aid olduğu klasterin daxilindəki digər verilənlər arasındakı orta məsafə, $b(i)$ – $i$-ci verilən ilə ən yaxın qonşu klasterin verilənləri arasındakı orta məsafədir:

$$a(i) = \frac{1}{|C_p|-1}\sum_{X_j \in C_p, j \neq i} dist(X_i, X_j),$$

$$b(i) = \min_{q=1,\dots,k; q \neq p} \frac{1}{|C_q|}\sum_{X_j \in C_q} dist(X_i, X_j).$$

Siluet indeksi $[-1,1]$ aralığında qiymət alır və onun -1 qiyməti ən pis, 1 qiyməti isə klasterləşdirmə üçün ən yaxşı hal hesab edilir.

B. Devis-Boldin (DB) indeksi

$$\text{DB} = \frac{1}{k}\sum_{p=1}^{k} R_p,$$

$$R_p = max_{q=1,\dots k, q \neq p}\, R_{pq}$$

$$R_{pq} = \frac{var(C_p) + var(C_q)}{dist(O_p, O_q)},$$

$$var(C_p) = \sqrt{\frac{1}{|C_p|}\sum_{i=1}^{|C_p|} dist(X_i, O_p)},$$

burada $k$ – klasterlərin sayı, $C_p - p$-ci klaster, $X_i - C_p$ klasterindəki nöqtə, $|C_p| - C_p$ klasterindəki nöqtələrin sayı, $O_p - C_p$ klasterinin mərkəzidir. DB indeksi 0 və 1 aralığında qiymətlər alır. Bu indeksin kiçik qiyməti klasterləşdirmə üçün yaxşı hesab edilir.

Cədvəl 2-də *m_k-means*, *m_k-means++* və *m_mini batch k-means* (*elbow* metodunun köməyilə klasterlərin optimal sayının müəyyən edilməsi və *PCA* metodunun tətbiqilə atributların sayının azaldılması) ilə uyğun olaraq *k-means*, *k-means++* və *mini batch k-means* alqoritmlərinin təklif olunmuş yanaşmalara nəzərən alınmış nəticələri göstərilmişdir.

Cədvəl 2.

*k-means*, *k-means++* , *mini batch k-means* alqoritmlərinin eksperimental nəticələri

| ***Sales Transactions Dataset Weekly*** | | | | ***KEGG Metabolic Relation Network*** | | | |
|---|---|---|---|---|---|---|---|
| Metod | Siluet | Devis-Boldin | Vaxt (san.) | Metod | Siluet | Devis-Boldin | Vaxt (san.) |
| ***m_k-means, k*=3** | **0.4647** | 0.7643 | **0.0850** | ***m_k-means, k*=3** | **0.8279** | 0.7263 | **0.7094** |
| *k-means, k*=4 | 0.4304 | 0.8279 | 0.3255 | *k-means, k*=4 | 0.8008 | **0.4440** | 1.5910 |
| *k-means, k*=5 | 0.4480 | 0.7753 | 0.1330 | *k-means,* k=5 | 0.7397 | 0.4887 | 2.0113 |
| k-means, *k*=6 | 0.4646 | 0.7535 | 0.1511 | *k-means,* k=6 | 0.7049 | 0.5468 | 2.7908 |
| k-means, *k*=7 | 0.4468 | **0.7517** | 0.1410 | *k-means,* k=7 | 0.6877 | 0.5302 | 4.3599 |
| ***m_k-means++, k*=3** | 0.4646 | 0.7650 | **0.0850** | ***m_k-means++, k*=3** | **0.8789** | **0.3793** | **0.4412** |
| *k-means++, k*=4 | 0.4322 | 0.8242 | 0.1360 | *k-means++, k*=4 | 0.8008 | 0.4440 | 0.8971 |
| *k-means++, k*=5 | 0.4480 | 0.7753 | 0.1541 | *k-means++, k*=5 | 0.7403 | 0.4889 | 1.1537 |
| *k-means++, k*=6 | 0.4617 | 0.7611 | 0.1941 | *k-means++, k*=6 | 0.7292 | 0.4936 | 1.2279 |
| *k-means++, k*=7 | **0.4766** | **0.6953** | 0.1651 | *k-means++, k*=7 | 0.6880 | 0.5303 | 1.4590 |
| ***m_mini batch k-means, k*=3** | **0.4651** | **0.7697** | **0.0410** | ***m_mini batch k-means, k*=3** | **0.8047** | **0.6745** | **0.1901** |
| *mini batch k-means, k*=4 | 0.4207 | 0.8327 | 0.0740 | *mini batch k-means, k*=4 | 0.6649 | 0.7234 | 0.3582 |
| *mini batch k-means, k*=5 | 0.4477 | 0.7733 | 0.0930 | *mini batch k-means, k*=5 | 0.5487 | 0.7283 | 0.3972 |
| *mini batch k-means, k*=6 | 0.4605 | 0.8073 | 0.1150 | *mini batch k-means, k*=6 | 0.5441 | 0.7708 | 0.3832 |
| *mini batch k-means, k*=7 | 0.4305 | 0.7729 | 0.0880 | *mini batch k-means, k*=7 | 0.5187 | 0.7632 | 0.4332 |
| ***Facebook Live Sellers in Thailand*** | | | | ***Gender by Name*** | | | |
| Metod | Siluet | Devis-Boldin | Vaxt (san.) | Metod | Siluet | Devis-Boldin | Vaxt (san.) |
| ***m_k-means, k*=4** | **0.4208** | **0.7652** | **0.1260** | ***m_k-means, k*=4** | 0.5225 | 0.7749 | 1.1967 |
| *k-means, k*=3 | 0.3565 | 0.9257 | 0.1561 | *k-means, k*=3 | **0.5568** | 0.6645 | **1.0131** |
| *k-means, k*=5 | 0.3774 | 0.8605 | 0.1561 | *k-means, k*=5 | 0.5366 | 0.5829 | 2.0233 |
| *k-means, k*=6 | 0.3696 | 0.9144 | 0.2501 | *k-means, k*=6 | 0.5179 | 0.6476 | 2.1514 |
| *k-means, k*=7 | 0.3696 | 0.9179 | 0.2321 | *k-means, k*=7 | 0.5275 | **0.5369** | 2.5627 |
| ***m_k-means++, k*=4** | **0.4207** | **0.7653** | **0.1530** | ***m_k-means++, k*=4** | **0.6122** | **0.4677** | 1.0396 |
| *k-means++, k*=3 | 0.3565 | 0.9257 | 0.1811 | *k-means++, k*=3 | 0.6082 | 0.5264 | **0.8225** |
| *k-means++, k*=5 | 0.3774 | 0.8631 | 0.2011 | *k-means++, k*=5 | 0.5697 | 0.4904 | 1.7541 |
| *k-means++, k*=6 | 0.3711 | 0.9472 | 0.2771 | *k-means++, k*=6 | 0.5761 | 0.4841 | 2.5917 |
| *k-means++, k*=7 | 0.3696 | 0.9179 | 0.2461 | *k-means++, k*=7 | 0.5782 | 0.5005 | 2.2915 |
| ***m_mini batch*** | **0.4209** | **0.7650** | 0.1270 | ***m_mini batch*** | 0.4959 | 0.6500 | 0.9126 |

| ***k-means, k*=4** | | | | ***k-means, k*=4** | | | |
|---|---|---|---|---|---|---|---|
| *mini batch k-means*, k=3 | 0.3529 | 0.9421 | 0.1821 | *mini batch k-means, k*=3 | **0.5516** | **0.6157** | 1.0547 |
| *mini batch k-means, k*=5 | 0.3492 | 1.1504 | 0.1230 | *mini batch k-means, k*=5 | 0.4851 | 0.7657 | 1.1427 |
| *mini batch k-means, k*=6 | 0.3393 | 0.9847 | **0.1150** | *mini batch k-means, k*=6 | 0.5230 | 0.9048 | **0.8491** |
| *mini batch k-means, k*=7 | 0.3488 | 0.9574 | 0.1811 | *mini batch k-means, k*=7 | 0.5144 | 0.6332 | 0.8541 |
| **Anuran Calls** | | | | **Query Analytics Workloads Dataset** | | | |
| Metod | Siluet | Devis-Boldin | Vaxt (san.) | Metod | Siluet | Devis-Boldin | Vaxt (san.) |
| ***m_k-means, k*=3** | **0.6369** | **0.4878** | **0.0640** | ***m_k-means, k*=3** | **0.4355** | **0.8701** | 2.3703 |
| *k-means, k*=4 | 0.5485 | 0.6586 | 0.1010 | *k-means, k*=4 | 0.3625 | 0.9424 | **2.3165** |
| *k-means, k*=5 | 0.5790 | 0.5872 | 0.1240 | *k-means, k*=5 | 0.3507 | 0.9700 | 5.8429 |
| *k-means, k*=6 | 0.5497 | 0.5647 | 0.1390 | *k-means, k*=6 | 0.3658 | 0.9243 | 4.8282 |
| *k-means, k*=7 | 0.5776 | 0.5190 | 0.1841 | *k-means, k*=7 | 0.3724 | 0.8791 | 9.0265 |
| ***m_k-means++, k*=3** | **0.6368** | **0.4892** | **0.0770** | ***m_k-means++, k*=3** | **0.4393** | **0.8635** | **1.8021** |
| *k-means++, k*=4 | 0.6127 | 0.5913 | 0.1250 | *k-means++, k*=4 | 0.3625 | 0.9424 | 2.7660 |
| *k-means++, k*=5 | 0.5757 | 0.5826 | 0.1591 | *k-means++, k*=5 | 0.3507 | 0.9700 | 4.1593 |
| *k-means++, k*=6 | 0.5584 | 0.5568 | 0.1791 | *k-means++, k*=6 | 0.3658 | 0.9276 | 6.3562 |
| *k-means++, k*=7 | 0.5886 | 0.5129 | 0.1951 | *k-means++, k*=7 | 0.3729 | 0.8776 | 8.1301 |
| ***m_mini batch k-means, k*=3** | **0.6369** | **0.4878** | 0.1961 | ***m_mini batch k-means, k*=3** | **0.3854** | 0.9604 | **1.0516** |
| *mini batch k-means, k*=4 | 0.6107 | 0.5401 | **0.0910** | *mini batch k-means, k*=4 | 0.3567 | 0.9562 | 1.1188 |
| *mini batch k-means, k*=5 | 0.5751 | 0.5860 | 0.1110 | *mini batch k-means, k*=5 | 0.3512 | **0.9110** | 1.1348 |
| *mini batch k-means, k*=6 | 0.5641 | 0.5390 | 0.1390 | *mini batch k-means, k*=6 | 0.3625 | 0.9484 | 1.2028 |
| *mini batch k-means, k*=7 | 0.5625 | 0.5705 | 0.2061 | *mini batch k-means, k*=7 | 0.3502 | 0.9044 | 1.1438 |

Qeyd etmək lazımdır ki, qiymətləndirilmə üçün istifadə edilmiş Devis-Boldin indeksinin və zamanın minimal qiymətləri, siluet indeksinin isə maksimal qiyməti daha keyfiyyətli nəticəni ifadə edir. Təklif edilən metodlar üçün ən yaxşı göstərici "*KEGG Metabolic Relation Network*" verilənlərinin nəticələridir. Belə ki, bu verilənlərin nəticələrinə əsasən, Siluet və "Devis-Boldin" indeksinə görə ən yaxşı nəticəni *m_k-means*++ alqoritmi, zamana görə isə *m_mini batch k-means* alqoritmi verir.

*k-means* alqoritminin ilkin mərkəzin təsadüfi seçilməsindən asılılığı onun dəqiq hesablamalar aparmasının qarşısını alır. *k-means*++ alqoritmi vasitəsilə bu asılılığı aradan qaldırmaq mümkündür. *k-means*++ alqoritmin əsas üstünlüyü odur ki, təsadüfi seçilən ilkin klaster mərkəzinə əsasən digər klaster mərkəzləri təyin olunur. *Mini batch k-means* alqoritmin əsas üstünlüyü isə verilənlərin paketlərə bölünməklə analiz edilməsidir ki, bununla emal prosesində texniki təminata qoyulan yüksək tələbləri minimallaşdırmaq olar. Beləliklə, eksperimentin nəticələrini əks etdirən cədvəl 2-dən də göründüyü kimi, təklif edilən metodlardan istifadə edilən zaman böyük ölçülü verilənlərdə *k-means*++ alqoritminin effektivliyi artır.

**Nəticə**

Məqalədə *Data mining* və onun tətbiq sahələri, BV-də klasterləşmə alqoritmləri, *k-means* alqoritminin üstünlükləri və çatışmazlıqları araşdırılmışdır. Qeyd edilən məsələlərin həllinə nail olmaq üçün *k-means, k-means*++ və *mini batch k-means* alqoritmlərinin eksperimental nəticələri

müqayisəli analiz edilmişdir. *PCA* və *elbow* metodu vasitəsilə bu alqoritmlərin effektivliyi artırılmışdır. *k-means* alqoritmi və onun modifikasıyalarının eksperimental nəticələrinin müxtəlif indekslər üzrə qiymətləndirilməsi aparılmışdır. Təklif edilən *elbow* metodunun əsas məqsədi *k-means* alqoritminin klaster sayının təsadüfi seçilməsindən asılılığını aradan qaldırmaqla klasterləşmənin keyfiyyətini artırmaq, *PCA* metodunun isə əsas məqsədi verilənlərin ölçüsünün azaldılması ilə vizuallaşdırmanı asanlaşdırmaq və alqoritmin işləmə prosesini sürətləndirməkdir.

Ənənəvi *k-means, k-means++ , mini batch k-means* alqoritmləri vasitəsilə müxtəlif ölçülü verilənlər üzərində eksperimentlər həyata keçirilmişdir. *k-means* alqoritminin modifikasıyaları vasitəsilə klasterləşdirmənin nəticələrinin keyfiyyəti yaxşılaşdırılmışdır. Eksperimentin nəticələri bəzi qiymətləndirmə indekslərinə və zamana əsasən müqayisə edilmişdir. Araşdırmalar onu göstərdi ki, bu alqoritmlərin əsas üstünlüyü ənənəvi metodla müqayisədə daha sürətli olmasıdır. Eksperimentlərin nəticələrindən də müşahidə etmək olar ki, təklif olunan yanaşmadan istifadə etməklə *k-means* alqoritminin modifikasıya edilmiş versiyaları vasitəsilə böyük həcmli verilənlərin klasterləşdirilməsində effektivliyi artırmaq mümkündür.

**Ədəbiyyat**


1. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011, 744 p.
2. Sanse K., Sharma M. Clustering methods for Big data analysis // International Journal of Advanced Research in Computer Engineering & Technology, 2015, vol.4, no.3, pp.642-648.
3. Chen C.L.P., Zhang C-Y. Data-ıntensıve applıcatıons, challenges, technıques and technologıes: a survey on bıg data // Information Sciences, 2014, vol.275, pp.314-347.
4. Alguliyev R.M, Aliguliyev R.M., Sukhostat L.V. Parallel batch k-means for Big data clustering // Computers & Industrial Engineering, 2021, vol.152.
5. Aliguliyev R.M., Hajırahimova M.Sh., Aliyeva A.Sh. Biġ Data-nın aktual elmi-nazari problemlari // Informasiya Jamiyyati Problemlari, 2016, no.2, pp.37-49.
6. Alguliyev R., Aliguliyev R., Bagirov A., Karimov R. Batch clustering algorithm for big data sets / 2016 IEEE 10th International Conference on Application of Information and Communication Technologies, 2016, pp.79-82.
7. Alguliyev R.M., Aliguliyev R.M., Sukhostat L.V. Weighted consensus clustering and its application to big data // Expert Systems with Applications, 2020, vol.150.
8. Alguliyev R.M., Aliguliyev R.M., Sukhostat L.V. Efficient algorithm for big data clustering on single machine // CAAI Transactions on Intelligence Technology, 2020, vol.5, no.1, pp.9-14.
9. Alıguliyev R., Tahirzada Sh. “Boyuk hajmli fardi malumatların analizi üchün iterativ chakili k-means algoritmi” / “Informasiya tahlukasizliyinin aktual multidissiplinar elmi-praktiki problemlari” V respublika konfransı, 29 noyabr 2019-ju il.


**УДК 004.056**

**Ахмедов Эльтон Я.**
Институт Информационных Технологий НАНА, Баку, Азербайджан
eltunehmedov95@gmail.com

**Сравнительный анализ алгоритмов k-means, k-means++ и mini batch k-means в среде Python**

В этой статье обсуждается применение алгоритма k-средних и его модификации к наборам данных различных измерений в среде Python. При этом были изучены текущее состояние, возможности, недостатки, проблемы традиционного алгоритма кластеризации k-средних и его модификаций и даны предложения по их решению. Алгоритм k-средних ++ устраняет недостаток случайного выбора начальных центров традиционным методом k-средних. Используя мини-пакетный алгоритм *k*-средних, большие данные анализировались путем их разделения на пакеты, что ускоряло процесс анализа больших и сложных данных. Был

предложен гибридный метод PCA и локтя, чтобы уменьшить размерность во время кластеризации данных и найти оптимальное количество кластеров. Чтобы оценить эффективность этого подхода, алгоритмы были протестированы на нескольких наборах данных разного размера. Результаты эксперимента показали, что предложенный подход более эффективен при кластеризации больших данных. Предлагаемый гибридный метод PCA и локтя создает новые возможности для решения задач, требующих больших вычислительных ресурсов в процессе анализа больших многомерных данных.
***Ключевые слова:*** *интеллектуальный анализ данных, кластеризация, k-means, k-means++, mini batch k-means, elbow, PCA.*

**Elton Y. Ahmadov**
Institute of Information Technology of ANAS, Baku, Azerbaijan
eltunehmedov95@gmail.com

**Comparative Analysis of K-Means, K-Means++ and Mini Batch K-Means Algorithms in Python Environment**

This article discusses the application of k-means algorithm and its modifications to datasets of different dimensions in the Python environment. At the same time, the current state, opportunities, shortcomings, problems of the traditional k-means clustering algorithm and its modifications are studied and suggestions for their solution are given. The k-means ++ algorithm eliminates the disadvantage of the traditional k-means method's random selection of starting centers. Using the mini-batch k-means algorithm, big data is analyzed by dividing it into packets, which accelerates the process of analyzing large and complex data. A hybrid PCA and elbow method are proposed to reduce the dimensionality during data clustering and to find the optimal number of clusters. To evaluate the effectiveness of this approach, algorithms are tested on several sets of data of different sizes. The silhouette and Davis-Boldin indices are used to evaluate the efficiency of the algorithms. The results of the experiment show that the proposed approach is more efficient when clustering big data. The proposed hybrid PCA and elbow method create new opportunities for solving problems that require large computing resources in the process of analysing large, multidimensional data
***Keywords:*** *data mining, clustering, k-means, k-means++, mini batch k-means, elbow, PCA.*