Böyük verilənlərdə anomaliyaların aşkarlanması üçün çoxkriteriyalı optimallaşdırma üsulu

Rasim Əliquliyev[1], Ramiz Alıquliyev[2], Yadigar İmamverdiyev[3], Fərqanə Abdullayeva[4]

[1,2,3,4] AMEA İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan

[1]rasim@science.az, [2]r.aliguliyev@gmail.com, [3]yadigar@iit.science.az, [4]a_farqana@mail.ru

***Xülasə*— anomaliyaların aşkarlanmasında klasterləşmə üsullarının tətbiqi effektiv yanaşmalardan hesab edilir. Məşhur k-orta və digər klassik klasterləşmə alqoritmlərində klasterin ilkin mərkəzinin seçilməsi və lokal optimumun tapılması əsas problemlərdən biridir və anomaliyaların aşkarlanmasında dəqiq nəticələr əldə etməyə imkan vermir. Məqalədə anomaliyaların aşkarlanmasında dəqiqliyi artırmaq üçün PSO (particle swarm optimization) və k-orta alqoritmlərinin birləşməsinə əsaslanan yeni çəkili klasterləşmə üsulu təklif edilmişdir. Təklif edilmiş üsul Yahoo! S5 verilənlər bazası üzərində test edilmişdir və alınmış nəticələrin k-orta alqoritmi ilə müqayisəli təhlili aparılmışdır. Eksperimentlərin nəticəsi göstərir ki, təklif edilmiş üsul k-means alqoritmi ilə müqayisədə daha dayanıqlıdır və dəqiq nəticələr əldə etməyə imkan verir.**

***Açar sözlər*— sürü intellektinə əsaslanan optimallaşdırma üsulu (PSO); verilənlərin klasterləşməsi; klaster mərkəzi; k-orta klasterləşmə üsulu**

I. GİRİŞ

Verilənlərin klasterləşməsi öyrədilməyən klassifikasiya üsuludur, məqsədi verilənlər çoxluğunu verilənlər obyektləri arasındakı yaxınlığa görə klasterlərə bölməkdir. Klaster analizi verilənlərin analizi, obrazların tanınması, maşın təlimi, şəkil seqmentləşdirilməsi, neyron hesablamaları və digər elm sahələrinin mühüm vasitəsinə çevrilmişdir. Klasterləşmə alqoritminin məqsədi klasterlər arasındakı məsafəni maksimallaşdırmaq və klaster daxilindəki məsafəni minimallaşdırmaqdır.

Məşhur k-orta (k-means) alqoritmi bir çox praktik klasterləşmə məsələlərinə tətbiq edilmişdir [1, 2]. Bu üsulun məqsədi verilənlər çoxluğunu avtomatik olaraq k sayda qrupa bölməkdir. K-orta alqoritmi sürətli və effektiv nəticələr generasiya edə bilir. Sadə k-means alqoritmində klasterləşmə hər bir nöqtə ilə bu nöqtəyə ən yaxın olan mərkəz arasındakı orta kvadratik məsafə minimallaşdırılmaqla həyata keçirilir. K-orta məşhur alqoritm olduğuna baxmayaraq bu alqoritmin çox sayda çatışmazlıqları vardır. K-orta alqoritmi ilkin klaster mərkəzinin seçilməsindən ciddi asılıdır və klaster mərkəzlərinin əvvəlcədən təyin olunmasını tələb edir. K-means alqoritminin məqsəd funksiyası qabarıq olmadığı üçün bu alqoritmin çox sayda lokal minimum nöqtələri vardır. K-means alqoritminin malik olduğu bu tip problemləri aradan qaldırmaq üçün təkamül alqoritmlərindən istifadə edirlər [3]. PSO təkamül alqoritmlərindən biridir, sürünün təfəkküründə və sosial davranışına əsaslanır.

Ədəbiyyatların analizi göstərir ki, PSO əsasında klasterləşmə mövcud klasterləşmə üsulları ilə müqayisədə daha yüksək nəticələr əldə etməyə imkan verir [4].

PSO əsasında təklif olunan mövcud klasterləşmə alqoritmlərinin bir-birindən fərqi onların məqsəd funksiyalarındadır. [3]-də PSO alqoritmini tətbiq etməklə verilənlərin klasterləşməsini təmin edən üsul təklif edilir. Burada aparılan eksperimentlərdən aydın görünür ki, kriteriyalara çəki əmsallarının verilməməsi səbəbindən instansiyaların səhv identifikasiya edilməsinə çox yol verilmişdir.

Mövcud alqoritmlərdə verilənlərin klasterləşməsi üçün minimallaşdırılacaq kriteriyaların çəkisi eyni götürüldüyündən, məqsəd funksiyasının daha optimal olmasını tənzimləmək mümkün olmur. Bu səbəbdən verilənlərin klasterləşməsi prosesində kriteriyaların vaciblik dərəcəsini göstərmək üçün onlara müvafiq çəki əmsalları verilməlidir [5]. Burada çəki əmsalları verilənlərin klasterləşməsi prosesində kriteriyaların vacibliyini göstərmək üçün istifadə olunur. Kriteriyalara çəkilərin verilməsi daha yaxşı optimal həll tapmağa imkan verir.

Təqdim olunan məqalədə verilənlərin klasterləşməsinin yuxarıdakı problemlərini (klasterləşmə dəqiqliyinin yüksəldilməsi, çox sayda lokal minimum nöqtələrinin olması, klaster mərkəzlərinin əvvəlcədən təyin olunması) aradan qaldırmaq məqsədi ilə çəkili PSO alqoritmi əsasında qurulmuş çoxkriteriyalı optimallaşdırma üsulu təklif edilir. Üsulda klaster daxili məsafənin minimallaşdırılması və klasterlər arasındakı məsafənin maksimallaşdırılması optimallaşdırma kriteriyaları kimi seçilmişdir. İşin əsas yeniliklərinə aşağıdakılar aiddir:

- Bulud mühitində baş verən anomaliyaların çəkili aşkarlanması üsulu təklif olunur. Burada optimallaşdırma məsələsi iki kriteriyanın minimallaşdırılması hesabına təmin olunur. Üsul klaster daxili məsafənin və klasterlər arasındakı məsafələrin cəminin minimallaşdırılmasını məqsəd funksiyası kimi istifadə etmişdir.
- Təklif edilmiş üsula əsasən verilənlərin klasterləşməsi üçün PSO alqoritmi qurulur.

DOI: 10.25045/NCInfoSec.2017.01

- Təklif edilmiş üsulun imkanları Matlab proqram paketində qiymətləndirilir.

- Standart PSO alqoritminin istənilən ölçülü verilənləri klasterləşdirə bilməsi imkanı nümayiş etdirilir;

- PSO alqoritminə əsaslanan yeni klasterləşmə alqoritmi təklif edilir. Burada ilkin sürünü formalaşdırmaq üçün k-orta alqoritmi istifadə edilir.

II. SÜRÜ INTELLEKTİNƏ ƏSASLANAN OPTİMALLAŞDIRMA ÜSULU

Sürü intellektinə əsaslanan optimallaşdırma alqoritmi (PSO) 1995-ci ildə Eberhart və Kennedy tərəfindən yaradılmışdır, popalyasiya tipli stoxastik axtarış prosesi alqoritmidir, quş sürülərinin sosial davranışına əsasən modelləşdirilmişdir [6, 7]. Alqoritmin əsasını hissəciklərin populyasiyası təşkil edir. Bu hissəciklərin hər biri optimallaşdırma məsələsinin mümkün həllini göstərir.

PSO alqoritmində sürü optimallaşdırma məsələsinə bir neçə mümkün həllər generasiya edir. Bu mümkün həllərin hər biri hissəcik adlanır. PSO alqoritminin məqsədi qoyulmuş məqsəd funksiyasını (fitness, objective) daha yaxşı ödəyən hissəciyi (həlli) tapmaqdır.

Hər bir hissəcik N_d ölçülü fəzada bir pozisiyanı göstərir və çoxölçülü axtarış fəzasında öz pozisiyasını aşağıdakılara əsasən tənzimləyərək hərəkət edir (uçur):

- Hissəciyin özünün ən yaxşı pozisiyası (best position);

- Həmin hissəciyin qonşularının ən yaxşı pozisiyası.

Hər bir i -ci hissəcik aşağıdakılardan ibarət olur:

- x_i –hissəciyin cari pozisiyası;

- v_i –hissəciyin cari sürəti;

- y_i –hissəciyin personal ən yaxşı pozisiyası.

Hissəciyin pozisiyası aşağıdakı parametrlərə əsasən tənzimlənir:

$$v_{i,k}(t+1) = \omega\ v_{ik}(t) + c_1\, r_{1k}(t)\ (\overrightarrow{y_{ik}}(t) - x_{ik}(t)) + \\ + c_2\ r_{2k}(t)\ (\overrightarrow{y_k}(t) - x_{ik}(t)) \quad (1)$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (2)$$

burada, ω - inersiya əmsalıdır ($\omega = 0.7298$), c_1 və c_2 sürətləndirmə sabitləridir, $r_{1,j}(t), r_{2,j}(t) \sim U$, $k = 1,..,N_d$.

i -ci hissəciyin personal ən yaxşı pozisiyası aşağıdakı kimi hesablanır :

$$y_i(t+1) = \begin{cases} y_i(t) & f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) & f(x_i(t+1)) < f(y_i(t)) \end{cases} \quad (3)$$

III. KLASTERLƏŞMƏ MƏSƏLƏSININ QOYULUŞU

R^n fəzasında klasterləşmə məsələsi aşağıdakı kimi qoyulur: n nöqtədən ibarət $x_1, x_2, ..., x_n$ nöqtələr çoxluğunun yaxınlıq meyarına görə k (məlum sabit ədəd) sayda $G_1, G_2, ..., G_k$ çoxluğuna bölünməsi həyata keçirilir.

Burada bölünmə zamanı aşağıdakı şərtlər ödənməlidir:

1) $G_i \neq \varnothing, \quad i = 1, 2, ..., k;$ (4)

2) $G_i \cap G_j = \varnothing, \ i, j = 1, 2, ..., k; i \neq j$ (5)

3) $\bigcup_{i=1}^{k} G_i = \{x_1, x_2, ..., x_n\}$ (6)

Klasterləşmə bir-birinə yaxın obyektlərin eyni klasterə təyin edilməsini təmin etməklə obyektlər çoxluğunu bir neçə klasterdə qruplaşdıran prosesdir. K-orta alqoritmi ən məşhur və geniş istifadə olunan klasterləşmə üsuludur. K-orta alqoritmi hər bir nöqtədən ona ən yaxın klaster mərkəzinə qədər olan məsafənin kvadratının cəminin minimum olduğu $C_1, C_2, ..., C_k$ klasterlər mərkəzlərini tapmağa cəhd edir (7).

$$D = \sum_{i=1}^{n} \left[\underbrace{\min}_{k=1, 2, ..., K} d(x_i, c_k) \right]^2 \quad (7)$$

burada d hər hansı bir məsafə funksiyasıdır. Təqdim olunan məqalədə bu funksiya Evklid götürülmüşdür.

Ənənəvi k-orta alqoritmini aşağıdakı addımlar təşkil edir:

Addım 1. k sayda klasterlərin təyin edilməsi və hər bir klaster üçün $\left(C_1^{(0)}, C_2^{(0)}, ..., C_k^{(0)}\right)$ mərkəzin elan edilməsi. Hər bir klaster mərkəzi m ölçülü vektordur məsələn, $C_i^{(0)} = \left\{C_{i1}^{(0)}, C_{i2}^{(0)}, ..., C_{im}^{(0)}\right\}$.

Addım 2. i -ci verilənlər çoxluğu ilə (m ölçülü fəzada nöqtə) k -cı klaster mərkəzi arasında $d_{ki}^{(t-1)}$ məsafəsinin hesablanması. Məsafə meyarı kimi (8) düsturunda verilmiş Evklid məsafəsi istifadə edilmişdir.

$$d_{ki}^{(t-1)} = \left\| x_i - C_k^{(t-1)} \right\| = \sqrt{\sum_{j=1}^{m} \left(x_{ij} - C_{kj}^{(t-1)} \right)^2} \quad (8)$$

Addım 3. Hər bir x_i verilənlər obyektinin ən yaxın C_k klaster mərkəzinə təyin edilməsi.

Addım 4. Hər bir $C_k^{(t)}$ klaster mərkəzinin ona daxil olan bütün x_i nöqtələrinin ortalama qiymətini hesablayan (9) düsturu vasitəsi ilə yenilənməsi.

$$C_k^{(t)} = \frac{\sum_{x_i \in k} x_i}{n_k} \quad (9)$$

burada n_k k -cı klasterə daxil olan nöqtələrin sayıdır.

Addım 5. (7) düsturu vasitəsi ilə klasterdaxili D məsafənin hesablanması.

Addım 6. Əgər D -nin qiyməti qənaətbəxşdirsə yekun klaster mərkəzlərinin seçilməsi. Əks halda $t = t + 1$ -ci iterasiyaya keçid edərək 2-ci addıma qayıtmaq.

IV. PSO KLASTERLƏŞMƏSI ÜÇÜN TƏKLIF EDILƏN OPTIMALLAŞDIRMA FUNKSIYASI

PSO klasterləşmə alqoritmində hər bir $Y_i = (y_1, y_2, ..., y_k)$ həlli (particle) k sayda sinfin mərkəzlərini əks etdirir. Burada həllər sürüsü bir neçə namizəd kimi klassifikasiya olunmuş plandan ibarət olur. Optimallaşdırma alqoritmində bu namizəd kimi klassifikasiya olunmuş planlardan qoyulmuş şərti ödəyən planın seçilməsi üçün məqsəd funksiyasından istifadə edilir. Bu məqsədlə təqdim olunan məqalədə aşağıdakı məqsəd funksiyası təklif edilir:

$$f = (1-\alpha) \times \sum_{i=1}^{n} \sum_{j=1}^{c} \left\| x_i - c_j \right\| + \alpha \times \sum_{k,j=1}^{c} \left\| c_k - c_j \right\| \to \min \quad (10)$$

Burada məqsəd (10) düsturu vasitəsi ilə verilmiş qiymətləndirmə funksiyasının aldığı qiyməti minimallaşdırmaqdır. Yəni (10) funksiyasının minimum qiymətində klasterləşmənin daha effektiv aparılacağı fərz edilir. $(1-\alpha)$ və α uyğun olaraq J_1 və J_2 faktorlarının çəki əmsallarıdır, J_1 və J_2 faktorlarının qiymətləndirməyə təsirini göstərir. Aparılan bir sıra eksperimentlərin nəticəsində çəki əmsalının $\alpha = 0.731$ qiymətində klasterləşmənin nəticəsi nisbətən sabit və daha yaxşı olmuşdur. Bu səbəbdən məqalədə çəki əmsalının qiyməti $\alpha = 0.731$ götürülmüşdür.

f funksiyasının minimal qiyməti eyni sinifdə nöqtələr arasındakı məsafənin kiçik olması və siniflər arasındakı məsafənin böyük olması şərtlərini ödəyir. f funksiyasının qiyməti minimum olan klassifikasiya planı ən yaxşı hesab edilir.

Məqsəd funksiyasını təşkil edən iki kriteriya aşağıdakılardır.

a) Klaster daxili məsafə (inner-cluster distance) – klasterin bütün nöqtələrindən onun mərkəzinə qədər olan məsafədir, alqoritmin məqsədi klaster daxili məsafəni minimallaşdırmaqdır. Bu kriteriya aşağıdakı düsturla hesablanır:

$$J_1 = (1-\alpha) \times \sum_{i=1}^{n} \sum_{j=1}^{c} \left\| x_i - c_j \right\| \quad (11)$$

burada, c_j j -cu klasterin mərkəzidir, x_i - c_j klasterinə daxil olan nöqtələrdir.

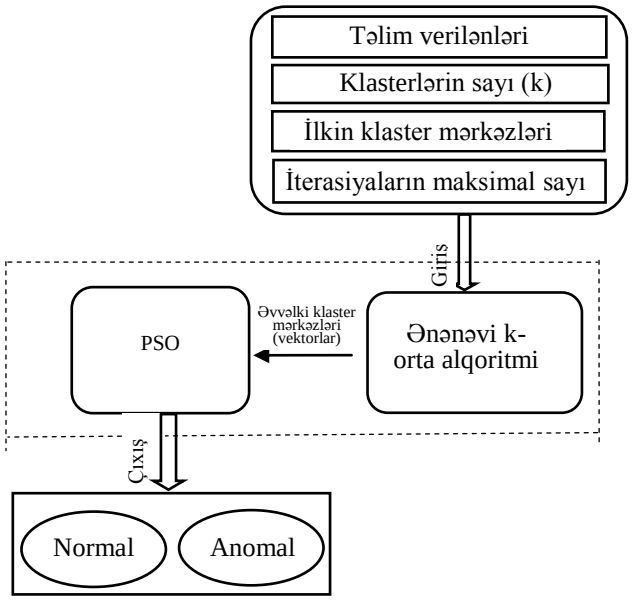
b) Klasterlər arasındakı məsafə (inter-cluster distance) – klasterlərin mərkəzləri arasında məsafədir, məqsədi klasterlər arasındakı məsafəni maksimallaşdırmaqdır. Bu kriteriya aşağıdakı düsturla hesablanır:

$$J_2 = \alpha \times \sum_{k,j=1}^{c} \left\| c_k - c_j \right\| \quad (12)$$

burada, c_k və c_j uyğun olaraq k -cı və j -cu klasterlərin mərkəzidir.

V. ANOMALİYALARIN AŞKARLANMASI ÜÇÜN TƏKLİF EDİLMİŞ OPTIMALLAŞDIRMA MODELI

Anomaliyaların aşkarlanması üçün təklif edilmiş optimallaşdırma modeli şəkil 1-də təsvir edilmişdir. Şəkildən göründüyü kimi model PSO və ənənəvi k-orta alqoritmlərinin birləşdirilməsi ideyasına əsaslanır.



Şəkil 1. Anomaliyaların aşkarlanması üçün optimallaşdırma modeli

Bu struktura uyğun olaraq anomaliyaların aşkarlanması üçün PSO əsasında optimallaşdırma məsələsinin alqoritmi aşağıdakı addımlardan ibarətdir:

Addım 1. Populyasiyanın ölçüsünün m, c_1 və c_2 sürətləndirmə əmsallarının, iterasiyaların sayının, kiçik təsadüfi pozisiyalardan ibarət ilkin x_p həllər populyasiyasının, inersiya əmsalının (ω), klasterlərin sayının k , n nöqtədən ibarət verilənlər çoxluğunun daxil edilməsi.

Addım 2. k-orta alqoritmi vasitəsi ilə populyasiyanın hər bir hissəciyi üçün aşağıdakı addımların yerinə yetirilməsi:

1) (8) düsturundan istifadə etməklə p -ci klaster mərkəzi (həll) ilə i -ci nöqtə arasında d_{pi} Evklid məsafəsinin hesablanması.
2) Hər bir x_i nöqtəsinin ən yaxın X_p klaster mərkəzinə təyin edilməsi.

Addım 3. Verilənlər obyektlərini minimum məsafə kriteriyasına əsasən qruplaşdırdıqdan sonra, yəni klasterlərin sinifləri tapıldıqdan sonra klasterləşmənin dəqiqliyini artırmaq

üçün təklif edilmiş (10) düsturu vasitəsi ilə məqsəd funksiyasının qiymətləndirilməsi.

Addım 4. Məqsəd funksiyasının qiymətinə görə qiymətləndirmənin nəticəsinin həllin əvvəlki P_{best} ən yaxşı qiyməti ilə müqayisə edilməsi. Əgər cari pozisiya qiyməti (klaster mərkəzinin mövqeyi) P_{best} -dən yaxşı olarsa, cari pozisiyanı P_{best} -in yerinə təyin edilməsi, əks halda P_{best} -in əvvəlki qiymətinə bərabər saxlanması. Bu proses populyasiyanın hər bir həllinə tətbiq olunur.

Addım 5. P_{best} yeniləndikdən sonra məqsəd funksiyasının ən yaxşı qiymətinin (məqsəd funksiyasının qiyməti minimum olanın) seçilməsi və onun G_{best} kimi təyin edilməsi. G_{best} $k \times m$ ölçülü vahid həlldir (particle). k klasterlərin sayıdır, verilənlər bazasını bölmək üçün təyin edilmişdir.

Addım 6. Hər bir həllin sürəti və pozisiyası uyğun olaraq (1) və (2) düsturları vasitəsi ilə yenilənir.

Addım 7. Konvergentlik kriteriyasının yoxlanması. Bu kriteriyalar kimi məqsəd funksiyasının ən yaxşı qiyməti və ya iterasiyaların maksimal sayı götürülür. Əgər konvergentlik kriteriyası ödənilibsə G_{best} qiyməti optimal klaster mərkəzi kimi qeyd olunur, əks halda iterasiyaların sayı $t = t + 1$ addım artırılır və addım 2-yə qayıdılır.

VI. EKSPERIMENTLƏR

Bu bölmədə effektivliyi qiymətləndirmək üçün təklif edilmiş üsulun və k-means alqoritminin Yahoo! S5 real verilənlər bazası üzərində klasterləşmə nəticələrinin müqayisəli analizi aparılır. PSO əsasında klasterləşmə alqoritmində sürətləndirmə əmsalı $C_1 = C_2 = 1.4962$ götürülmüşdür.

Test prosesində populyasiyanın ölçüsü 50 götürülərək məqsəd funksiyasının qiyməti (10) düsturuna əsasən hesablanmışdır.

İstifadə edilmiş A1Benchmark Yahoo! S5 verilənlər bazasının real_2.csv faylı 1440 sətirdən ibarətdir. Bu bazanın əlamətlər vektorunu vaxt qeydiyyatı və qiymət kimi iki parametr təşkil edir və burada həmçinin əlamətin normal və ya anomal olduğu da göstərilir. Yəni baza təsnif olunmuş bazadır.

Eksperimentlərin aparılması üçün Yahoo! S5 bazasından ümumi olaraq 84 nöqtə götürülmüşdür, onlardan 68-i normal 16-sı anomal nöqtələrdir. Bu baza üzərində aparılan eksperimentlərdə k-means alqoritminin nəticəsinə görə verilənlərin 79-u normal 5-i anomal kimi identifikasiya edilmişdir. Burada 11 nöqtə yanlış identifikasiya edilmişdir.

Təklif edilmiş PSO alqoritmində isə ümumi verilənlərin 72-si normal 12-si anomal kimi identifikasiya edilmişdir. Bu alqoritmdə 4 nöqtə yanlış identifikasiya edilmişdir.

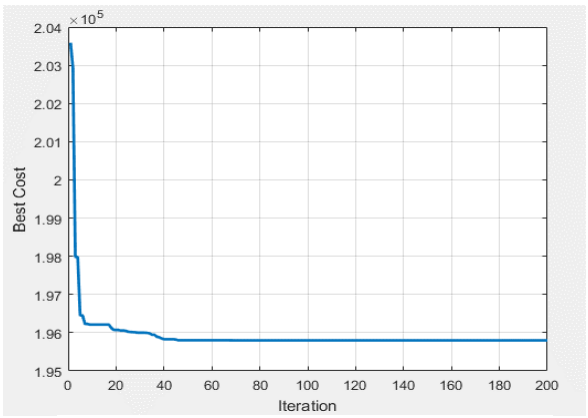
Klasterləşmənin effektivliyi dörd metrika üzərində qiymətləndirilmişdir: Dunn`s index, Silhouette index, Purity index, Entropy Index (cədvəl 1) [8, 9, 10].

	Ümumi	*Normal*	*Anomal*	*Səhv normal*	*Səhv anomal*	*Dunn`s index*	*Silhouette index*	*Purity index*	*Entropy index*
K-means	84	79	5	11	-	0,05	0,39	0,87	0,58
PSO	84	72	12	4	-	0,38	0,87	0,95	0,31
Real	84	68	16	-	-	-	-	-	-

CƏDVƏL 1. $\alpha = 0.731$ QIYMƏTINDƏ PSO ALQORITMININ KLASTERLƏŞMƏ NƏTICƏLƏRI

Burada klasterləşmənin qiymətləndirilmə metrikalarına görə k-means alqoritminin Dunn indeksi 0,0510, PSO alqoritminin Dunn indeksi isə 0,3847 təşkil etmişdir. Qeyd edək ki, Dunn indeksi nə qədər böyük olsa alqoritm bir o qədər effektiv hesab edilir. Silhouette indeksinə görə k-means alqoritmi 0,3899, PSO alqoritmi 0,8722 olmuşdur. Təklif edilmiş üsuldan digər metrikalar üzrə də yaxşı nəticələr əldə edilmişdir. Belə ki, Purity indeksinə görə k-means alqoritmi 0,8690, PSO alqoritmi 0,9524 qiymət almışdır. Entropiyanın hesablanması zamanı isə k-means alqoritminin entropiyası 0,5821, PSO alqoritminin entropiyası isə 0,3096 təşkil etmişdir. Klasterləşmə məsələsində entropiyanın az olması üsulun daha effektiv olduğunu göstərir.

Təklif edilən PSO alqoritminin iterasiyalarının sayı 200 götürülmüşdür və alınmış nəticələr get-gedə yaxşılaşaraq optimal həll (BestCost) tapılmışdır (şəkil 2).



Şəkil 2. PSO əsasında optimal həll dinamikası

NƏTİCƏ

Araşdırmalar göstərir ki, PSO alqoritmi, PSO-nun modifikasiyaları və onun müxtəlif alqoritmlərlə hibridləşdirilməsi optimallaşdırma məsələsinin həllində effektivlik və dəqiqlik baxımından çox yaxşı nəticələr verir. Bu alqoritmin verilənlərin klasterləşməsi məsələsinə tətbiqi klasterlərin dəqiq yaradılmasına və verilənlərin dəqiq proqnoz və klasterləşməsinə imkan verir. Məqalədə verilənlərin

klasterləşməsi üçün PSO alqoritminə əsaslanan çoxkriteriyalı optimallaşdırma metodu təklif edilmişdir. Metodun k-orta alqoritmi ilə müqayisəli dəqiqliyi Yahoo! S5 verilənlər bazası üzərində test edilmişdir. Eksperimentlərin nəticəsi PSO əsasında klasterləşmə üsulunun k-orta alqoritmindən daha yaxşı olduğunu göstərmişdir.

MİNNƏTDARLIQ

Bu iş Azərbaycan Respublikasının Prezidenti yanında Elmin İnkişafı Fondunun maliyyə yardımı ilə yerinə yetirilmişdir— **Qrant № EİF-KETPL-2-2015-1(25)-56/05/1**

ƏDƏBİYYAT

[1] S.Z. Selim, M.A. Ismail, “K-means-type algorithms: A generalized convergence theorem and characterization of local optimality,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, vol 6, no. 1, pp. 81-87.

[2] H.T. Sarma, P. Viswanath, B.E. Reddy, “A hybrid approach to speed-up the k-means clustering method,” International Journal of Machine Learning and Cybernetics, 2013, vol 4, no. 2, pp. 107-117.

[3] S. Rana, S. Jasola, R. Kumar, “A boundary restricted adaptive particle swarm optimization for data clustering,” International Journal of Machine Learning and Cybernetics, 2013, vol. 4, no. 4, pp 391-400.

[4] R.J. Kuo, M.J. Wang, T.W. Huang, “An application of particle swarm optimization algorithm to clustering analysis,” Soft Computing, 2011, vol. 15, no. 3, pp. 533-542.

[5] R.M. Alguliyev, Y.N. Imamverdiyev, F.C. Abdullayeva, “Multıcrıterıa optımızatıon method for load balancıng ın cloud computıng,” Problems of information technology, 2017, № 2, pp. 3-15.

[6] J. Kennedy, R. Eberhart, “Particle Swarm Optimization,” Proc. of the lEEE International Conference on Neural Networks, 1995, vol. 4, pp. 1942-1948.

[7] R. Eberhart, Y. Shi, J. Kennedy, “Swarm Intelligence,” 1st edition, 2002, 512 p.

[8] J.C. Dunn, “Well Separated Clusters and Optimal Fuzzy Partitions,” Journal of Cybernetics, 1974, vol. 4, no. 1, pp. 95-104.

[9] S. Saitta, B. Raphael, F.C. Smith, “A Bounded Index for Cluster Validity,” Proc. of the International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2007, pp. 174-187.

[10] R.M. Aliguliyev, “Performance evaluation of density-based clustering methods,” Information Sciences, 2009, vol. 179, no. 20, pp. 3583-3602.

MULTI-CRITERION OPTIMIZATION METHOD FOR ANOMALY DETECTION ON BIG DATA

Rasim Alguliyev[1], Ramiz Aliguliyev[2],
Yadigar Imamverdiyev[3], Fargana Abdullayeva[4]
[1,2,3,4]Institute of Information Technology of ANAS,
Baku, Azerbaijan
[1]r.alguliev@gmail.com, [2]r.aliguliyev@gmail.com,
[3]yadigar@iit.science.az, [4]a_farqana@mail.ru

Abstract – The use of clustering methods in anomaly detection is considered as an effective approach. The choice of the cluster primary center and the finding of local optimum in the well-known k-means and other classic clustering algorithms are considered as one of the major problems and do not allow to get accurate results in anomaly detection. In this paper to improve the accuracy of anomaly detection based on the combination of PSO (particle swarm optimization) and k-means algorithms, the new weighted clustering method is proposed. The proposed method is tested on Yahoo! S5 dataset and a comparative analysis of the obtained results with the k-means algorithm is performed. The results of experiments show that the proposed method is more robust compared with the k-means algorithm and allows to get more accurate results.

Keywords – particle swarm optimization (PSO); data clustering; cluster center; k-means