РУХАНИ ЖАҢҒЫРУ

20 АСТАНА

ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ

Студенттер мен жас ғалымдардың
«ҒЫЛЫМ ЖӘНЕ БІЛІМ - 2018»
XIII Халықаралық ғылыми конференциясы

СБОРНИК МАТЕРИАЛОВ

XIII Международная научная конференция
студентов и молодых ученых
«НАУКА И ОБРАЗОВАНИЕ - 2018»

The XIII International Scientific Conference
for Students and Young Scientists
«SCIENCE AND EDUCATION - 2018»

12th April 2018, Astana

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ
Л.Н. ГУМИЛЕВ АТЫНДАҒЫ ЕУРАЗИЯ ҰЛТТЫҚ УНИВЕРСИТЕТІ

Студенттер мен жас ғалымдардың «Ғылым және білім - 2018» атты XIII Халықаралық ғылыми конференциясының БАЯНДАМАЛАР ЖИНАҒЫ

СБОРНИК МАТЕРИАЛОВ XIII Международной научной конференции студентов и молодых ученых «Наука и образование - 2018»

PROCEEDINGS of the XIII International Scientific Conference for students and young scholars «Science and education - 2018»

2018 жыл 12 сәуір

Астана

УДК 378

ББК 74.58
Ғ 96

Ғ 96
«Ғылым және білім – 2018» атты студенттер мен жас ғалымдардың XIII Халықаралық ғылыми конференциясы = XIII Международная научная конференция студентов и молодых ученых «Наука и образование - 2018» = The XIII International Scientific Conference for students and young scholars «Science and education - 2018». – Астана: http://www.enu.kz/ru/nauka/nauka-i-obrazovanie/, 2018. – 7513 стр. (қазақша, орысша, ағылшынша).

ISBN 978-9965-31-997-6

Жинаққа студенттердің, магистранттардың, докоранттардың және жас ғалымдардың жаратылыстану-техникалық және гуманитарлық ғылымдардың өзекті мәселелері бойынша баяндамалары енгізілген.

The proceedings are the papers of students, undergraduates, doctoral students and young researchers on topical issues of natural and technical sciences and humanities.

В сборник вошли доклады студентов, магистрантов, докторантов и молодых ученых по актуальным вопросам естественно-технических и гуманитарных наук.

УДК 378

ББК 74.58

ISBN 978-9965-31-997-6

ДЕРЕКТЕРДІ КЛАСТЕРЛЕУДЕ R БАҒДАРЛАМАЛАУ ТІЛІНІҢ МҮМКІНДІКТЕРІ

Ерназар Б.Н.
e_bisultan@mail.ru,
Ғылыми жетекші М.С.Сауханова
m.saukhanova@mail.ru
Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан

Кластерлеу статистика, тиімділеу, деректерді интеллектуалды талдау, бейнелерді сегменттеу, бейнелерді тану т.б. ғылым салаларында кеңінен қолданысқа ие.

Кластерлеу - берілген объектілер (деректер) жиынын кластерлер деп аталатын ішкі жиындарға бөлу. Кластер - алдын-ала анықталған қасиеттері бойынша өзара ұқсас объектілер тобы. Қасиеттер ретінде объектінің қандайда бір сандық мінездемесі қарастырылады [1].

Нақты есеп ерекшелігіне байланысты кластерлеу әртүрлі мақсаттарды көздеуі мүмкін:

- объектілерді ұқсас топтарға бөлу арқылы деректер жиынының құрылымын анықтау;
- ешқандай кластерге жатқызуға болмайтын объектілерді оқшаулау;
- деректермен жұмыс істеуді жеңілдету, яғни кластерлердегі барлық деректермен емес, тек олардың типтік өкілдерімен ғана жұмыс істеу.

Деректерді кластерлеу келесі кезеңдерден тұрады:

- объектілер қасиеттерін/мінездемелерін анықтау (ұқсастықтарын анықтау);
- метриканы анықтау;
- объекттерді топтарға бөлу;
- нәтижені визуализациялау.

Объектілер қасиеттерін/мінездемелерін анықтау кезеңінде объектілердің әрқайсысын айқын көрсететін сипаттамалары анықталады. Ол сандық (координаталар, интервалдар т.б.) немесе сапалық (түсі, статусы, дәрежесі т.б., ескерте кетеміз, өңдеу барысында сапалық сипаттама да сәйкес сандық баламаға ауыстырылады) болуы мүмкін.

Содан соң сипаттамалар аймағын қысқартуға ұмтылу, яғни, объектілердің ең маңызды сипаттамаларын белгілеу керек. Сипаттамалар аймағын қысқарту кластерлеу процесін жеделдетеді және көп жағдайда визуалды түрде баға беруге мүмкіндік тудырады.

Ары қарай объектілер сипаттамалық векторлар күйінде беріледі.

Кластерлеудің келесі кезеңі, объектілердің ұқсастықтарын көрсететін, *метриканы анықтау* болып саналады. Метриканы таңдау мәселесі кластерлеу есебінің нәтижесіне үлкен әсер ететін факторлардың бірі. Сипаттамалар сандық болмаған жағдайда ең қарапайым Хэмминг метиркасын қолдануға болады. Көп жағдайда объектілердің сандық сипаттамалары үшін классикалық Евклид метрикасы қолданылады.

Объектілерді топтарға (кластерлерге) бөлу үшін әртүрлі кластерлеу алгоритмдері қолданылады (иерархиалық, k-means, жақын көршілер алгоритмі, бұлдыр кластерлеу алгоритмдері, нейрондық желілер т.б.)

Нәтижені визуализациялау кезеңінде кластерлеу алгоритмдері арқылы топтарға (кластерлерге) жіктелген деректерді графикалық кескін түрінде бейнелейді (иерархиялық алгоритмдер үшін дендрограммалар, гистограммалар, диаграммалар т.б.).

Деректерді талдауда R бағдарламалау тілін қолдану

Кластерлеу есептерін шығаруда R бағдарламалау тілі деректерді және алгоритмдер жұмыстарын талдау, кластерлеу нәтижелерін ұсыну мүмкіндіктерін ұтымды қамтамасыз етеді [2]. Сондықтан R бағдарламалау тілінде деректерге кластарлеу жүргізу мысалдарын қарастыру көзделді.

Мысал ретінде «DBforR.txt» файлына жазылған деректер қорын қолданамыз. Файл мектеп оқушыларының 2006-2009 жылдар аралығында әлеуметтік желі арқылы бір-біріне

жазған хабарламаларының сипаттамаларынан тұрады. Деректер қорында 30 000 дерек және 40 баған бар. Алғашқы төрт баған хабарлама жазылған уақыт, жіберген оқушының жасы, жынысы, достарының саны секілді ақпараттардан тұрса, қалған 36 бағанда оқушылардың жазған хабарламаларын талдауға көмектесетін арнайы сөздердің хабарламаларда қанша рет кездескені көрсетілген. Анонимдікті сақтау мақсатында, хабарламалардың авторлары көрсетілмеген [3].

Файлды R бағдарламалау ортасына жүктеу:

> teens<-read.csv("D:/ DBforR.txt ")

>fix(teens)

	Уакыты	жынысы	жасы	достары	баскетбол	футбол	доп	регби	волейбол	бассейн	черлидер
1	2006	M	18.982	7	0	0	0	0	0	0	0
2	2006	F	18.801	0	0	1	0	0	0	0	0
3	2006	M	18.335	69	0	1	0	0	0	0	0
4	2006	F	18.875	0	0	0	0	0	0	0	0
5	2006		18.995	10	0	0	0	0	0	0	0
6	2006	F	NA	142	0	0	0	0	0	0	0
7	2006	F	18.93	72	0	0	0	0	0	0	0
8	2006	M	18.322	17	0	0	0	1	0	0	0
9	2006	F	19.055	52	0	0	0	0	0	0	0
10	2006	F	18.708	39	0	0	0	0	0	0	0
11	2006	F	18.543	8	0	0	0	0	0	0	0
12	2006	F	19.463	21	0	1	0	0	0	0	0
13	2006	F	18.097	87	0	0	0	0	0	0	0
14	2006		NA	0	0	0	0	0	0	0	0
15	2006	F	18.398	0	0	0	0	0	0	0	0
16	2006		NA	0	0	0	0	0	0	0	0
17	2006		NA	135	0	0	0	0	0	0	0
18	2006	F	18.087	26	0	0	0	0	0	0	0

Сурет 1. fix(teens) командасы арқылы алынған кесте

Талдау жасауды бастамас бұрын деректерді белгілі бір нормаға келтіру қажет. Ол үшін summary() командасын пайдалану арқылы әр бағанға сипаттама жүргіземіз:

```
> summary(teens)
     Уакыты        жынысы           жасы            достары
 Min.   :2006   F   :22054   Min.   :  3.086   Min.   :  0.00
 1st Qu.:2007   M   : 5222   1st Qu.: 16.312   1st Qu.:  3.00
 Median :2008   NA's: 2724   Median : 17.287   Median : 20.00
 Mean   :2008                Mean   : 17.994   Mean   : 30.18
 3rd Qu.:2008                3rd Qu.: 18.259   3rd Qu.: 44.00
 Max.   :2009                Max.   :106.927   Max.   :830.00
                             NA's   :5086
   баскетбол          футбол             доп              регби
 Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
 1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
 Mean   : 0.2673   Mean   : 0.2523   Mean   : 0.2228   Mean   : 0.1612
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
 Max.   :24.0000   Max.   :15.0000   Max.   :27.0000   Max.   :17.0000
```

Сурет 2. summary() командасының орындалуы

Алынған нәтижеде оқушылардың жастарына мән беретін болсақ, ең кішісі жас 3-ке, ал ең үлкені 106 жасқа тең, яғни қорда деректері қате енгізілген оқушылардың бар екенін көреміз. Сондықтан бізге керекті 13 пен 20 жас аралығындағыларды қалдырып, қалғанына «белгісіз» (NA) мәнін беріп шығамыз:

```
> teens$жасы<-ifelse(teens$жасы >= 13 & teens$жасы <= 20, teens$жасы, NA)
> summary(teens$жасы)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  13.03   16.30   17.27   17.25   18.22   20.00    5523
```

Сурет 3. Жастары көрсетілген бағанды нормалау

Келесі туындайтын ең маңызды сұрақ: «белгісіз» деректерді қалай нормалаймыз? Әрине, ең оңай жолы, «белгісіз» мәні бар жолдарды жойып тастау. Бірақ бұл жол қажетті көптеген деректерден айырылып қалуға алып келеді. Сондықтан, әдетте, оларды «жалған деректермен» алмастырады. Жалған деректер ретінде әр жылдағы оқушылардың орташа жасы алынады:

```
> ave_age <- ave(teens$жасы, teens$Уакыты, FUN =function(x) mean(x, na.rm = TRUE))
> teens$жасы <- ifelse(is.na(teens$жасы), ave_age, teens$жасы)
> summary(teens$жасы)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.03   16.28   17.24   17.24   18.21   20.00
```

Сурет 4. Жастары көрсетілген бағанды нормалау

Енді оқушылар жастарының ішінде белгісіз мән жоқ.

Деректерді нормаға келтіргеннен соң кластерлік талдауға көшуге болады. Бізге оқушыларды қызығушылықтары бойынша кластерлерге жіктеу қажет. Қызығушылықтарын сипаттайтын 36 бағанды (5-тен 40-қа дейін) бөлек алып, оларды *стандартты* түрге келтіреміз. Деректерді стандартты түрге келтіру дегеніміз – оларды орташа мәні 0-ге, орташа квадраттық ауытқуы 1-ге тең болатындай етіп түрлендіру. Стандарттауды R тілінде scale() функциясын шақыру арқылы оңай жүзеге асыруға болады:

```
> interests <- teens[1:30000,5:40]
> interests_z <- as.data.frame(lapply(interests, scale))
> summary(interests_z )
    баскетбол           футбол              доп               регби
 Min.   :-0.3322   Min.   :-0.3577   Min.   :-0.2429   Min.   :-0.2179
 1st Qu.:-0.3322   1st Qu.:-0.3577   1st Qu.:-0.2429   1st Qu.:-0.2179
 Median :-0.3322   Median :-0.3577   Median :-0.2429   Median :-0.2179
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
 3rd Qu.:-0.3322   3rd Qu.:-0.3577   3rd Qu.:-0.2429   3rd Qu.:-0.2179
 Max.   :29.4923   Max.   :20.9081   Max.   :29.1937   Max.   :22.7642
```

Сурет 5. 5-тен 40-қа дейінгі бағандарды стандарттау

5-ші суреттен көріп отырғанымыздай барлық бағандардың орташа мәні 0-ге тең болып тұр. Енді k-means функциясын шақырып деректерді k=5 кластерлерге жіктейік:

```
> teen_clusters <- kmeans(interests_z, 5)
> teen_clusters
K-means clustering with 5 clusters of sizes 675, 25301, 986, 2436, 602

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Сурет 6. k-means функциясын қолдану арқылы деректерді талдау.

Деректер 5 кластерге жіктелінді. Кластерлеу нәтижесін teen_clusters объектісіне жаздық. Бұл объектінің бірнеше негізгі атрибуттары бар. Бірнешеуін қарап көрейік:

> teen_clusters$size

[1] 675 25301 986 2436 602

> teen_clusters$centers

```
> teen_clusters$centers
    баскетбол      футбол         доп        регби    волейбол     бассейн
1  0.14092994  0.05397322  0.03494031  0.036430852  0.03330160  0.08963562
2 -0.05838929 -0.05970126 -0.02888039 -0.027759138 -0.03325367 -0.04299989
3  0.16435904  0.22463886  0.09769303  0.075487396  0.21374657  0.25989404
4  0.52379370  0.50015879  0.27539872  0.246023902  0.26898914  0.30776977
5 -0.09275704  0.05679227 -0.09979815  0.006641763 -0.07830402  0.03563688
     черлидер     бейзбол      теннис       спорт     суйкімді       жаным
1  0.01158639  0.03455766  0.06605559  0.03623664  0.007504876 -0.02419020
2 -0.04510290 -0.03355393 -0.02153125 -0.06555159 -0.102147167 -0.09924199
3  0.46508503  0.03797609  0.15673077  0.08179563  0.409062924  0.02304765
4  0.30029279  0.35122649  0.13202082  0.66405820  0.898580209  1.03801692
5 -0.09428488 -0.11197774  0.03992583 -0.10670296 -0.021458053 -0.04000715
     тартымды         зор      суйісу          би       талпа         шеру
1 -0.06537811  0.00168510 -0.05136662  0.01021042  0.07252572 -0.04337254
2 -0.06274222 -0.06146298 -0.13844159 -0.07843610 -0.11591653 -0.11264545
3  0.12629603  0.40253483  0.06818860  0.23884336 -0.10552371 -0.11315732
4  0.62565123  0.48458152  1.43027120  0.70419232  0.21581514 -0.05705503
5 -0.02831103 -0.03887460 -0.02324169  0.04436493  4.08998351  5.19913301
      музыка         рок        кудай        мешіт   пайгамбар        куран
1  0.2093371  0.11259574  2.375621323  2.196368390  2.45089976  4.18140327
2 -0.1055958 -0.09780300 -0.089997318 -0.083304223 -0.06722452 -0.10424551
3  0.1081088  0.03811592 -0.003613504  0.007670277 -0.00609823 -0.07946606
4  0.8721219  0.92729081  0.258711405  0.237469948  0.01835201 -0.03203110
5  0.4971643  0.16951622  0.077773541  0.064936545  0.01295263 -0.04742565
```

7-сурет. teen_clusters объектісінің centers атрибутын қарау.

Біздің деректер стандартталғандықтан әр сөздің кластер нөмірі бойынша теріс мәндері, сол сөздің жалпы кездесу жиілігінің орташа мәнінен төмен екенін көрсетсе, оң мәндер керісінше орташа мәнінен жоғары екенін көрсетеді. Бұл ақпараттарды пайдаланып, әр кластердегі оқушылардың қызуғышылықтарын анықтауға болады. Мысалы 1-кластер «кудай», «мешіт», «пайгамбар», «куран» сөздері бойынша жоғары оң сандарды көрсетіп тұр (2.37, 2.19, 2.45, 4.18), яғни 1-кластердегі оқушылардың көбін «религия» тақырыбы қызықтыратынын біле аламыз (size атрибутынан 1-кластерге 30000 хабарламаның 675-і кірістірілгенін көруге болады).

Бастапқы деректер қорындағы әр хабарлама қай кластерге жататынын teen_clusters$cluster бағанын пайдаланып біле аламыз:

```
> teens$cluster <- teen_clusters$cluster
> teens[1:20, c("cluster", "жынысы", "жасы", "достары")]
   cluster жынысы     жасы достары
1        2       M 18.98200       7
2        4       F 18.80100       0
3        2       M 18.33500      69
4        2       F 18.87500       0
5        4    <NA> 18.99500      10
6        2       F 18.65586     142
7        3       F 18.93000      72
8        2       M 18.32200      17
9        2       F 19.05500      52
10       1       F 18.70800      39
11       2       F 18.54300       8
```

8-сурет. teens объектісіне cluster бағанын қосу.

Әр кластердегі оқушылардың орташа жасын қарап көрейік:

```
> aggregate(data = teens, жасы ~ cluster, mean)
  cluster    жасы
1       1 17.35298
2       2 17.26222
3       3 16.88607
4       4 17.05335
5       5 17.38101
```

Орташа жастары барлық кластерде шамамен бірдей екені көрініп тұр. Ол оқушылардың жасы қызығушылықтарына аса қатты ықпал етпейтіндігін көрсетеді. Енді жыныстарының ықпалын қарап көрейік. Оны тағы бір жаңа баған қосу арқылы жүзеге асырамыз, себебі деректер қорында жыныстары үшін де «белгісіз»(NA) ақпараттар кездеседі:

```
> teens$female <- ifelse(teens$жынысы == "F", 1, 0)
> aggregate(data = teens, female ~ cluster, mean)
cluster    female
1       0.4063492
2       0.5963546
3       0.9162996
4       0.2966870
5       0.7738516
```

Яғни, үшінші кластерде басым бөлігі қыздар екені көрініп тұр.

Қорытынды

Алынған нәтижелер R бағдарламалау тілін пайдаланып деректерді нормалауға, стандартты түрге келтіруге, кластерлерге жіктеуге, кластерлер көлемін, кластер центрлерінің сипаттамасын алуға, осы сипаттама негізінде оқушылардың қызығушылығын анықтауға, әрбір хабарламаның қай кластерге жататындығын білуге т.б. маңызды ақпараттарды табуға болатындығына көзімізді жеткізді. R бағдарламалау тілі KDnuggets.com сайтының деректерді талдауда кеңінен қолданылатындығы да аян. Деректерді талдау есебін шығаруда R бағдарламалау тілі мүмкіндіктерін меңгеру болашақ IT-мамандар үшін маңызы зор екені айқын.

Қолданылған әдебиеттер

1._Brett Lantz. Machine Learning with R. Packt Publishing, Birmingham - Mumbai, 2013

2. https://edu.kpfu.ru/mod/resource/view.php?id=110193&redirect=1

3. https://github.com/stedy/Machine-Learning-with-R-datasets/blob/master/snsdata.csv