Торайғыров университетінің хабаршысы
ҒЫЛЫМИ ЖУРНАЛЫ

НАУЧНЫЙ ЖУРНАЛ
Вестник Торайгыров университета

Торайғыров университетінің ХАБАРШЫСЫ

Энергетикалық сериясы

1997 жылдан бастап шығады

ВЕСТНИК Торайгыров университета

Энергетическая серия

Издается с 1997 года

ISSN 2710-3420

№ 1 (2021)

Павлодар

НАУЧНЫЙ ЖУРНАЛ
Вестник Торайгыров университета

Энергетическая серия
выходит 4 раза в год

СВИДЕТЕЛЬСТВО
о постановке на переучет периодического печатного издания,
информационного агентства и сетевого издания
№ 14310-Ж
выдано
Министерство информации и общественного развития
Республики Казахстан

Тематическая направленность
публикация материалов в области электроэнергетики,
электротехнологии, автоматизации, автоматизированных и
информационных систем, электромеханики и теплоэнергетики

Подписной индекс – 76136

https://doi.org/10.48081/OGVZ5983

МРНТИ 20.53.17

https://doi.org/10.48081/UOVU7003

А. Д. Кубегенова, К. Т. Искаков
Л. Н. Гумилев атындағы Еуразия ұлттық университеті,
Қазақстан Республикасы, Нұр-Сұлтан қ.

DATA MINING ТЕХНОЛОГИЯСЫНЫҢ КӨМЕГІМЕН, МЕДИЦИНАДА ИНТЕЛЛЕКТУАЛДЫ ТАЛДАУ ЖАСАУ ЖӘНЕ ӘДІСТЕРДІ ҚОЛДАНУ АСПЕКТІЛЕРІ

Мақалада деректерді зерттеу, деректерді терең талдау, білім алу, білім базасындағы мәліметтерді өңдеу жолдары, медицина саласында Data Mining технологиясының интелектуалды талдау әдістері мен қолдану аспектілері қарастырылған. АИТВ инфекциясын жұқтырған науқастардың, тобын анықталып аурулар тарихымен талдау жасалды, Модельдер әзірленіп және іс-қимыл алгоритмі құрылып (кіріс деректері), деректерді іздеу әдістері арқылы талдау жасап эксперименттер жүргізілді. Барлық аурулар сандық векторлар жиынтығы ретінде ұсынылып, кластерлерге топтастырылған деректердің бейімділігі сипатталған әдіснамаға сәйкес осы бөлу арқылы Хопкинс статистикасының мәні есептелді. Кластерлеудің өзі sklearn кітапханасының әдеттегі құралдарын қолдану арқылы жүзеге асырылды. Екі өлшемді жазықтықта көп өлшемді деректерді ұсынудың әртүрлі әдістері ұсынылды негізгі компоненттер әдісі, Кохонен желісі және т. б.

Кластерлеудің екі түрлі тәсілі қарастырылды: k – орта әдісі (Python тілінің sklearn кітапханасынан Kmeans функциясын қолдана отырып), АВТО-конфигурациясы бар тығыздыққа негізделген кластерлеу әдістері қолданылды (Python тілінің hdbscan кітапханасынан HDBSCAN функциясынан). Салыстырған жағдайда бір алгоритмнің әртүрлі параметрлерін өзгерту арқылы кластерлердің құрылымын бағалайды (мысалы, k топтарының саны); Алынған және дайындалған объектілерде модель құрылып (немесе бірнеше) және оның параметрлері реттелді. Содан кейін тестілеу және нәтижелерді талдау жүргізілді.

Кілтті сөздер: кластеризация,векторизация, корреляция, sklearn, манипуляция.

Кіріспе

Қазіргі заманғы мәліметтерді сақтау мен өңдеу әдістерінің дамуы, жинақталған, талдауды керек ететін ақпараттардың тез өсуіне алып келуде. Жинақталған мәліметтердің соншалықты көптігі оны адам күшімен өңдеуге мүмкіндік бермейді, әрі бұл өңделмеген мәліметтердің ішінде, маңызды шешімдер қабылдауға керекті ақпараттар бар екені анық. Сол себептен, мәліметтерді автоматты талдау жасау үшін Data Mining технологиясын қолдану керек болады.

Data Mining қолданбалы статистика, білімді тану, жасанды интеллект, мәліметтер базасының теориясы және т. б. сияқты ғылымдар негізінде пайда болған және дамитын көп салалы аймақ деп білеміз.

DataMining – бұл жасырын заңдылықтарды (ақпарат шаблондарын) деректерден іздеуге негізделген шешім қабылдауды қолдау процесі.

Data Mining – ті анық емес, объективті және тәжірибеде пайдалы заңдылықтардың үлкен көлемін іздеуге арналған технология ретінде сипаттауға болады:

– анық емес, өйткені табылған заңдылықтар ақпаратты өңдеудің стандартты әдістерімен немесе сараптамалық жолмен анықталмайды;

– объективті, өйткені анықталған заңдылықтар әрдайым субъективті болып табылатын сараптамалық білімнен айырмашылығы шындыққа толығымен сәйкес келеді;

– іс жүзінде пайдалы, өйткені қорытындылар практикалық қолдануға болатын нақты мағынаға ие.

Материалдар мен әдістер

Медициналық мәліметтер мұрағатында нақты аурулардың әртүрлі жағдайлары, оларды диагностикалау әдістері туралы көптеген ақпарат бар. Үлгілерді іздеу көптеген медициналық зерттеулердің міндеттерінің бірі болып табылады. Мұндай мәселелерді шешу үшін деректерді автоматты түрде талдау әдістері жиі қолданылады.

Data Mining тұжырымдамасы медицинаны қоса алғанда, әртүрлі қызмет салаларында шешім қабылдау үшін қажет деректердегі іс жүзінде пайдалы білімді анықтауға әдістерін жиынтығын белгілеу үшін пайдаланылады. Деректерді іздеу -бұл статистика, машиналық оқыту, жасанды интеллект сияқты ғылымдардың буынында пайда болған және дамитын кең сала.

Қысқаша сипаттамасын қарастырсақ. Статистика - бұл ақпарат жинайтын және оларды әрі қарай талдау және өңдеу үшін нысандарды мұқият зерттейтін ғылым. Машиналық оқытуды оқуға қабілетті алгоритмдерді құру әдістерін зерттейтін процесс ретінде сипаттауға болады. Жасанды интеллект – бұл ғылыми сала, әдістерді өндейді, интелектуалды мәселелерді электрондық компьютерде шешуге мүмкіндік береді, егерде осы мәселелерді адамдар шешіп жүрсе.

Data Mining жасанды нейрондық желілер, шешім ағаштары, корреляция, кластерлік талдау, сызықтық регрессия, байес желілері және басқалары сияқты әдістер мен алгоритмдерді біріктіреді. Жіктеу, кластерлеу, болжау сияқты міндеттер шешіледі.

Статистикалық әдістер көбінесе сызықтық регрессия теңдеулерін шешуге азаяды. Алайда, мұндай тәсілмен байланыс табу әрдайым мүмкін емес. Мұндай жағдайларда Машиналық оқыту әдістері қолданылады. Медицинада диагноз қоюға арналған көптеген сараптамалық жүйелер бар, олар әртүрлі аурулардың белгілерінің үйлесімін сипаттайтын заңдылықтар мен ережелерге негізделген.

Бұл ережелер науқастың қалай ауырғанын, қандай ем тағайындау керектігін анықтауға, тағайындалған емнің нәтижесін болжауға, әртүрлі патологиялардың себептерін зерттеуге көмектеседі. Деректерді іздеу технологиялары ережелер мен схемалар сияқты медициналық деректерді табуға мүмкіндік береді. Диагностика әдістерін әзірлеу медицинаның өзекті міндеті болып табылады, бұл өз кезегінде жіктеу міндеттеріне жатады.

Бұл талдаудың маңыздылығы – дұрыс диагнозды уақтылы қою және олардың клиникалық формаларының біріне сәйкес келетін қажетті емдеуді жүргізу.Уақытылы ем алмау денсаулықтың нашарлауына әкеледі және аурудың асқынуы мүмкін.

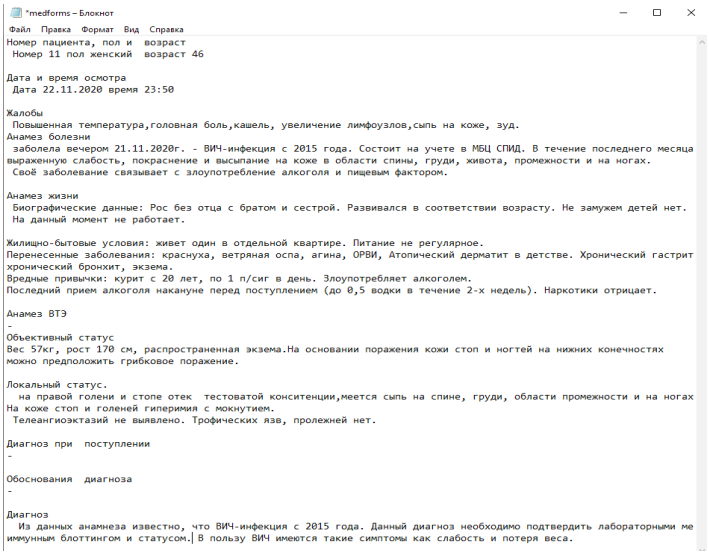
Талдау нысаны – АИТВ инфекциясын жұқтырған әртүрлі клиникалық түрлерімен ауырған науқастардың мәліметтерін жинақтау, болжамды модельді құру және деректерді іздеу әдістері арқылы талдау жасап эксперименттер жүргізу.

Талдаудың мақсаты – деректерді талдауға арналған мамандандырылған бағдарламалық қамтамасыз етудің көмегімен АИТВ-инфекциясы бар пациенттер тобын анықтау болып табылады.

Алынған нәтижелерді шешім қабылдау үшін, диагноз қою кезінде мамандар қолдана алады.Модельдерді әзірлеу барысында іс-қимыл алгоритмі құрылды және кіріс деректері енгізіледі.

Ең алдымен бастапқы деректер ретінде ем алып жатқан АИТВ инфекциясын жұқтырған, науқастардың мәліметтер тарихынан талдау жасалынды.

Әр науқас үшін міндетті талдау жасау № 1 суретте құжаттардың бірінен үзінді көрсетілген



Сурет 1 – Кіріс деректері

Көріп отырғанымыздай, құжатта құрылымданбаған жазбалар бар (айқын белгілерді қоспағанда, оларды жай шаблондарға бөліп, талдауға болады). Науқастын өз еркімен берген мәліметтері деректер қорына енгізіледі. Сонымен қатар, кейбір бөлімдер науқастын мәліметтер қорында болмауы мүмкін. Мысалы, барлық құжаттарда науқастың диагнозы мен шағымдары туралы толық мәлімет берілмеген. Егер тест нәтижелерінің блоктары параметрлерді алуға болатын бірнеше түрлі шаблондар түрінде ұсынылса, онда шағымдар мен медициналық ауру тарихы сияқты бөлімдер толығымен еркін түрде толтырылады. Сонымен қатар, кластерлік модель құруға кіріспес бұрын, бұл мәселенің ықтималды шешілуін зерттеу керек.

Векторизация

Мәтіндерді осындай манипуляциялармен орындауға болатын форматта ұсыну Python тілінің sklearn кітапханасынан TfidfVectorizer функциясын қолдану арқылы жүзеге асырылды.Бұл әдістің негізіндегі TF-IDF статистикалық өлшемі сөздің маңыздылығын бағалау үшін қолданылады. Барлық аурулар сандық векторлар жиынтығы ретінде ұсынылады, олармен әрі қарай манипуляциялар жасалуы мүмкін [4].

Кластерлеуді бастамас бұрын, бірінші кластерлерге топтастырылған деректердің бейімділігін анықтау қажет. Ол үшін Хопкинс статистикасы таңдалады.Ол іс жүзінде деректер топтастыруға бейім емес деген нөлдік гипотезаға негізделген. Оның мәнін есептеу үшін бастапқы деректер

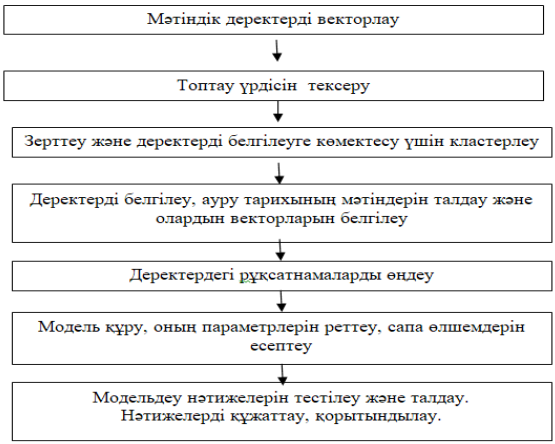
жиынымен бірдей стандартты ауытқумен бөлу негізінде кездейсоқ пайда болатын, бірнеше жалған деректер жиынтығы жасалады. Әрбір *i* бақылау үшін *n*-нен *k*-ге дейінгі орташа қашықтық есептеледі: жасанды нысандар мен олардың ең жақын нақты көршілері арасындағы нақты объектілер мен q_i арасындағы ω_i.(1)

$$H_{ind} = \frac{\Sigma_n \omega_i}{\Sigma_n q_i + \Sigma_n \omega_i} \quad (1)$$

Содан кейін Хопкинс статистикасы 0,5-тен асатын *q* ұқсас, ал топтастырылған нысандар кездейсоқ және біркелкі болып бөлінеді де нөлдік гипотезаға сәйкес келеді. Мәні H .25 ind <0 90 % сенімділік деңгейінде деректерді топтастыру тенденциясын көрсетеді. Егер бұл статистика нөлдік гипотезаның дұрыс еместігін көрсетсе және біздің кірістеріміз кластерленуге бейім болса, онда кластерлеуге көшеміз [1].

Кластерлеу алгоритмдері

Кластерлеудің екі түрлі тәсілі қарастырылды: *k* – орта әдісі (Python тілінің sklearn кітапханасынан Kmeans функциясын қолдана отырып), АВТО-конфигурациясы бар тығыздыққа негізделген кластерлеу әдістері қашықтық (Python тілінің hdbscan кітапханасынан HDBSCAN функциясын қолдану). Алынған және дайындалған объектілерде модель құру (немесе бірнеше) және оның параметрлерін реттеу қажет. Содан кейін тестілеу және нәтижелерді талдау жүргізіледі. № 2 суретте талдау жасау жұмыс реті көрсетілген.



Сурет 2 – Талдау жұмыс реті

Тапсырма бойынша жұмыстың әртүрлі кезеңдерінде визуализация қажет болып matplotlib Python кітапханасындағы pyplot модуліне негізделген әртүрлі функциялар қолданылды [2].

Атап айтатын болсақ:

1 NumPy – Python-да ғылыми есептеулер жүргізу үшін қажет іргелі кітапхана.

2 Matplotlib – екі өлшемді графиктермен жұмыс істеуге арналған кітапхана

3 Pandas – құрылымдық деректер мен уақыт қатарларын талдау құралы.

4 Scikit – learn - классикалық Машиналық оқыту алгоритмдерінің интеграторы.

5 SciPy – математика, ғылым және инженерия саласында қолданылатын кітапхана

6 Jupyter – интерактивті есептеу ортасы.

Ең алдымен, деректерді талдап белгілеу үшін кластерлеуіміз қажет. Деректер қарапайым болу үшін барлық қажет емес тыныс белгілері жойылып, Python тұрақты өрнек кітапханасы осы тапсырманы орындау үшін пайдаланылады. Тазартылған құжатты бос символдарға және жеке сөздерге бөлу құжаттарды лексемаларға бөлу әдісі болып келеді.

Мысалға:

```
def correct_known_words(story):
    dict_ = {‘сопут диагноз’: ‘сопутствующий диагноз’,
    ‘табл ‘:’таблеток ‘,
        ‘ ст. вторичных’: ‘ стадии вторичных’,
        ‘голов. боль’: ‘головная боль’
```

Алдыңғы қадамда жасалған манипуляциялардан кейін толық мәтінді түрлендірілген және тазартылған деректер шеңберін жүктеп, барлық оқиғаларды өңдеу жеткілікті болып келеді, деректер жақтауының ұяшықтарының мәндері сияқты, индекстер тарихы файлдарының атаулары болып табылады. Sklearn кітапханасынан NearestNeighbors функциясы, әдетте бақыланбайтын және басқарылатын модельдерді құру үшін қолданылады, бұл жағдайда біздің векторланған мәтінімізге ұқсас жалған объект жасауға мүмкіндік береді.

Нәтижелер мен талқылау

Жоғарыда сипатталған әдіснамаға сәйкес осы бөлу арқылы біз Хопкинс статистикасының мәнін есептей аламыз. Кластерлеудің өзі sklearn кітапханасының әдеттегі құралдарын қолдану арқылы жүзеге асырылады. Екі өлшемді жазықтықта көп өлшемді деректерді ұсынудың әртүрлі әдістері бар: негізгі компоненттер әдісі, Кохонен желісі және т. б.

PCA tSNE бұл әдістер әртүрлі принциптерге негізделгендіктен, егер біреуі кластерлер арасындағы визуалды айырмашылықтарды көрсетпесе,

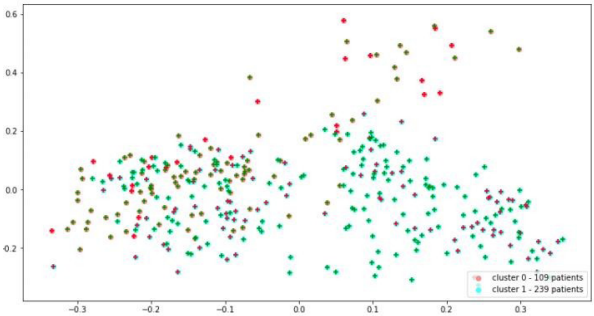
екіншісі жақсы жұмыс істеуі мүмкін. Екі әдіс де sklearn кітапханасында қолданылады. Sklearn-ді іске асыруда PCA және tSNE сіз жұмыс істеген кезде көрсеткен осьтер санына сәйкес келетін векторлар жиынтығын қайтарады.

Біздің жағдайда жазықтықпен жұмыс, бұл екі вектор болады олардың әрқайсысының ұзындығы бар, оқиға бар деректер шеңберіндегі жолдардың санына тең.Бұл векторларды нақты класс белгілерін және кластерлік нөмерлерін анықтап визуализация функциясына тапсыру керек.

Нәтижесінде график жасалынды, онда әр түрлі кластерлердегі нүктелер әртүрлі түстерге ие болады, ал егер оларға қажетті параметрлерді орнатсақ, әр түрлі түстердің крестері олардың үстінде болады.

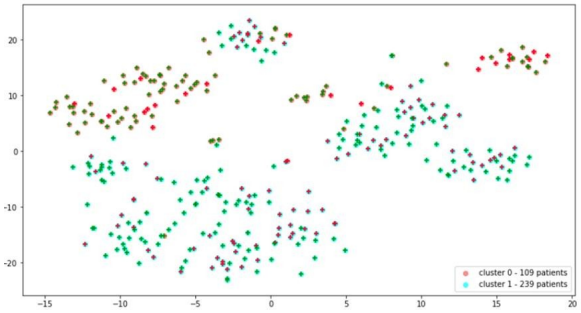
Кластерлік шешімді жасағаннан кейін оның қаншалықты тұрақты және статистикалық маңызды екендігі туралы сұрақ туындайды. Тұрақты топтау кластерлеу әдістері өзгерген кезде сақталуы тиіс: мысалы, егер иерархиялық кластерлік талдау нәтижелерінде *k*-орташа әдісін пайдалана отырып топтау кезінде 70 %-дан астам сәйкестіктер үлесі болса, онда орнықтылыққа жол беру қабылданады [4].

Салыстырмалы тексеру бір алгоритмнің әртүрлі параметрлерін өзгерту арқылы кластерлердің құрылымын бағалайды (мысалы, *k* топтарының саны); Бұл sklearn кітапханасының әдеттегі құралдарымен жасалады.

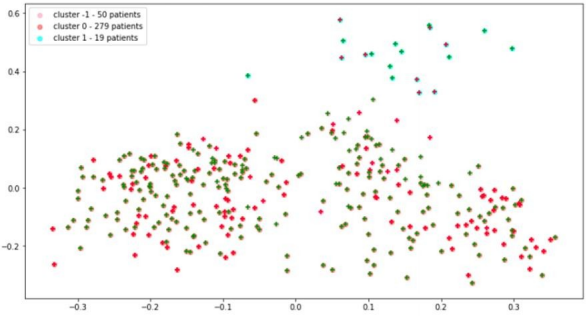


Сурет 3 – Кластерлеу мысалы

Хопкинс статистикасының құрылымданбаған мәтіндік деректердің бастапқы жиынтығы бойынша есептеулері кез-келген экспериментте H мәні 0,150-ден аспағанын көрсетті, бұл деректерде кластерлік құрылымның болуы туралы болжамдарымыздың дұрыстығын көрсетеді. Негізгі компоненттер әдісімен визуализациялау кезінде кластерлер арасында бөлу сызығын визуалды түрде салу мүмкін болмаса да, визуалды бөлінудің белгілі бір тенденциясы барын көреміз [5–6].



A)



B)

Сурет 4 – Кластерлік эксперимент

Нақты заңдылықтар анықталған жоқ: А) және В) кластерлер көптеген нүктелерде мүлдем басқа науқастарға да ұқсас, бірақ сонымен бірге бір және әртүрлі кластерлерде орналасқан. № 4 суретте көрсетілген. Деректерден ақпарат алу үшін визуализация мен бөлінуді іздеудің кластерлік эксперименттің В) әрекеті tSNE әдісін қолдану арқылы жүзеге асырылды. Бұл визуализация әдісі кластер құрылымының айқын ауырлығын көрсетеді, алайда оны көршілер әдісімен (метод соседей) анықтау мүмкін емес. Бұл экспериментте осы визуализацияға нақты бөлінген аймақтардың санына тең кластерлер санын анықтауға болады.Алайда визуалды кластерлермен сәйкестікке қол жеткізу мүмкін емес. Ол кезде кластерлеуге түбегейлі өзгеше тәсіл қолданылды – қашықтықты автоматты түрде реттейтін тығыздық негізінде. (атап айтқанда, hdbscan іске асыру).

Қорытынды

Интелектуалды талдау әдістері мен қолдану аспектілері ретінде Data Mining технологиясын медицина саласында қолдану тиімді болып келеді. АИТВ инфекциясын жұқтырған науқастардың топтарын

анықтаудағы міндеті, іске асырудың барлық негізгі сәттері жазылды, TF-IDF статистикалық өлшемін бағалау үшін Python тілінің sklearn кітапханасынан Tfidf Vectorizer функциясы қолданылып эксперементте қарастырылды. Хопкинс статистикасы тандалып, топтастырылған деректердің бейімділігі анықталып, науқастын ауру тарихы бойынша кластерлеу жүргізілді. Кластерлеудің екі түрі қарастырылды. Дайындалған объектілерден әдіснамаға сәйкес модельдер құрылып график жасалынды.

Нәтижесінде негізгі компоненттер әдісімен визуалды бөлінудің белгілі бір тенденциясы бары анықталды. Барлық тұжырымдар мен алынған өлшемдер жазылып, талданды.

Жұмыста пайдаланылған және сипатталған барлық шешімдер ауқымды және де эксперименттерде қолдануға мүмкіндік беретін немесе басқа аурулары бар, бірақ ұқсас болып келетін науқастар тобын іздеп жүзеге асыру үшін қолданылады.

Пайдаланған деректер тізімі

1 **Лисинин, А. В.,Файзулин, Р. Т**. Применение метаэвристических алгоритмов к решению задач кластеризации методом k-средних [Текст] // Компьютерная оптика. – 2015. – Т. 39, №. 3. – С. 406–412.

2 **Андреас Мюллер** Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными [Текст] // Мюллер Андреас. – М. : Альфа-книга, 2017.

3 **Нейский, И. М.** Классификация и сравнение методов Кластеризации. [Текст] // Интеллектуальные технологии и системы. Сборник учебно-методических работ и статей аспирантов и студентов. – М. : НОК «CLAIM», 2006. – Выпуск 8. – С. 130–142.

4 **Барсегян, А. А., Куприянов, М. С., Степаненко, В. В., Холод, И. И.** Методы и модели анализа данных [Текст] // OLAP и DataMining. СПб. : БХВ-Петрбург, 2008. – 336 с.

5 **Петрунин, Ю. Ю.** Информационные технологии анализа данных. [Текст] // Анализ данных. – М. : КДУ, 2010.

6 **Плас, Джейк Вандер** Python для сложных задач. Наука о данных и машинное обучение. Руководство. // Плас Джейк Вандер. – М. : Питер, 2018.

7 **Charikar, M. E**. Incremental clustering and dynamic information retrieval. [Текст] // SIAM Journal on Computing. – 2004. – Vol. 33, №. 6. – P. 1417–1440.

8 **Мокина, Е. Е., Марухина, О. В., Шагарова, М. Д., Дубинина, И. А**. Использование методов Data Mining при принятии медицинских диагностических решений. [Текст] // фундаментальные исследования. – 2016. – № 5–2. – с. 269–274.

9 **Бююль, А., Цеффель, П.** SPSS: Искусство обработки информации. [Текст] – М., 2005 • Глава 20. Кластерный анализ

References

1 **Lisinin, A. V., Fayzulin, R. T.** Primeneniye metaevristicheskikh algoritmov k resheniyu zadach klasterizatsii metodom k-srednikh [Application of metaheuristic algorithms to solving clustering problems using the k-means method] [Text] – Komp'yuternaya optika. – 2015. – T. 39, №. 3. – P. 406–412.

2 **Andreas Myuller.** Vvedenie v mashinnoe obuchenie s pomoshh'yu Python. Rukovodstvo dlya specialistov po rabote s danny'mi. [An introduction to machine learning with Python. A guide for data scientists] [Text] – Moscow : Al'fa-kniga, 2017.

3 **Neyskiy, I. M.** Klassifikatsiya i sravneniye metodov Klasterizatsii. [Classification and comparison of methods Clustering.] [Text]. Intellektual'nyye tekhnologii i sistemy. Sbornik uchebno-metodicheskikh rabot i statey aspirantov i studentov – Moscow : NOK «CLAIM», 2006. – Vypusk 8. – P. 130–142.

4 **Barsegyan, A. A., Kupriyanov, M. S., Stepanenko, V. V., Kholod, I. I.** Metody i modeli analiza dannykh. [Data Analysis Methods and Models] [Text]. OLAP i Data Mining. SPb : BKHV-Peterburg, 2008. – 336 p.

5 **Petrunin, YU. YU.** Informatsionnyye tekhnologii analiza dannykh. [Information technology data analysis] [Text]. Analiz dannykh. – Moscow : KDU, 2010.

6 **Plas, Dzheyk Vander,** Python dlya slozhnykh zadach. Nauka o dannykh i mashinnoye obucheniye. Rukovodstvo. [Python for complex tasks. Data Science and Machine Learning. Leadership.] [Text] – Moscow : Peter, 2018.

7 **Charikar, M. E**. Incremental clustering and dynamic information retrieval. [Текст] // SIAM Journal on Computing. – 2004. – Vol. 33, №. 6. – P. 1417–1440.

8 **Mokina, E. E., Maruxina, O. V.,Shagarova, M. D.,Dubinina, I. A.** Ispol'zovaniye metodov Data Mining pri prinyatii meditsinskikh diagnosticheskikh resheniy. [Using Data Mining Methods in Making Medical Diagnostic Decisions] [Text]. Fundamental'nyye issledovaniya. – 2016. – № 5–2. – P. 269–274.

9 **Byuyul', A., Tsefel', P.** SPSS: Iskusstvo obrabotki informatsii. [The art of information processing.] [Text]. Klasternyy analiz – Moscow, 2005. Glava 20.

Материал 19.03.21 баспаға түсті.

А. Д. Кубегенова, К. Т. Искаков

Аспекты интеллектуального анализа и применения методов в медицине с помощью технологии data mining

Евразийский национальный университет имени Л. Н. Гумилева,
Республика Казахстан, г. Нур-Султан.
Материал поступил в редакцию 19.03.21.

A. D. Kubegenova, K. T. Iskakov

Aspects of intellectual analysis and application of methods in medicine using data mining technology

L. N. Gumilyov Eurasian National University,
Republic of Kazakhstan, Nur-Sultan.
Material received on 19.03.21.

В статье рассматриваются аспекты исследования данных, глубокого анализа данных, получения знаний, способов обработки данных в базе знаний, методов интеллектуального анализа и применения технологии Data Mining в области медицины. Выявлена группа ВИЧ-инфицированных больных, проведен анализ с историей заболеваний, разработаны модели и разработан алгоритм действий (входные данные), проведен анализ и эксперименты с методами поиска данных. Все болезни были представлены в виде набора числовых векторов и были сгруппированы в кластеры, согласно описанной методологии, с помощью этого распределения было рассчитано значение статистики Хопкинса. Сама кластеризация осуществлялась с использованием обычных инструментов библиотеки sklearn. Предложены различные методы представления многомерных данных в двумерной плоскости метод основных компонентов, линия Кохонена и др.

Были рассмотрены два разных способа кластеризации: метод k-Medium (с использованием функции Kmeans из библиотеки sklearn языка Python), методы кластеризации на основе плотности с автоконфигурацией (из функции HDBSCAN из библиотеки Hdbscan языка Python). В случае сравнения оценивают структуру кластеров путем изменения различных параметров одного алгоритма (например, количество групп k); на полученных и подготовленных объектах строится модель (или несколько) и корректируются ее параметры. Затем было проведено тестирование и анализ результатов.

Ключевые слова: кластеризация, векторизация, корреляция, sklearn, манипуляция.

The article discusses aspects of data research, deep data analysis, knowledge acquisition, methods of data processing in the knowledge base, methods of data mining and application of Data Mining technology in the field of medicine. A group of HIV-infected patients was identified, an analysis with a history of diseases was carried out, models were developed and an algorithm of actions (input data) was developed, analysis and experiments with data search methods were carried out. All diseases were represented as a set of numerical vectors and were grouped into clusters, according to the described methodology, using this distribution, the value of Hopkins statistics was calculated. Clustering itself was performed using the usual tools of the sklearn library. Various methods for representing multidimensional data in a two-dimensional plane are proposed: the principal component method, the Kohonen line, etc.

Two different clustering methods were considered: the k-Medium method (using the Kmeans function from the Python sklearn library), and density-based clustering methods with autoconfiguration (from the HDBSCAN function from the Python Hdbscan library). In the case of comparison, the cluster structure is evaluated by changing various parameters of one algorithm (for example, the number of groups k); a model (or several) is built on the obtained and prepared objects and its parameters are adjusted. Then testing and analysis of the results were carried out.

Keywords: clustering, vectorization, correlation, sklearn, manipulation.

Теруге 19.03.2021 ж. жіберілді. Басуға 29.03.2021 ж. қол қойылды.
Электрондық баспа
17,4 Mb RAM
Шартты баспа табағы 21,0. Таралымы 300 дана. Бағасы келісім бойынша.
Компьютерде беттеген: А. К. Шукурбаева
Корректор: А. Р. Омарова
Тапсырыс № 3746

Сдано в набор 19.03.2021 г. Подписано в печать 29.03.2021 г.
Электронное издание
17,4 Mb RAM
Усл. печ. л. 21,0. Тираж 300 экз. Цена договорная.
Компьютерная верстка: А. К. Шукурбаева
Корректор: А. Р. Омарова
Заказ № 3746

«Toraighyrov University» баспасынан басылып шығарылған
«Торайғыров университет»
коммерциялық емес акционерлік қоғамы
140008, Павлодар қ., Ломов к., 64, 137 каб.

«Toraighyrov University» баспасы
«Торайғыров университет»
коммерциялық емес акционерлік қоғамы
140008, Павлодар қ., Ломов к., 64, 137 каб.
8 (7182) 67-36-69
E-mail: kereku@tou.edu.kz
www.vestnik.tou.edu.kz