

Международный научно-популярный журнал

ISSN 2073-333X

Наука и жизнь Казахстана Қазақстанның ғылымы мен өмірі

№ 3(58) 2018



Оңтүстік Қазақстан облысы,
Мақтаарал ауданы, Жетісай қаласындағы
«Сырдария» университетіне - 20 жыл

Құрылтайшы:
«ҚҰҚЫҚТЫҚ МИССИЯ» ҚОҒАМДЫҚ ҚОРЫ
ХАЛЫҚАРАЛЫҚ ҚАЗАҚСТАН КРИМИНОЛОГИЯЛЫҚ КЛУБЫ

Учредитель:
ОБЩЕСТВЕННЫЙ ФОНД «ПРАВОВАЯ МИССИЯ»
МЕЖДУНАРОДНЫЙ КАЗАХСТАНСКИЙ КРИМИНОЛОГИЧЕСКИЙ КЛУБ

Founder:
PUBLIC FOUNDATION «LEGAL MISSION»
INTERNATIONAL KAZAKHSTAN CRIMINOLOGY CLUB

ҚАЗАҚСТАННЫҢ ҒЫЛЫМЫ МЕН ӨМІРІ
НАУКА И ЖИЗНЬ КАЗАХСТАНА
SCIENCE AND LIFE OF KAZAKHSTAN
Халықаралық ғылыми-көпшілік журнал
Международный научно-популярный журнал
International popular-science journal

№3(58) 2018

ӨНЕРТАНУ - ЭКОНОМИКА
ТЕХНИКА

Астана 2018

Тюлепбердинова Г.А.

ф.м.ғ.к., доцент м.а.; Өл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы қаласы
tyulepberdinova@mail.ru

Адилжанова С.А.

аға оқытушы; Өл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы қаласы

Ғазиз Г.Г.

аға оқытушы; Өл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы қаласы

Гусманова Ф.Р.

ф.м.ғ.к., доцент; Өл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы қаласы

Жумартов М.А.

аға оқытушы; Өл-Фараби атындағы Қазақ Ұлттық Университеті, Алматы қаласы

ДЕРЕКТЕРДІ ТАЛДАУДАҒЫ ТЕХНОЛОГИЯЛАР МЕН ӘДІСТЕР

Түйін: Бұл мақалада білім – адамзатпен жинақталған деректер, түсінік, ережелер, әдістер және т.б. мен принциптер жиыны ретінде немесе ойлау қызметінің нәтижелер жиыны ретінде немесе хабардар жағдай ретінде анықталатын ақпарат қарастырылады. Білім термині интелект ұғымымен тығыз байланысты. Әдістемелер деректер қорында әртүрлі есептерді пайдалануға мүмкіндік береді, зерттелетін проблемалардың шешімдер жиынын болжау үшін жіктеу және кластеризациялау есептері пайдаланылады. Сонымен қатар, олар әдістер дұрыс жорамал жиілігі мен болжаудың дәлелділігі тұрғысынан болжанған шешімдерге арналған сенімді интервалдарды бағалауға мүмкіндік береді. Деректерді талдау әртүрлі айналымның үлкен көлемін, статикалық және ақпараттық талдауды, шешімдер ағашын, сәйкестік бағасын, нейрондық желіні, генетикалық алгоритмді, айқын емес логика әдістерін және т.б. қамтиды. Мақалада осындай әдістерді қарастыратын боламыз.

Кілт сөздер: интелект, зияткерлер, нейрон, үлкен деректер қоры, Data Mining, деректер қорында білімді анықтау, болжайтын модель, сипаттайтын модель, жіктеу, регрессия, уақытша қатарлар талдауы, кластерлеу, модельдің жинақтылығы немесе жалтыламалылығы, ассоциация ережесі.

Резюме: В данной статье рассматриваются вопросы знания, накопленных человечеством данных, понятие, правила, методы и т.д. результаты деятельности и мышления, также рассматриваются принципы определяемые в качестве информации. Термин образования тесно связана с понятием интеллекта. Использование различных методик позволяет дать итоги в базе данных, используя отчеты для решения проблем, прогнозирования и классификации, исследуемых кластеризации схода. Кроме того, они надежно с точки зрения прогноза и методов интервалов для решения будущее и позволяет оценить обоснованность правильного предположения. Анализ оборота больших объемов различных данных включает в себя информационный и статический анализ, дерево решений, оценку соответствия, нейронных сетей, генетических алгоритмов, методов и логики, и т.д. Данной статье будем рассматривать такие методы.

Ключевые слова: интелект, нейрон, большая база данных, Data Mining, выявление знаний в базах данных, модель, предполагающая, характеризующая модель, классификация, регрессия, анализ временных рядов, кластеризация, комплектность, правила ассоциации.

Summary: This article discusses the issues of knowledge accumulated by mankind data, the concept, rules, methods, etc., the results of activity and thinking, also discusses the principles defined as information. The term education is closely related to the concept of intelligence. The use of different methodologies allows us to give the results in the database using reports for problem-solving, prediction and classification, the investigated clustering of the gathering. In addition, they are reliable from the point of view of prognosis and methods intervals for the solution of the future and allows us to assess the validity of the correct assumptions. The analysis of the turnover of large amounts of various data includes information and static analysis, decision tree, evaluation of conformity, neural networks, genetic algorithms, methods and logic, etc. This article will consider these methods.

Keywords: intelligence, neuron, large database, Data Mining, detection of knowledge in databases model, which assumes a characterizing model, classification, regression, time series analysis, clustering, completeness and rules of the Association.

Әдістемелер (техникалар, әдістер, алгоритмдер) ДМ-да әртүрлі есептерді пайдалануға мүмкіндік береді, зерттелетін проблемалардың шешімдер жиынын болжау үшін жіктеу және кластеризациялау есептері пайдаланылады. Сонымен қатар, олар (әдістер) дұрыс жорамал жиілігі мен болжаудың дәлелділігі тұрғысынан болжанған шешімдерге арналған сенімді интервалдарды бағалауға мүмкіндік береді. Деректерді талдау әртүрлі айналымның үлкен көлемін, статикалық және ақпараттық

талдауды, шешімдер ағашын, сәйкестік бағасын, нейрондық желіні, генетикалық алгоритмді, айқын емес логика әдістерін және т.б. қамтиды. Төменде кейбір әдістерді қарастырамыз [1].

Статистика (Statistics) математиканың үлкен бөлімдерінің бірі, әртүрлі әдіс-тәсілдерді пайдалану арқылы сандық деректерді талдау және жинақтаумен айналысады. Төменде DM-де жиі қолданылатын кейбір статистикалық әдістер көрсетілген:

- *Нүктелік бағалау* (Point estimation) деректерді орта мән, дисперсия, мода, медиана сияқты іріктеудегі бір мәннің статистикалық әдісін пайдалана отырып есептеу үрдісі жүргізіледі.

- *Мәндерді жинақтау* (Data summarization) үлкен деректер банкісіндегі әрбір элементті зерттеп талдамай-ақ, қысқаша ақпаратты ұсыну үрдісі. Дербес жағдайда, әртүрлі статистикалық әдістерді (атап айтқанда, орта мәнін есептеу, медиана және т.б.) пайдаланып гистограммалар құру, мәндерді жинақтаудың өте тиімді тәсілі болып табылады.

- *Байес әдісі* (Bayesian techniques) статистика және ықтималдылық аппараттарын (әдістерін) пайдаланып, деректер банкіндегі дерек элементтерін жіктеу әдісі. Байес әдістерінің ең қарапайымы Байес теоремасын пайдалану қазіргі таңдағы DM әдістерінің әртүрлі ықтималдықтарының дамуына әкеледі.

- *Гипотезаны тексеру* (Testing a hypothesis) іріктелген деректерді қолдана отырып гипотезаны мақұлдау немесе теріске шығару әдістемесі.

- *Корреляция* (Correlation) деректер үлестірілімінің нүктелік бағаларын пайдаланып екі немесе оданда көп әртүрлі сандық деректер жиынтықтары арасындағы байланыстың тығыздығын анықтау әдісі.

- *Регрессия* (Regression): деректер жиынтығындағы айнымалының басқа айнымалыға тәуелділігі арқылы сол айнымалының мәнін бағалау әдісі, сондай-ақ, бұл тәуелділік үздіксіз функция арқылы сипатталады.

Шешімді қабылдау ағаштары (Decision trees) – болжам және жіктеу моделі ретінде жиі пайдаланылатын ағаш тәрізді графтар. Шешімдер ағашында әрбір бұтақ жіктеуге қатысты сұрақты, ал әрбір жапырағы кластардың біреуін білдіреді. Түйіндері деректер атрибуттарының атауын қамтитын шешімдер ағашының бұтағын бойлай қарасақ объектілерді жіктеу туралы сұраққа жауап алуға болады. Шешімдер ағашы әдетте кластеризациялаумен және жіктеумен байланысты тапсырмаларға пайдаланылады. Олар сонымен қатар қабылданатын шешімдерді таңдауға пайдаланылуы мүмкін логикалық ережелерді құруға көмектеседі [2].

Машиналық оқыту (Machine Learning) - деректерді өз еркінше жинауға және жалпылауға қабілетті, содан кейін пайдалы білімдерді түрлендіретін компьютерлік жүйелерді құру үрдісі. Машиналық оқытудың негізгі мақсаты компьютерлік жүйені үйрету арқылы өз тәжірибесінде үйреніп, бақылау нәтижелерін талдай алатын адам сияқты әрекет етуге дайындау. Осылайша жүйе өзін-өзі үздіксіз жетілдіріп, өнімділігі мен тиімділігін арттырып отыруы тиіс. Бұл әдістерді DM-да қолданудың нәтижесі машиналық оқытудың жаңа құрылымдарын және деректер жиынтығында бұрын белгісіз болған заңдылықтарды ашуға септігін тигізеді.

Жасырын Марковтық модельдер (Hidden Markov models) – уақытша қатарларда болуы мүмкін болашақтағы іс-әрекеттерді болжауға мүмкіндік беретін модельдер. Модельдер іс-әрекеттердің қазіргі және бұрынғы деректеріне сүйене отырып олардың болашақта болу ықтималдығын бағалауға жол ашады. Жай сөзбен айтқанда Марков модельдерінің нәтижесі тікелей уақытпен байланысты болатын уақыт модельдері болып табылады. Марковтық жасырын модельдердің шектеуінің бірі болашақ оқиғалардың өзгерісі толығымен қазіргі және бұрынғы оқиғаларға негізделген (яғни, қазіргі және зерттеліп отырған келер шақ арасында болатын әрекеттердің әсері ескерілмейді).

Жасанды нейронды желілер (Artificial Neural Networks) биологиялық-нейрондық желі құрылымына жуықталған түрде ұқсайтын оқу және жаттығу тренингіне пайдаланылатын сызықты емес болжамдық модель. Жасанды нейрондық желілерде болашақ жағдайларды болжау мақсатымен бұрын белгілі болған үлкен көлемді деректерді меңгеру үшін талданады және пайдаланылады. Олар кең шоғырлы қосымшаларда пайдаланылады. Атап айтқанда, кредит карточкаларымен онлайн транзиттік транзакциялық үрдістердегі алаяқтарды анықтау, әскери операциялардың нәтижелілігінің ықтималдылығын бағалау үшін, сондай-ақ биологиялық үрдістерді симуляциялауды қажет ететін зерттеулерде пайдаланылады және т.б.

Генетикалық алгоритмдер (genetics algorithms) аргументтері сандық тізбектерден тұратын, көпөлшемді функцияның экстремумдарын табуға мүмкіндік беретін іздеу алгоритмі (әдетте екілік кодта жазылады). Ол үшін жасанды мутация, будандастыру және селекция операциялары қолданылады. Генетикалық алгоритмдері кластеризациялау үшін және ассоциация ережелері үшін пайдаланылады. Егер нақты бір деректер жиынтығы бар болса, онда генетикалық алгоритм іріктелген

деректерге сәйкес келетін көптеген модельдер жиынынан ең үздігін таңдап алады. Іріктелген деректерді беру үшін генетикалық алгоритмді қолданудың алдында ерікті түрде модель таңдалып алынады. Сосын жаңа модельді жасау үшін таңдалған модельге қатысты жоғарыда айтылған операциялар орындалады. Бұлардың ішінен белгілі бір критерийге байланысты мейілінше сәйкес модель таңдалып алынады [3].

Талдау және басқару құралы ретінде деректерді талдаудың қолданылуына мысалдар. Қазіргі уақытта DM біздің күнделікті өміріміздегі әртүрлі салаларда, атап айтсақ күнделікті көптеген деректерді түрлендіретін шағын, орта және ірі мекемелерде пайдаланылады. Негізгі мақсаты сәйкес қызметті талдау мен басқаруда DM қолданылуына мысалдар келтіру.

- *Бизнес.*

Бизнес-талдау және басқару: DM мекемелердің қызметін жақсартуға, бизнес жүргізу құнын төмендетуге, табысты молайтуға және тұтынушыларға қызмет көрсету сапасын жоғарылатуға бағытталған, бизнес құруға көметеседі.

Электронды (және дәстүрлі) сауда.

Сатуды, клиенттер топтарын, өнімдерді, уақытты және аймақты көп өлшемді талдауда DM кросс-сатуды ұйымдастыруда, тұтынушылардың ұқсас топтарын анықтауда және тұтынушыларды қызықтырып тарту функцияларын атқаратын әртүрлі шешім қабылдауды қолдау жүйелерін ұсынады. Осылайша DM маркетинг мақсатында, яғни тұтынушылармен өзара жақсы қарым-қатынаста болу, әрі ұсынылатын қызмет деңгейін жоғарылату және ең негізгісі клиенттермен ұзақ мерзімді қатынас орнау үшін клиенттерді топтастыру және жіктеуге мүмкіндік береді. Сауда-саттық компанияларының тиімділігін талдау арқылы үнемі сатып алынатын тауарлар жиынтығынан кластерлер қалыптастыруға септігін тигізеді.

- *Компьютерлік қауіпсіздік.*

Деректерді интеллектуалды талдау компьютерлік қауіпсіздік мамандары мен желі әкімшілікке компьютерлер мен желілердің қауіпсіздігіне қатер төндіретін алаяқтық іс-әрекеттерді, теріс пайдалануларды, рұқсатсыз өзгертулерді анықтауда аналитикалық әдістердің мүмкіндіктерін біріктіруге жағдай туғызады.

- *Банктік және қаржылық өңдеу.*

Кредиттерді қайтару және клиенттерге кредит беру шешімін болжауда, клиенттің кредиттік рейтингін анықтау және кредиттік төлем сенімділігіне әсер ететін күшті немесе әлсіз болған көп факторлы кредиттік саясатты талдауда, сонымен қатар «ақшаның ағылуы» және басқада қаржылық қылмыстарды байқауда DM әдістері қолданылады.

- *Биоинформатика, Медицина, Денсаулық сақтау.*

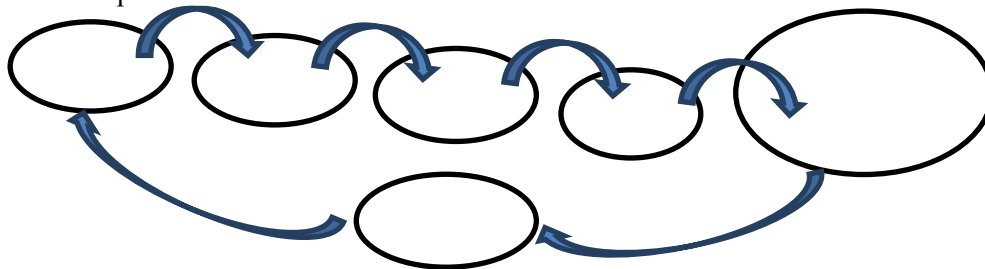
DM молекулярлы-генетикалық және гендік инженерия (гендік зерттеулер, адамның дезоксирибонуклеин қышқылы (ДНК) тізбектерін зерттеуде, емделу тиімділігін жоғарылату және әртүрлі ауруларға байланысты гендерді анықтау) салаларында қолданылады. Денсаулық сақтау және фармацевтикалық мекемелер өздерінің клиникалық және диагностикалық зерттеулерінде үлкен көлемдегі деректер жинайды. Мұнда DM әдістері медициналық ЭЖ-лерді құру құралы ретінде пайдаланылады. Деректерді интеллектуалды талдау осындай мекемелерге машиналық үйрену техникасын пайдалану және фармацевтикалық және медициналық талдаулар арқылы жаңа дәрілер жасауға керекті ақпараттарды алу мүмкіндігін тудырады [4].

- *Жаңалықтар және ойын-сауық*

Медиа жаңалықтар мен ойын-сауық индустриясы алуан түрлі демографиядағы адамдардың көруі мен оқуы үшін мәтіндер, графикалар, аудио-видео жазбалар түріндегі үлкен көлемдегі деректерді шығарып отырады. Деректерді интеллектуалды талдау (DM) әртүрлі коллекциядағы деректерді талдауға және зерттеуге мүмкіндік беру арқылы болашақтағы оқырмандар немесе көрермендер аудиториясын өздеріне қызықтырып жүзеге асырады. Интернеттен адамзат қызметінің ғылыми және инженерлік деректерді іздеу, телекоммуникация, инвестиция, сақтандыру, қылмыстарды зерттеу сияқты салаларында да DM-ды қолданудың көптеген мысалдарын кездестіруге болады. Деректерді талдау үрдістерінің модельдері мен сұлбаларына тоқталайық. Ұйымдар мен кәсіпорындарда қолданылатын DM тәжірибелік қосымшалар саласында жұмыс істейтін зерттеушілер үлкен көлемді деректерден (мысалы, үлкен ДҚ-нан) қажетті деректі алу үшін жүйелілік тәсілді пайдаланады. Төменде DM үрдістерінің бірнеше белгілі модельдері көрсетіледі.

1. *«5A» үрдісінің моделі* (SPSS Inc-пен ұсынылған және пайдаланылған, Чикаго, АҚШ). SPSS Inc. Бұл модельді DM-ға дайындық қадамы ретінде әзірледі және пайдаланды. Одан кейін жиналған ақпараттардың талдаулары үшін CRISP_DM деген атаумен белгілі үрдістің кеңейтілген моделі пайдаланылады. «5A» моделі проблемамен (Asses) байланысты деректерге қатынауға және

толықтыруға қатысты есептің анықталуы мен қойылуынан бастайды. Бұл сатыдан кейін деректердің талдануы және Asses (Analyze) әртүрлі әдістерін пайдалану арқылы қажетті деректі алу мен мәселені шешуге, нәтижелерді жүзеге асыру (Act) жүргізіледі. Зерттеу соңында, сәйкес ПҚ-ды құру арқылы оқу үрдісін автоматтандыру мүмкіндігі бағаланады (Automate). «5A» үрдісі моделінің кадамдарының сұлбасын 1-суретте келтірілген.

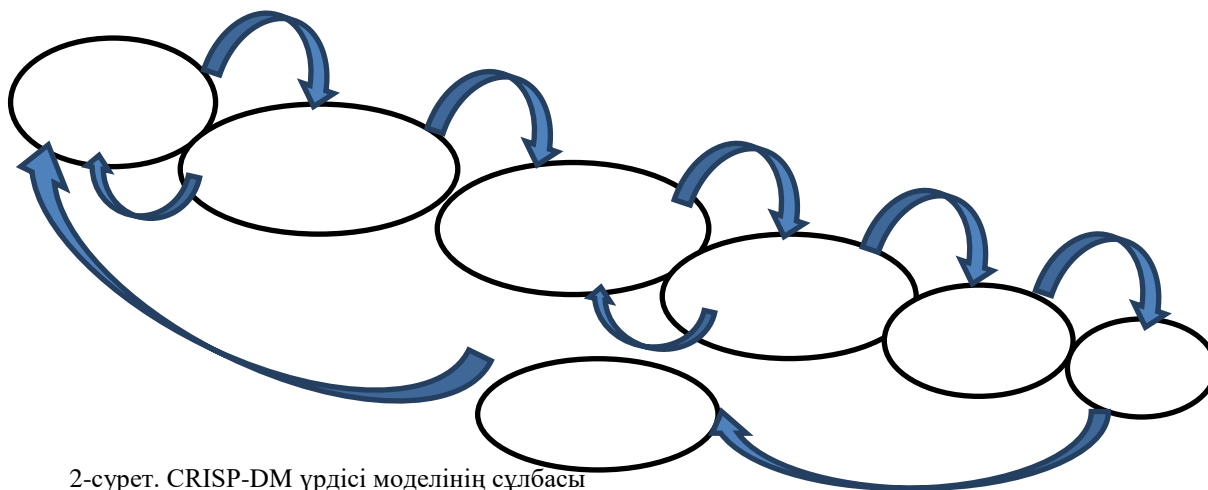


1-сурет. «5A» үрдісінің моделінің сұлбасы

2. *CRISP-DM*. (Cross-Industry Standard Process for Data Mining) – халықаралық провайдерлер тобымен ұсынылған салааралық деректерді талдаудың стандартты үрдісі (Дания, Германия, Ұлыбритания, Нидерланды).

CRISP-DM үрдіс моделі төмендегі кезеңдерден тұрады (2-сурет).

- DM бизнес-мәселелері үшін, қажетті талаптардың бастапқы жиыны мен мақсатын түсіну;
- бірінші кезеңде жиналған деректерді талдау, есепті шешу моделін ұсыну үшін сәйкес шаблондарды анықтау;
- модельдеудің түрлі әдістері мен техникаларына кіріс дерек ретінде пайдаланылатын таңдалған модель үшін деректердің толық жиынтығын құру;
- модельдеу мен талдаудың алдыңғы фазасындағы ДҚ-н пайдалану арқылы деректерді талдаудың әртүрлі әдістерін таңдау және қолдану;
- нақтыланған деректерді талдау үшін алдыңғы кезеңдерде құрылған модельді бағалау;
- енгізу: соңғы пайдаланушыларға алынған білімді ұйымдастыру және жеткізу.



2-сурет. CRISP-DM үрдісі моделінің сұлбасы

Үрдістің басқа екі моделін қысқаша қарастырамыз.

3. *SEMMA* (Sample-Explore-Modify-Model-Analysis). SAS Institute үрдісінің моделі. (Деректерді жинақтау-Зерттеу-Нақтылау-Модель-Талдау):

- проблеманы түсіну.
- деректерді жинақтау (таңдау).
- тенденцияларды, ерекшеленіп тұратын шамаларды, т.б. іздеу үшін түрлі статистикалық техникаларды пайдалану арқылы таңдалған деректердің бөлігін зерттеп нақтылау;
- деректерді жетілдіру – кластеризациялау, ассоциациялар, визуализация, түрлендіру;
- деректердің моделін құру және оны бағалау.

4. «6 σ» *Алты сигма*. Шығарылатын өнімнің сапасын арттыратын үрдістің моделі (Motorolla, GE, АҚШ):

- өнімді шығарудың негізгі кезеңдерін анықтау;

- сапаны тексеру әдістерін анықтау;
- өнім сапасын бағалау тұрғысынан технологиялық мәселелердің талданылуы;
- технологияларды жақсарту;
- бақылау.

Әдебиеттер тізімі

1. Газуль С.М., Ананченко И.В., Кияев В.И. Совершенствование образовательного процесса в вузе: активные методы обучения и гибридные информационные системы на основе виртуализации // *Современные проблемы науки и образования*. – 2015. – № 2; URL: www.science-education.ru/122-20856 (дата обращения: 16.08.2015).
2. Тюлепбердинова Г.А., Унайбаева Р.К., Шанляякова А.С. Процесс компьютеризации и информатизации общества // ТРУДЫ Международной научно-практической конференции «Информационные и телекоммуникационные технологии: образование, наука, практика», Посвященной 50-летию Института информационных и телекоммуникационных технологий Алматы, Казахстан, 5-6 декабрь, 2012 стр. 204-207.
3. Б.А. Урмашев, Ф.Р. Гусманова, Г.Г. Газиз, Г.А. Тюлепбердинова, М.Ж. Сақыпбекова, А. Алтыбай. Ақпараттық-коммуникациялық технологиялар: Оқу құралы. Алматы, КазНУ, 2017.
4. Сейтбекова Г.О., Тюлепбердинова Г.А. Ақпараттық технологиялар: оқу құралы / Г.О.Сейтбекова, Г.А.Тюлепбердинова - Алматы: «Эверо» баспасы, 2015. - 256 б.