Дәріс №9

Зияткерлік деректерді талдау

Қарастырылатын сұрақтар:

1. Зияткерлік деректерді талдаудың мақсаты мен міндеттері
2. Data Mining технологиясы

1. Зияткерлік деректерді талдаудың мақсаты мен міндеттері

Деректерді талдаудың мақсаты (ағылш. Datamining, аударманың басқа нұсқалары - "деректерді өндіру"," деректерді қазу") - бұл мәліметтер жиынтығындағы анық емес заңдылықтарды анықтау. Деректерді талдаудың заманауи технологиялары – Джон Тьюки ұсынған принциптерге негізделген. Ол деректерді зерттеу үдерісінің жаңа парадигмасын ұсынды:

- Талдау - бұл деректердің болу тәсілі. Оның материалдық негізі – «адам-машина» жүйесі.
- Сол деректерге бірнеше рет оралу принципі.
- Ықтимал модельдердің көптігі принципі.
- Бейресми шешім қабылдау рәсімдеріне негізделген нәтижелердің көптігі және таңдау принципі.
- Эндогендік ақпаратты толық пайдалану және экзогендік ақпаратты максималды есепке алу принципі.

Кейбір жағдайларда ДТ жасанды интеллект технологияларына сәйкес құрылады және жүзеге асырылады.

Жасанды интеллект (AI, artificial intelligence) - бұл «есептеу машинасының әдетте адамның ақыл-ойымен байланысты функцияларды орындау арқылы ойлау процесін модельдеу қабілетін» сипаттайтын жалпы түсінік: сараптамалық жүйелерді құру және пайдалану, логикалық қорытынды, табиғи тілдерді түсіну, көру және есту қабілеті.

Сараптама жүйесі (СЖ, expert system) - бұл жасанды жүйе

ережелер жиынтығы бар білім базасын және пайдаланушы ұсынған фактілер негізінде жағдайды тануға, шешім тұжырымдауға немесе ұсыныс беруге мүмкіндік беретін шығару машинасын (inference engine) қамтитын интеллект. Әдетте СЖ қосымша пайдаланушының жұмыс интерфейсін қамтиды, ол арқылы сарапшы компьютермен өзара әрекеттеседі.

Осылайша, СЖ - бұл сарапшының шешім қабылдау қабілетін еліктейтін компьютерлік жүйе.

Деректерді зияткерлік талдау (ДЗТ) – жасанды интеллект әдістерін пайдаланатын және жүйеге жасанды интеллект қасиеттерін беруге бағдарланған деректерді зерттеу. Есептеу техникасы, ең алдымен, деректерді өңдеу үшін жасалды. Олар деректерді талдаудың күнделікті бөлігін шешімдерді қолдау жүйелеріне (SPPR, DSS) – тиімді басқару шешімін табу үшін белгілі бір пәндік аймақтан деректерді Енгізу, сақтау және талдау құралдары бар жүйелерге ауыстыруға тырысады.

Мұндай жүйелер дұрыс шешімдер жасамайды, бірақ талдаушы маманға зерттеу мен талдауға ыңғайлы түрде мәліметтер береді. Интеллектуалды SPPR-де AI әдістеріне негізделген функциялар бар. Олардың басты айырмашылығы - өзін-өзі дамыту қабілеті, ол бастапқы алгоритмдер мен бағдарламада қарастырылмаған сапалы жаңа шешімдерді құруда көрінеді.

<u>Қорытындыларды болжауға, талдауға және ұсынуға арналған құралдар</u>

Интеллектуалды талдау тәжірибесі жоқ пайдаланушылар келесі тапсырмалар үшін қарапайым құралдарды қамтитын Excel-ге арналған кестелерді талдау құралынан бастау керек:

* нәтижеге әсер ететін факторларды талдау;
* деректер санаттарын анықтау;

* берілген мысалдарға негізделген мәндерді енгізу;
* бірқатар мәліметтерге негізделген болжамдар жасау;
деректердегі ықтимал жаман мәндерді анықтау;
Деректерді іздеу технологиясы
* ықтимал гипотетикалық нұсқаларды талдау;
* көрсетілген мақсатқа жету үшін талаптарды сәйкестендіру
мақсаты;
* бағалауды есептеу үшін пайдаланылуы мүмкін парақты жасау;
* жиі бірге сатып алынатын өнім үлгілерін талдау.

Деректерді іздеумен таныс немесе болжамды талдау үшін неғұрлым күшті құралдарды қажет ететін пайдаланушылар зияткерлік клиент ұсынатын шеберлер мен диалогтық терезелерді қолдануы керек

Excel үшін деректер. Клиенттің мүмкіндіктерін пайдалана отырып, зияткерлік талдау құрылымдары мен модельдерін жасауға және тексеруге болады
деректерді сақтау кезінде оларды басқаруға мүмкіндік береді.
Excel үшін деректерді іздеу клиенті пайдалы
мынадай міндеттерді:
* Деректерді дайындау: зерттеу, тазалау, қайта белгілеу
және деректерді бөлу.
* Талдау: деректерді жіктеу, трендтерді болжау, корреляцияны анықтау және кластерлерді іздеу.
* Тексеру және бағалау: дәлдікті талдау үшін диаграммалар құру
деректерді іздеу және графикалық шешімдер
нәтижелерді жалпы статистикалық өлшемдермен бірге ұсыну.
* Ұсыну: пайдаланушы қарау құралдарын пайдаланып нәтижелерді шолу. Аналитикалық процестерді бақылау және басқару үшін кірістірілген құжаттама шебері қолданылады.
* Кеңейтілген деректерді іздеу: көп жақты талдауды қолдайтын деректер құрылымын құру және деректерді іздеудің жеке модельдерін құру. Интерактивті пайдаланушы интерфейсін қолдана отырып, жеке деректерді іздеу сұрауларын жасау.
* Басқару: SQL Server Analysis Services данасында сақталған деректерді іздеу шешімдерін қарау және басқару.

Деректерді өндірудің негізгі міндеттері.

Жіктеудің міндеті - әр нұсқа үшін оған тиесілі санат немесе сынып анықталады. Мысал ретінде әлеуетті қарыз алушының несие қабілеттілігін бағалауға болады: мұнда тағайындалған сыныптар «несие қабілетті» және «несие қабілетсіз» болуы мүмкін. Айта кету керек, мәселені шешу үшін көптеген сыныптар алдын-ала белгілі және түпкілікті және есептелетін болуы керек.

Регрессия мәселесі көбінесе жіктеу міндетіне ұқсас, бірақ оны шешу барысында сандық мәнді анықтау үшін шаблондар ізделеді. Басқаша айтқанда, мұнда болжанған параметр, әдетте, үздіксіз диапазондағы Сан болып табылады.

Сандық тізбектің қолда бар мәндеріне негізделген жаңа мәндерді болжау міндеті бөлек бөлінеді (немесе корреляция байқалатын мәндер арасындағы бірнеше реттілік). Бұл ретте бар үрдістер (трендтер), маусымдылық, басқа да факторлар ескерілуі мүмкін. Классикалық мысал-биржадағы акциялардың бағасын болжау.

Мұнда аздап шегіну қажет. Зияткерлік талдау мәселесін шешу әдісіне сәйкес екі сыныпқа бөлуге болады: мұғаліммен оқыту (ағылшын тілінен. supervisedlearning) және мұғалімсіз оқыту (ағылш. unsupervisedlearning). Бірінші жағдайда, деректерді іздеу моделі жасалып, оқытылатын мәліметтер жиынтығы қажет. Дайын модель сыналады және кейіннен жаңа деректер жиынтығындағы мәндерді болжау үшін қолданылады. Кейде сол жағдайда олар басқарылатын интеллектуалды алгоритмдер туралы айтады. Жіктеу және регрессия міндеттері дәл осы түрге жатады.

Екінші жағдайда, мақсат-қолданыстағы мәліметтер жиынтығындағы заңдылықтарды анықтау. Бұл жағдайда Оқу үлгісі қажет емес. Мысал ретінде, зерттеу барысында көбінесе бірге сатып алынатын тауарлар анықталған кезде тұтыну қоржынын талдау міндетін келтіруге болады. Кластерлеу міндеті сол сыныпқа жатады.

Сондай-ақ, мақсатты деректерді іздеу тапсырмаларын жіктеу туралы айтуға болады, соған сәйкес олар сипаттамалық (сипаттамалық) және болжамды (алдын-ала) болып бөлінеді. Сипаттамалық есептерді шешудің мақсаты - зерттелген деректерді жақсы түсіну, олардағы заңдылықтарды анықтау, тіпті егер олар басқа мәліметтер жиынтығында кездеспесе де. Болжамды есептер үшін оларды шешу барысында белгілі нәтижелері бар мәліметтер жиынтығы негізінде жаңа мәндерді болжауға арналған модель құрылатындығымен сипатталады.

Енді деректерді іздеу тапсырмаларының тізіміне оралайық

Кластерлеу міндеті-параметрлер бойынша ұқсас көптеген объектілерді топтарға (кластерлерге)бөлу. Бұл ретте жіктеуден айырмашылығы кластерлердің саны мен олардың сипаттамалары алдын ала белгісіз болуы мүмкін және кластерлерді құру барысында біріктірілген объектілердің параметрлер жиынтығы бойынша жақындық дәрежесіне сүйене отырып анықталуы мүмкін.

Бұл тапсырманың тағы бір атауы - сегментация. Мысалы, интернет-дүкен өз клиенттерінің базасына осындай талдау жүргізуге қызығушылық танытуы мүмкін, содан кейін олардың ерекшеліктерін ескере отырып, арнайы топтар үшін арнайы ұсыныстар жасайды.

Кластерлеу дегеніміз - мұғалімсіз оқыту міндеттері (немесе «басқарылмайтын» міндеттер).

Ассоциативті ережелерді табу міндеті деп аталатын қатынастарды анықтау міндеті көптеген ұқсас жиынтықтар арасында жиі кездесетін объектілер жиынтығын анықтау болып табылады. Классикалық мысал - көбінесе бір тапсырыста (немесе бір чекте) кездесетін тауарлар жиынтығын анықтауға мүмкіндік беретін тұтынушы себетін талдау. Бұл ақпарат тауарларды сауда алаңына орналастыру кезінде немесе байланысты тауарлар тобы үшін арнайы ұсыныстар қалыптастыру кезінде пайдаланылуы мүмкін.
Бұл міндет «мұғалімсіз оқыту» класына да қатысты.

Бірізділікті талдау немесе дәйекті талдауды кейбір авторлар алдыңғы тапсырманың нұсқасы ретінде қарастырады, басқалары бөлек бөлінеді. Бұл жағдайда мақсат-оқиғалар тізбегіндегі заңдылықтарды анықтау. Мұндай ақпарат, мысалы, ақпараттық жүйенің сәтсіздігін болдырмауға, көбінесе осы типтегі сәтсіздіктің алдында болатын оқиғаның басталуы туралы сигнал алуға мүмкіндік береді. Қолданудың тағы бір мысалы - веб-сайт пайдаланушыларының беттері бойынша өтулердің реттілігін талдау.

Ауытқуларды талдау көптеген оқиғалардың ішінен нормадан айтарлықтай ерекшеленетін оқиғаларды табуға мүмкіндік береді. Ауытқу қандай да бір ерекше оқиғаны (эксперименттің күтпеген нәтижесі, банктік картадағы алаяқтық операция ...) немесе, мысалы, оператордың деректерді енгізудегі қателігін білдіруі мүмкін.

2. Data Mining технологиясы

Деректерді сақтаудың машиналық формасы пайдалы ақпаратты жасырын түрде қамтиды, оны алу және ыңғайлы түрде ұсыну үшін арнайы әдістерді қолдану қажет. Data Mining технологиясы деректер базасында жаңа білімді табу процестерін зерттейді. Оның негізінде:

- Қолданбалы статистика;
- Жасанды интеллект теориясы сияқты мәліметтер базасының жүйесі жатыр;

Деректерді іздеу мақсаты (ағылш. Datamining, аударманың басқа нұсқалары – «деректерді өндіру», «деректерді қазу») - бұл мәліметтер жиынтығындағы анық емес заңдылықтарды анықтау. Ғылыми бағыт ретінде ол XX ғасырдың 90-шы жылдарында белсенді дами бастады, бұл ақпаратты автоматтандырылған өңдеу технологияларының кең

таралуына және компьютерлік жүйелерде үлкен көлемдегі деректердің жиналуына байланысты болды. Қолданыстағы технологиялар, мысалы, дерекқордан қажетті ақпаратты тез табуға мүмкіндік берсе де, бұл көптеген жағдайларда жеткіліксіз болды. Математикалық статистика, мәліметтер базасы теориясы, жасанды интеллект теориясы және басқа да бірқатар салалардағы әдістерді қажет ететін үлкен көлемдегі деректер арасында жеке оқиғалар арасындағы байланысты іздеу қажеттілігі туындады.

Бағыттың негізін қалаушылардың бірі Григорий Пятицкий-Шапиро берген анықтама классикалық болып саналады: DataMining - бұрын белгісіз, тривиалды емес, іс жүзінде пайдалы, түсіндіруге болатын жасырын білімнің шикі деректерінде «машинаны» (Алгоритмдер, жасанды интеллект құралдары) зерттеу және анықтау.

Data mining әдістерінің негізі шешім ағаштарын, жасанды нейрондық желілерді, генетикалық алгоритмдерді, эволюциялық бағдарламалауды, ассоциативті жадты, анық емес логиканы қолдануға негізделген жіктеу, модельдеу және болжаудың барлық түрлерінен тұрады. Data mining әдістері көбінесе статистикалық әдістерді қамтиды (сипаттамалық талдау, корреляциялық және регрессиялық талдау, факторлық талдау, дисперсиялық талдау, компонентті талдау, дискриминантты талдау, уақыт қатарларын талдау, өмір сүруді талдау, байланыстарды талдау). Мұндай әдістер data mining мақсаттарынан (бұрын белгісіз тривиалды емес және іс жүзінде пайдалы білімді анықтау) айырмашылығы бар талданған деректер туралы кейбір априорлық идеяларды қамтиды.

Data mining әдістерінің маңызды мақсаттарының бірі-есептеу нәтижелерін визуалды түрде ұсыну (визуализация), бұл арнайы математикалық дайындығы жоқ адамдарға data mining құралдарын пайдалануға мүмкіндік береді.

Деректерді талдаудың статистикалық әдістерін қолдану ықтималдықтар теориясы мен математикалық статистиканы жақсы меңгеруді талап етеді.

Деректерді ұсынудың әр түрлі формаларын, қолданылатын алгоритмдер мен қолдану салаларын ескере отырып, деректерді іздеу келесі сыныптардың бағдарламалық өнімдерін қолдана отырып жүргізілуі мүмкін:

- зияткерлік талдауға арналған мамандандырылған «қораптық» бағдарламалық өнімдер;
- математикалық;
- электрондық кестелер (және олардың үстіндегі әртүрлі қондырмалар);
- деректер базасын басқару жүйесіне интеграцияланған құралдар (ДҚБЖ);
- басқа бағдарламалық өнімдер.

Осы курс аясында бізді ДҚБЖ-мен біріктірілген құралдар қызықтырады. Мысал ретінде MicrosoftSQLServer ДҚБЖ және оның құрамына кіретін, пайдаланушыларды MSSQLServer 2000-де алғаш пайда болған on-line (OLAP) режимінде деректерді талдау және деректерді талдау құралдарымен қамтамасыз ететін AnalysisServices қызметтерін келтіруге болады.

Microsoft корпорациясы ғана емес, сонымен қатар ДҚБЖ-нің басқа да жетекші әзірлеушілері өздерінің арсеналында деректерді іздеу құралдарына ие.

Деректерді іздеу міндеттері

Деректерге зияткерлік талдау жүргізу барысында көптеген объектілерге (немесе нұсқаларға) зерттеу жүргізіледі. Көп жағдайда оны кесте түрінде ұсынуға болады, оның әр жолы нұсқалардың біріне сәйкес келеді, ал бағандарда оны сипаттайтын параметрлердің мәндері болады. Тәуелді айнымалы-мәні басқа параметрлерге (тәуелсіз айнымалыларға) тәуелді деп саналатын параметр. Шын мәнінде, бұл тәуелділікті деректерді іздеу әдістерін қолдана отырып анықтау керек.