№6 дәріс. Деректерді талдау. Деректерді басқару.

Қарастырылатын сұрақтар: Деректерді талдау негіздері. Жинау, жіктеу және болжау әдістері. Шешімдер ағаштары. Үлкен көлемді деректерді өңдеу. Data Mining әдістері мен кезеңдері. Data Mining міндеттері. Деректерді визуализациялау.

Деректерді талдау.

Деректерді талдау – пайдалы ақпаратты алу және шешім қабылдау үшін деректерді зерттеу, сүзу, түрлендіру және модельдеу процесі. Деректерді талдау ғылым мен қызметтің әртүрлі салаларындағы әртүрлі әдістерді қамтитын көптеген аспектілері мен тәсілдеріне ие.

Деректерді жинау жоспарын жасау үшін төмендегілер қажет:

1) Мәселелерді анықтау және зерттеу мақсаттарын тұжырымдау.
2) Қызықтыратын тақырыпты алдын ала зерттеуді жүзеге асыру.
3) Зерттеу тұжырымдамаларын жасау.
4) Зерттеуді егжей-тегжейлі жоспарлауды жүзеге асыру.
5) Ақпарат көздерін таңдау және қосымша мәліметтерді жинау.
6) Алынған деректерді бағалау және бастапқы деректердің қаншалықты қажет екенідігі жөнінде шешім қабылдау.
7) Алғашқы мәліметтерді жинау әдісін анықтау: сауалнама, бақылау, эксперимент.
8) Алғашқы ақпаратты тікелей жинау.
9) Зерттеу нәтижелерін ұсыну (презентация).

Деректерді болжау әдістері келесідей бөлінеді: интуитивті – сарапшылар пікірлер мен бағалауларына сүйенетін; формальды – әдебиетте бұрыннан сипатталған және олардың негізінде болжау үлгілері қазірдің өзінде құрылып жатырған.

Шешім ағаштары деректерді талдау саласында кеңінен қолданылады.

Шешім ағаштары – бұл әрбір нысан шешім шығаратын бір түйінге сәйкес келетін иерархиялық, тізбекті құрылымдағы ережелерді көрсету тәсілі.

Ағаш әдісі шешетін барлық есептерді келесі үш класқа біріктіруге болады:

Деректер сипаттамасы: Шешім ағаштары деректер туралы ақпаратты ықшам түрде сақтауға мүмкіндік береді, оның орнына объектілердің дәл сипаттамасын қамтитын шешім ағашын сақтауға болады.

Жіктеу: Шешім ағаштары жіктеу есептерін өте жақсы орындайды, яғни объектілерді бұрыннан белгілі класстардың біріне тағайындау. Мақсатты айнымалының мәндері дискретті болуы керек.

Регрессия: Егер мақсатты айнымалының үздіксіз мәндері болса, шешім ағаштары мақсатты айнымалының тәуелсіз (кіріс) айнымалыларға тәуелділігін орнатуға мүмкіндік береді. Мысалы, бұл класқа сандық болжау есептері кіреді (мақсатты айнымалы мәндерін алдын ала болжау).

Шешім ағаштарын жүзеге асыратын көптеген алгоритмдер бар, соның ішінде CART, C4.5, NewId, ITrule, CHAID, CN2 және т.б. Бірақ ең көп таралғандары мыналар:

CART (*Classification and Regression Tree*) – екілік шешім ағашын құру алгоритмі – дихотомиялық жіктеу моделі. Мұндай ағаштың әрбір түйіні бөлінген кезде екі ғана ұрпағы болады. Алгоритм классификация және регрессия есептерін шешеді.

C4.5 – түйіннің ұрпақтарының саны шексіз болатын шешім ағашын құру алгоритмі.

Data Mining негіздері

Деректердің үлкен көлемін өңдеу туралы айтқанда, деректердің көлемі өте үлкен екенін білдіретін Data Mining терминін қолданамыз.

Data Mining – бұл деректердегі жасырын заңдылықтарды (ақпарат үлгілерін) іздеуге негізделген шешімдерді қолдау процесі. Бұл деректердің үлкен көлемінде анық емес, объективті және тәжірибеде пайдалы заңдылықтарды іздеуге арналған технология.

Data Mining міндеттерін (*tasks*) кейде заңдылықтар (regularity) немесе техникалар (techniques) деп атайды. Data Mining негізгі міндеттеріне мыналар жатады: жіктеу, кластерлеу, болжау, ассоциациялау, визуализация, талдау және ауытқуларды анықтау, бағалау, қарым-қатынасты талдау, қорытындылау.

Data Mining әдістері мен алгоритмдері:

- жасанды нейрондық желілер;
- шешім ағаштары;
- символдық ережелер;
- ең жақын көрші және k-ең жақын көрші әдістері;
- тірек векторлар әдісі;
- байестік желілері;
- сызықтық регрессия;
- корреляциялық-регрессиялық талдау;
- кластерлік талдаудың иерархиялық әдістері;
- кластерлік талдаудың иерархиялық емес әдістері, соның ішінде k-орташалар және k-медианалар алгоритмдері;

- ассоциативтік ережелерді іздеу әдістері, соның ішінде Apriori алгоритмі;
- шектеулі іріктеу әдісі, эволюциялық программалау және генетикалық алгоритмдер, деректерді визуализациялаудың әртүрлі әдістері және басқа да көптеген әдістер.

Data Mining екі немесе үш кезеңнен тұруы мүмкін:

1-кезең. Заңдылықтарды анықтау (еркін іздеу).

2-кезең: Белгісіз мәндерді болжау (болжамдық модельдеу) үшін анықталған заңдылықтарды пайдалану.

3-кезең: Ерекшеліктерді талдау – заңдылықтарда кездесетін ауытқуларды анықтауға және түсіндіруге арналған кезең.

Data Mining құралдарының визуализациясы.

Data Mining алгоритмдерінің әрқайсысы арнайы визуализация тәсілін пайдаланады. Data Mining әдістерінің әрқайсысын пайдалану кезінде, дәлірек айтқанда, оны программалық қамтамасыз етуді жүзеге асыру кезінде біз сәйкес әдістер мен алгоритмдердің жұмысы нәтижесінде алынған нәтижелерді түсіндіре алатын визуализаторлармыз.

Шешім ағаштары үшін мұндай визуализатор шешім ағашы, ережелер тізімі немесе түйіндестілік кестесі болып табылады.

Нейрондық желілер үшін құралға байланысты бұл желі топологиясы, оқу процесін сипаттайтын қателік шамасының өзгеру графигі болуы мүмкін.

Кохонен карталары үшін: кіру, шығу карталары, басқа да арнайы карталар.

Сызықтық регрессия үшін визуализатор ретінде регрессия сызығы пайдаланылады.

Кластерлеу үшін: дендрограммалар, шашырау диаграммалары.

Сол немесе басқа әдістің өнімділігін бағалау үшін шашыраңқы графиктер мен диаграммалар жиі пайдаланылады.

Деректерді көрнекі түрде көрсетудің немесе бейнелеудің осы тәсілдерінің барлығы келесі функциялардың бірінің қызметін атқара алады:

- модельді тұрғызудың иллюстрациясы болып табылады (мысалы, нейрондық желінің құрылымын (графын) бейнелеу);
- алынған нәтижелерді түсіндіруге көмектесу;
- тұрғызылған модельдің сапасын бағалау құралы болып табылады;
- жоғарыда аталған функцияларды біріктіру (шешім ағашы, дендрограмма).

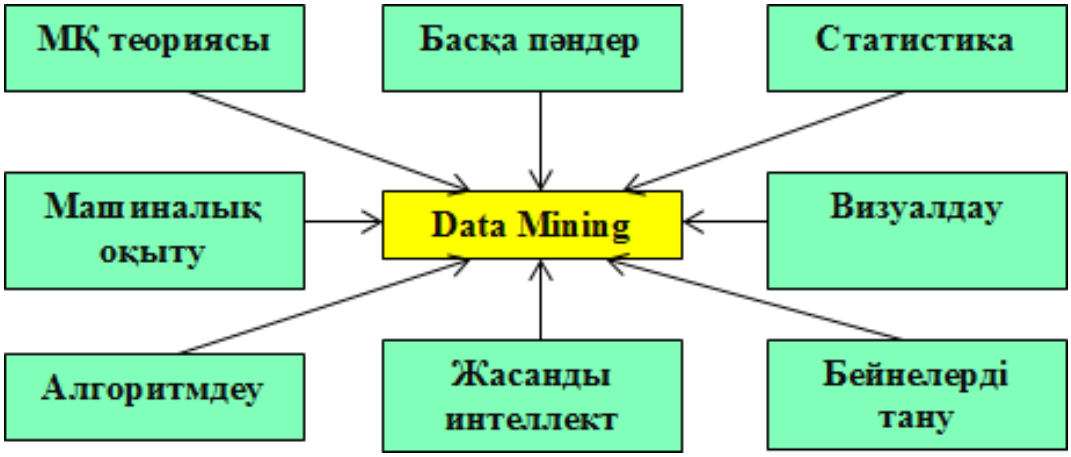
Data Mining технологиясын қолданар алдында оның мәселелерін, шектеулерін және онымен байланысты маңызды сұрақтарын мұқият талдау керек, сонымен қатар бұл технология не істей алмайтынын түсіну керек.

Data Mining талдаушыны (аналитик) алмастыра алмайды!

Қойылмаған сұрақтарға технология жауап бере алмайды. Ол аналитикті алмастыра алмайды, тек оның жұмысын жеңілдету және жақсарту үшін күшті құралды береді.

Data Mining қолданбасын әзірлеу және пайдаланудың күрделілігі

Бұл технология көппәнді сала болғандықтан, Data Mining-ті қамтитын қосымшаны әзірлеу үшін әртүрлі сала мамандарын тарту, сонымен қатар олардың жоғары сапалы өзара әрекеттесуін қамтамасыз ету қажет.



1-сурет. Data Mining көппәнді сала ретінде.

Бақылау сұрақтары

1. Деректерді талдау дегеніміз не?
2. Деректер дегеніміз не?
3. Деректерді болжау әдістері қалай жіктеледі?
4. Деректерді талдаудағы регрессия дегеніміз не?
5. Деректерді визуализациялау дегеніміз не?
6. Data Mining дегеніміз не?
7. Data Mining әдістері.
8. Шешім ағашы дегеніміз не?
9. Data Mining міндеттері қандай?
10. Шешім ағашын құрудың алгоритмдері қандай?