

Лекция 9. Data Mining модельдері

1. Лекция мақсаты мен міндеттері: Data Mining технологиясының мақсаты мен оқытатын қарапайым тәсілдер. Модельдердің түрлерін қарастыру.

2. Лекция мазмұны:

1. Data Mining технологиясы.

2. Модельдердің түрлері.

3. Оқытудың техникалық құралдары: Zoom онлайн оқыту платформасы

4. Лекция оқудың тәртібі, оқыту әдістері мен түрлері: баяндау, сұрақ – жауап, түсіндіру, кіріспе лекция

Data Mining технологиясының мақсаты - осындай модельді ақпараттарды табу, қарапайым тәсілмен табылмайтын тәсілдерге қатысты болып табылады. Модельдердің екі түрі бар: болжамдық және сипаттаушы.

Сипаттаушы модельдер. Болжау модельдері белгілі нәтижелері бар деректертер жиынтығының негізінде құрылады. Олар өзге деректертер жиынтығының негізінде нәтижелерді болжау үшін қолданылады. Бұл жағдайда, әрине, модельдің максималды нақты, статистикалық жағынан мәнді және дәлелді және т.б. болуы талап етіледі.

Осындай модельдерге келесідей жатады:

- Классификация модельдері - ережелер мен ережелер жиынтығын сипаттайды, соларға сәйкес қандайда бір жаңа объектіні класстардың біріне жатқызуға болады. Мұндай ережелер қазіргі бар объектілер туралы ақпараттарды кластарға бөлу жолымен құрылады;

- Реттілік модельдері - үздіксіз сандық параметрлерді болжауға мүмкіндік беретін функцияларды сипаттайды. Олар өткен уақыт кезеңінде кейбір параметрлерді өзгерту жөніндегі деректер негізінде құрылады.

Сипаттаушы (descriptive) модельдер деректер жиынтығында түрлі факторлардың өзара әсеріне, тәуелділік мәніне, яғни түрлі жүйелердің эмпириялық модельдерінің құрылымына назар аударады. Мұндай модельдердегі басты жол - адамның қабылдауы үшін жеңілдік пен анықтылық. Мүмкін табылған заңдылықтар нақты зерттелетін деректердің арнайы белгісі болып және өзге жерде кездеспейтін шығар, бірақ бұл қалайды падалы және сондықтан да мәлім болуы тиіс.

Мұндай модельдердің келесі түрлері бар:

- Регрессионды модельдер - тәуелді және тәуелді емес көрсеткіштер арасындағы және адамға түсінікті формаға ауыспалы қызметтік тәуелділіктерді сипаттайды. Мұндай модельдер тек үздіксіз сандық параметрлер арасында ғана емес, сондай-ақ санатты параметрлер арасындағы қызметтік тәуелділікті сипаттайтындығын ескере кеткен жөн;

- Кластер модельдері - топтарды (кластерлерді) сипаттайды, оларға деректері талдауға ұшырайтын объектілерді бөлуге болады. Объектілер (бақылаулар, оқиғалар) объектілердің мәнін сипаттайтын деректердің (қасиеттердің) негізінде топтастырылады. Кластер ішіндегі объектілер бір-біріне «ұқсас» және өзге кластерге кіргендерден өзгеше болуы мүмкін. Кластер ішіндегі объектілер неғұрлым бір-біріне «ұқсас» және кластерлер

арасындағы айырмашылықтар көп болған сайын, кластеризация нақтырақ бола түседі;

- Ерекшелік модельдер - жазбалардағы ерекше жағдайларды сипаттайды (мысалы жеке емделушітерді), олар негізгі көпшілік жазбалардан (науқастардың топтарынан) бірден өзгеше болып көрінеді. Ерекшелікті білу екі түрлі болуы мүмкін. Бәлкім, бұл жазбаларда оператордың шатасуынан кеткен кездейсоқ қателіктер болуы мүмкін. Сипатты жағдай: егер оператор қателесіп ондық нүктені басқа жерге қоя салса, онда мұндай қателік дереу құрылу тәртібіне әсер етеді. Мұндайда кездейсоқтықты зерттеуден мүлде алып тастау қажет, бұл тарау осындай «шығарындылардың» болуына өте сезімтал – дереу оқшау көрінетін нүктелерге, сирек кездесетін, әдеттегіден тыс жағдайларға. Басқа жағынан қарайтын болсақ ерекшелігі бар жазбалар зерттеудің қызығушылығын тудыруы мүмкін, яғни олар кейбір сирек кездесетін, бірақ тосын сырқаттардың көрсеткіші болуы мүмкін. Тіпті бұл жазбалардың теңестірілуі, оларды ары қарай талдау және бөлшектеп қарастыру жөнінде айтпағанда, зерттелуші объектілер мен олардың мәнін түсінуде өте пайдалы болуы мүмкін;

- Қорытынды модельдер - талданушы деректер массивіне шектеуді анықтау. Мысалы, миокард инфарктісін бастан кешкен, 30 жастан аспаған емделуші бойынша деректерді қарағанда, олардың күніне 5 қораптан аса шылым шегетіндігі немесе салмағының 95 килограмнан төмен еместігі анықталады. Мұндай шектеулер массив дерегін түсіну үшін өте маңызды, бұл талдаулардың негізінде алынған жаңа білім. Осылайша қорытынды модельдерді құру барлығы үшін немесе зерттелуші деректерді іріктегендегі барлық жазбаларда дұрыс қандайда бір деректерді табуға шоғырланады, бірақ осындай форматтағы жазбалардың алуан түрлілігі сирек кездесетін, сол белгіленген жолдарда орналастырылуы сирек кездестіріледі. Егер салыстыру үшін барлық емделушілер бойынша алатын болсақ, онда көп шылым шегетіндердің немесе өте семіз адамдардың пайызы аса көп емес. Жіктелудің анық емес тапсырмасы шешілуде деп айтуға болады, бірақ деректі түрде ұсынылған деректері бар бір ғана кластың тапсырмасы тапсырылған;

- Ассоциативті модельдер - байланысты оқиғалар арасындағы заңдылықтарды анықтау. Мұндай заңдылықтардың мысалы ретінде келесі ереже қызмет етеді, X оқиғадан Y оқиға туындайды. Мұндай ережелер ассоциативті деп аталады.

Қарастырылған модельдерді құрау үшін, Data Mining алгоритмі мен тәсілдері қолданылатын болады. Data Mining технологиясы статистика, ақпарат теориясы, машиналық оқыту және деректер базасы теориясы сияқты пәндер тоғысуында дамыды және дамып келетіндігін ескере отырып, Data Mining әдістерінің және алгоритмдерінің көпшілігі түрлі технологиялар мен тұжырымдар негізінде жасалған, заңдылықтарға сәйкес деуге болады. Ары қарай Data Mining әдістері тарататын технологияларды қарастырайық.

Базалы тәсілдер. Базалы Data Mining тәсілдерге ең алдымен алгоритмдерді жатқызу тура келеді. Барлық зерттелетін объектілерді оңай

таңдау, $2n$ операцияларға негізделетін болады, мұндағы N - объектілердің саны.

Деректердің көлемінің ұлғаюына қарай есептеулердің көлемі де ұлғаюда, қандайда бір тапсырманың көлемі көп болған жағдайда шешімі тәжірибе жүзінде мүмкін еместей көрінеді. Мұндай алгоритмдердегі есептеу күрделілігін қысқарту үшін артықшылықты қысқартуға әкелетін эвристиканың түрлерін қолданады. Осы алгоритмдерді оңтайландыру операциялар санын линиялық түрдегі қызметтің зерттелуші деректерінің санына тәуелді етуге әкеледі. Сондай-ақ атрибуттар санының тәуелділігі, әлбетте, экспоненциалды болып қалады. Олардың көп естігін ескрегенде (көптеген жағдайларда сирек кездесетіндігі) мұндай тәуелділікті құптауға болады.

Бұл алгоритмдер деректерінің негізгі қасиеті, оларды түсіну және қолдану көзқарасынан алғанда, қарапайымдылығы болып табылады. Кейбір жетіспеушіліктеріне орындай алгоритмдер құрудың негізінде формальды теорияның жоқтығын, оларды зерттеу мен дамытудағы қиындықтарды жатқызуға болады.

Data Mining базалы тәсілдерге статистика теориясындағы бағыттарды да жатқызуға болады. Осыған орай, Data Mining статистиканың дамуы болады, мұндай әдістер жеткілікті. Олардың негізгі идеясы: корреляционды және регрессионды негіздеме арқылы жүргізілетін болады. Басты жетіспеушілігі белгілерді орташалаандыру, ол деректердің ақпараттылығын жоғалтады Бұл, әрине, алынатын білімді төмендетеді.

Нақты емес логика. Деректерді талдауды зерттеудің басты әдісі болып олардың қалыпты тілге ауыстырылуы және алынған модельді талдау табылады. Жоқ ақпарат жөніндегі бірдеңгейсіздікті келесідей 3 топтар арқылы анықтауға болады:

- Анықталмағандық;
- Толықсыздық (жетімсіздік, адекватсыздық);
- Нақтылықсыздық.

Анықсыздық табиғи (оның қайнар көзі сыртқы орта) және лингвистикалық (сөздік қорыту нәтижесінде туындайды және сансыз жағдайлар санын шектеулі сөздермен шектеулі уақытта сипаттау қажеттілігімен шарттанады) болады.

Физикалық нақтылықсыздықтың екі түрі белгілі:

1. Дәлсіздік (табиғи құралдармен жасалатын, белгілі көлемдегі мәндердің өлшемінің дәлсіздігі);
2. Кездейсоқтық (немесе сыртқы ортада бірнеше мүмкіндіктің болуы, олардың әрқайсы кездейсоқ жағдаймен шынайы болуы мүмкін; мүмкіндіктерді таратудың лайықты заңдылықтарын меңгеруді алғышарттаған).

Лингвистикалық нақтылықсыздықтың екі түрі бар:

1. Сөздердің мәндерінің анықсыздығы (көп мәнділік, тиянақсыздық, түсініксіздік, дәлсіздік). Ол, бір ғана сөзбен белгіленетін тапсырманың басқару объектілері әртүрлі болған жағдайда туындайды;

2. Сөз орамдары мәнінің бірнеше мәндес болуы (синтаксистік және семантикалық болады).

Табиғи анықсыздықтарды өңдеу үшін мүмкіндік теориясының әдістері мен көптіктің классикалық теориясы тиімді қолданылады. Бірақ жүйелердің дамуымен қатар табиғи тілдің қатынасын және түсінігін өңдеуді талап ететін жасанды интеллект теориясының әдістерінде тапсырмалардың лингвистикалық анықсыздығын есепке алу мақсатында көптеген қалыпты әдістерді кеңейту қажеттігі туындады.

Дәл емес көптік теориясының негізін қалаушы болып басқару облысындағы танымал маман Л.Заде болуы кездейсоқтық емес. Өйткені, дәл емес логика туралы бастапқы идеяда шешімдерді қабылдау үрдістері мен басқару жөніндегі ұсыныстар жинақталды. Ал мұндай тапсырмалар барлық технологиялық процесстерде туындайды, осы теорияны дамытудағы қажеттілік және олардың мүмкіндіктері өте кең.

Бұл параметрлерді және ауыспалы сандарды ұлғайтумен, жеке ауыспалы күрделіліктерді жоғарылатумен байланысты. Нәтижесінде тепе-тең модельді құру мүлде мүмкін емес. Осының орнына Л. Заде лингвистикалық модельді ұсынды, онда математикалық өрнектер емес сапасын білдіретін сөздер қолданылады. Сөздік модельдерді қолдану математикалық модельдеуге ұқсас дәлдікті қамтамасыз етпейді, бірақ жақсы сапалы модельді құруға болады. Бұл жағдайда талқыға түсетін сипаттау жүйесі тіліндегі сөздердің нақты еместігі болады.

5. Деңгейлік тапсырмалар:

1-деңгей.

- Data Mining технологиясы.
- Модельдердің түрлері.

2-деңгей.

- Сипаттаушы модельдерге қандай модельдер жатады?
- Классификация модельдері дегеніміз не?

3-деңгей.

- Реттілік модельдері
- Регрессионды модельдер
- Кластер модельдері

6. БОӨЖ тапсырмалары:

Логистикалық регрессия. Градиентті түсу.

7. БӨЖ тапсырмалары:

Байесов реттелуі. Деректер мен модельдің бірлескен сенімділігінің максималды принципі. Квадраттық (Гаусс) және Лаплас реттегіштері.

8. Пайдаланылған әдебиеттер:

1. Ғ. Айғараева, Қ. Асанова, М. Нысанов. Сандық қондырғылар және микропроцессорлық жүйелер. Оқулық. 2010
2. Д. С. Козен. Есептеу теориясы. Сандық. Оқулық Алматы. 2013.
3. Асамбаев А.Ж. Жасанды интеллект негіздері.- Алматы: ЖШС РПБК "Дәуір", 2011-135б. Оқу құралы.

4. Рассел С., Норвиг П. Жасанды интеллект: Жаңашыл әдіс -Алматы, 2014..
Оқу құралы.
5. К. Пол. Цифровое искусство. Учебник. Москва , 2017, 220
6. Андреас Мюллер, Сара Гвидо. Введение в машинное обучение с помощью Python. –Москва. 2016.
7. В.Н.Негода. Машинно-ориентированное программирование. Учебное пособие Ульяновск 2015.