Дәріс 3

Деректерді жинау әдістемесі

Дәстүрлі түрде деректерді интеллектуалды талдау процессінде келесі кезеңдер бөлінеді:

1. Нәтижесінде талдаудың негізгі мақсаттары тұжырымдалатын пәндік облысты зерттеу.
2. Деректерді жинау
3. Деректерді алдын-ала өңдеу:

a. Деректерді тазалау – бастапқы деректердегі кездейсоқ «шулар» мен карама-кайшылықтарды жою.

b. Деректерді интеграциялау – бірнеше мүмкін болатын ақпарат көздерінен алынған деректерді бір қоймаға біріктіру.

c. Деректерді түрлендіру. Бұл кезеңде деректер талдау үшін лайықты түрге түрленеді. Көп жағдайда деректерді агрегаттау, атрибуттарды дискреттеу, деректерді қысу және мөлшерін қысқарту қолданылады.

4. Деректерді талдау. Осы кезең шеңберінде үлгілер алу мақсатында интеллектуалды талдау алгоритмдері қолданылады.
5. Табылған үлгілерді интерпретациялау. Бұл кезең алынған үлгілерді визуалды түрде көруге мүмкіндік береді.
6. Жаңа білімді қолдану

Әдетте деректерді интеллектуалды талдау жүйелерінде келесі негізгі компоненттер көрсетіледі:

1. Деректер базасы, деректер қоймасы немесе басқа да ақпарат қоймасы. Бұл тазарту және интеграциялау орындау мүмкін бола алатын бір немесе бірнеше деректер базасы, деректер қоймасы, электронды кестелер, қойманың басқа да түрлері бола алады.
2. Деректер базасының немесе деректер қоймасының сервері. Көрсетілген сервер пайдаланушы сұранысы негізінде маңызды деректерді алу үшін жауап береді.
3. Білім базасы. Бұл нәтижелік үлгілерді (паттерн) қалай іздеу керектігін және пайдалылығын бағалауды көрсететін пәндік облыс туралы білім.
4. Білім табу қызметі. Ол деректерді интеллектуалды талдау жүйесінің ажырамас бөлігі және сипаттамалау, қауымдастықты іздеу, классификация, кластерлі талдау, ауытқуды талдау сияқты тапсырмаларға функционалды модульдер жинағын қамтиды.
5. Үглілерді (паттерн) бағалау модульдері. Бұл компонент үлгінің қызығушылық мөлшерін немесе пайдалылығын есептеп шығарады.
6. Графикалық қолданушы интерфейсі. Бұл модуль қолданушы мен деректерді интеллектуалды талдау жүйесі арасындағы байланысқа, үлгілердің әртүрлі түрдегі визуализациясына жауап береді.

Үлкен деректерді талдаудың әдістемесі

Негізінде статистика мен информатикадан (мысалы, машиналық оқыту) алынған құралдар жатқан деректер жиынын талдаудың көптеген әртүрлі әдістері бар. Берілген тізімде барлық әдістер сипатталмаған, бірақ онда әртүрлі салада көбірек сұранысқа ие болатындар келтірілген. Бірақ бұл ретте зерттеушілер жаңа әдістемелерді ойлап табу үстінде және қолданыстағыны жетілдіру жұмыстарын жалғастырып жатқанын түсіну қажет. Сонымен қатар, келтірілген әдістемелердің ішінде кейбіреуі тек қана үлкен деректерге ғана қолданылуы міндетті емес және көлемі кішірек массивтер үшін де табысты қолданылуы мүмкін (мысалы, A/B тестілеу, регрессиялық талдау). Әрине, неғұрлым көлемді және алуан түрлі массив талдауға ұшыраса, шығуда соғұрлым дәл және орынды деректерді алу мүмкін болады [2].

– A/B testing. Бақылау сынамасы басқалармен кезекпен салыстырылатын әдістеме. Осылайша, жетістік үшін көрсеткіштердің оңтайлы комбинациясын анықтау мүмкін болады, мысалы, маркетингтік ұсынысқа тұтынушылардың ең үздік жауап реакциясы. Үлкен деректер орасан зор көп цикл жүргізуге мүмкіндік береді және, осылайша, статистикалық сенімді нәтиже алуға болады.

– Association rule learning. Өзара байланыстарды, яғни, деректердің үлкен массивіндегі айнымалы шамалар арасындағы ассоциативті ережелерді анықтауға арналған әдістеме жинағы. Data mining-те қолданылады.

– Classification. Белгілі бір нарық сегментінде тұтынушылардың мінез-құлқын болжауға мүмкіндік беретін әдістеме жинағы (сатып алу туралы шешімдер қабылдау, кетуі, тұтыну көлемі және т. б.). Data mining-те қолданылады.

– Cluster analysis. Алдын ала белгісіз, жалпы белгілерді анықтау арқасында объектілерді топтар бойынша жіктеудің статистикалық әдісі. Data mining-те қолданылады.

– Crowdsourcing. Ақпарат көзінің үлкен санынан деректерді жинау әдістемесі.

– Data fusion and data integration. Әлеуметтік желілерді қолданушылардың пікірлерін талдауға және оны нақты уақыт режимінде сату нәтижелерімен салыстыруға мүмкіндік беретін әдістеме жиынтығы.

– Data mining. Сатылатын тауарға немесе қызметке неғұрлым сезімтал тұтынушылар категориясын анықтауға мүмкіндік беретін, ең табысты қызметкерлердің артықшылығын анықтайтын, тұтынушылардың мінез-құлық моделін болжайтын әдістеме жинағы.

– Ensemble learning. Бұл әдісте көптеген предикативті модельдер іске қосылады, осының есебінен жасалған болжамдардың сапасы артады.

– Genetic algorithms. Бұл әдісте мүмкін болатын шешімдер, бірігіп және өзгеріп тұратын «хромосомалар» түрінде ұсынылады. Табиғи эволюция процесіндегідей мұнда неғұрлым бейімделген дарақ аман қалады.

– Machine learning. Деректердің эмпирикалық талдауы негізінде өздігінен

білім алу алгоритмдерін құру мақсатын көздейтін информатикадағы бағыт (тарихи түрде оған «жасанды интеллект» атауы бекітілген).

– Natural language processing (NLP). Адамның табиғи тілін танып білетін, информатика және лингвистикадан алынған әдістер жинағы.

– Network analysis. Желідегі түйіндер арасындағы байланысты талдайтын әдістеме жинағы. Әлеуметтік желілерге қолданылады, жекелеген пайдаланушылар, компаниялар, қауымдастықтар және т.б. арасындағы өзара байланысты талдауға мүмкіндік береді.

– Optimization. Бір немесе бірнеше көрсеткішті жақсарту үшін, күрделі жүйелер мен процестерді өзгертуге арналған әдістердің сандық жинағы. Стратегиялық шешімдерді қабылдауда көмектеседі, мысалы, нарыққа шығарылатын өнімнің құрамы, инвестициялық талдау өткізу және т.б.

– Pattern recognition. Тұтынушылардың мінез-құлық моделін болжау үшін арналған өздігінен білім алу элементі бар әдістеме жинағы.

– Predictive modeling. Алдын ала берілген оқиғалар дамуындағы ықтимал сценарийдің математикалық моделін құруға мүмкіндік беретін әдістемелер жинағы. Мысалы, CRM-жүйесінің деректер базасын талдауда абоненттерді провайдерді ауыстыруға итермелейтін мүмкін шарттар.

– Regression. Тәуелді және бір немесе бірнеше тәуелсіз айнымалы арасындағы өзгеру заңдылығын анықтауға арналған статистикалық әдістердің жинағы. Болжау мен болжам жасау үшін жиі қолданылады. Data mining-те қолданылады.

– Sentiment analysis. Тұтынушылардың көңіл-күйін бағалау әдістемесінің негізінде адамның табиғи тілін тану технологиясы жатыр. Олар жалпы ақпараттық ағыннан қызықтыратын затпен байланысты болатын хабарларды (мысалы, тұтынушылық өнімдер) бөліп алуға мүмкіндік береді. Бұдан әрі пайымдаулар полярлығын (оң немесе теріс), эмоция дәрежесін және т.б бағалауға болады.

– Signal processing. Шу фонында сигналды тану және оны одан әрі талдау мақсатын көздейтін, радиотехникадан алынған әдістеме жинағы.

– Spatial analysis. Кеңістік деректер – жер топологиясы, географиялық координаттар, нысан геометриясын талдайтын, статистикадан аздап алынған әдістеме жинағы. Бұл жағдайда, үлкен деректердің қайнар көзі ретінде жиі геоақпараттық жүйелер (ГАЖ) болады.

– Statistics. Пікіртерім жасап шығару және тәжірибелер жүргізуді қоса алғанда, деректерді ұйымдастыру және интерпретациялау, жинау туралы ғылым. Статистикалық әдістер сол немесе өзге де оқиғалар арасындағы өзара байланыс туралы пайымдауларды бағалау үшін жиі қолданылады.

– Supervised learning. Талданатын деректер массивіндегі функционалды өзара байланысты анықтауға мүмкіндік беретін, машиналық оқыту технологиясына негізделген әдістеме жинағы.

– Simulation. Күрделі жүйелердің мінез-құлықын модельдеу болжау, болжам жасау және жоспарлауда әртүрлі сценарийлер жасау үшін жиі пайдаланылады.

– Time series analysis. Деректер тізбегінің уақыт ағынында қайталануын талдайтын, статистика мен сигналдарды цифрлық өңдеуден алынған әдістеме жинағы. Айқын қолданудың бірі – бағалы қағаздар нарығын немесе пациенттердің сырқатын қадағалау.

– Unsupervised learning. Талданатын деректер массивінде жасырын функционалды байланысты анықтауға мүмкіндік беретін, машиналы оқыту технологиясына негізделген әдістеме жинағы. Cluster analysis-пен ортақ ерекшеліктері бар.

– Visualization. Алынған нәтижелерді түсінуді жеңілдету үшін үлкен деректерді талдау нәтижелерін диаграмма немесе анимацияланған суреттер түрінде ұсынудың графикалық әдісі.