

Деректерді компьютерлік өңдеу жағдайындағы модельдерді сәйкестендіру

Деректерді компьютерлік өңдеу

Клиенттің есептік ақпаратын өңдеу процесінде әртүрлі кластағы компьютерлерді (калькуляторларды қоспағанда) қолдануды көздейді. Экономикалық субъект оларды дербес немесе үшінші тараппен жасалған шарт бойынша қолдана алады.

Компьютерлік деректерді өңдеу:, деректерді бір форматтан немесе жүйеден екіншісіне түрлендіру процесі, мысалы, басып шығарылған құжатты (қатты көшірме) магниттік таспадағы ақпаратқа (магниттік таспа) түрлендіру кейде қайта форматтау деп аталады.





Сәйкестендіру мәселесін шешу-бұл математикалық модель түрінде жаңа білім ала отырып, деректерді өңдеудің интеллектуалды процесі. Data Mining жалпы терминімен анықталған шетелдік әдебиеттерде деректерді іздеу және алу әдістері мен алгоритмдерін жасау мәселесі жыл сайын ғылымның, өнеркәсіптің және экономиканың көптеген салаларында туындайтын нақты мәселелерді шешуде практикалық маңызы артып келеді.

Осы деректерді пайдалану үшін мамандандырылған әдістер мен алгоритмдерді қолдану негізінде пайдалы, тривиалды емес және бұрын белгісіз заңдылықтарды анықтау үшін деректерді автоматты түрде алудың және талдаудың жеке бағдарламалық құралдарын қолдану қажет. Осылайша, заманауи бағдарламалық жасақтамаға білім алу үшін ақпаратты интеллектуалды өңдеу міндеті қойылады. Білім фактілердің жиынтығы ғана емес, себебі олар деректер тізіміне емес, заңдылықтарға негізделген. Дәл осы міндетті заманауи Data Mining құралдары орындайды.



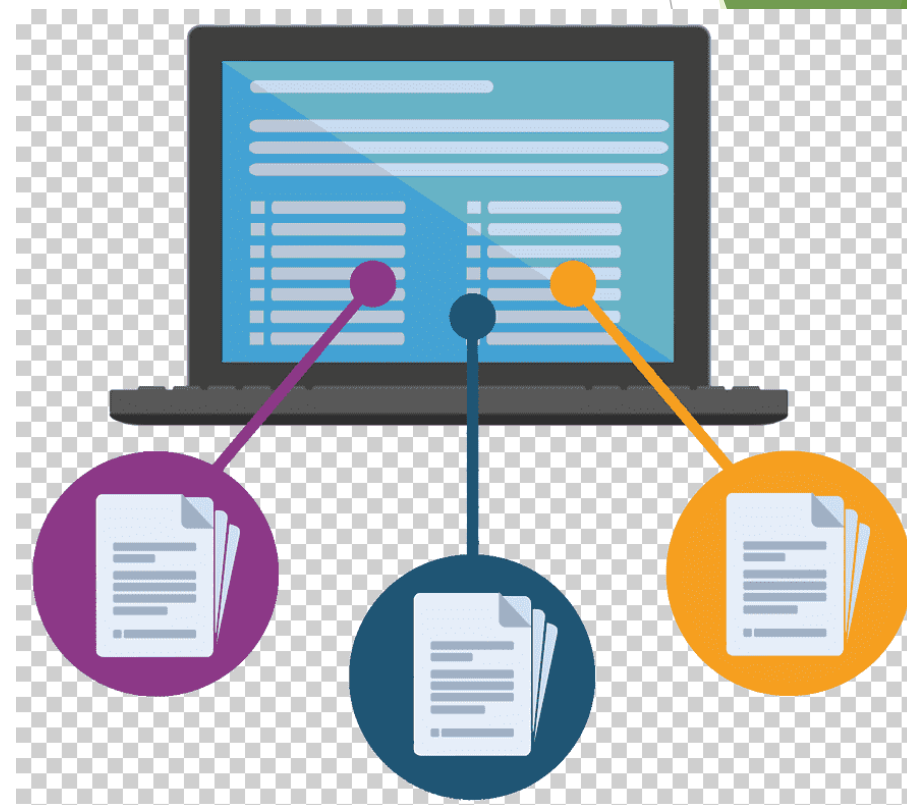
Шикі деректер толығымен немесе ішінара құрылымдалған болуы мүмкін, сондықтан бұл жүйелер реляциялық және басқа мәліметтер базасымен (мысалы, көп өлшемді мәліметтер базасы) жұмыс істеуі керек. Деректерді алу және талдау тиімділігін анықтайтын нақты математикалық аппарат ақпаратты сақтау әдісі мен формасына аз тәуелді.



Ақпаратты сақтаудың заманауи жүйелері келесі ерекшеліктерге ие көптеген деректерді қамтуы мүмкін:

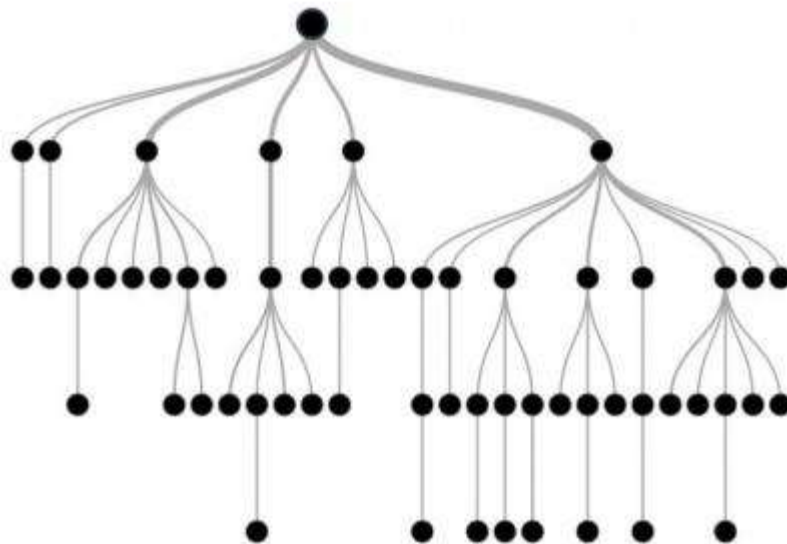
- ▶ Әдетте, объектілердің сипаттамасында кемінде 25-50 сипаттамалық белгілер немесе мәліметтер базасының өрістері болады, мұнда әр белгі дискретті (және 5-10 немесе одан да көп мәндерге ие) немесе үздіксіз болуы мүмкін.
- ▶ Мәндер мен белгілердің жиынтығы түпкілікті емес және өзгеруі мүмкін. - Объектілердің сапасын анықтайтын критерийлер формальды да, сараптамалық та сипатқа ие, яғни әрдайым айқын тәуелділікпен айтыла бермейді.

- Деректер көлемі (объектілер мысалдары) айтарлықтай үлкен және өсуде, сонымен қатар кейбір сәтті объектілер технологиялардың тұрақты дамуына байланысты уақыт өте келе сапасын жоғалтады.
- Тәжірибелі сарапшы сапалы нысанды анықтайтын заңдылықтың интуитивті біліміне ие және сәтті нысандарды сәтсіздерден ажырата алады, бірақ өз білімін нақты түрде білдіре алмайды.
- Объектілерді бағалау және саралау процесі айтарлықтай уақытты алады.



Қорытындылай келе, гетерогенді деректердің шексіз көлемімен жұмыс істейтін Data Mining жүйелері пайдаланушыға пайдалы, нақты, түсінікті және айқын емес нәтижелерді алу үшін қолдануға оңай құралдарды ұсынуы керек екенін атап өтуге болады. Осыны ескере отырып, деректерді іздеу технологиясын қолдана отырып, заңдылықтарды іздеу айтарлықтай есептеу ресурстарын және мәліметтер ағынынан білім алу процесін ұйымдастырудың әртүрлі әдістері мен тәсілдерін синтездеуді қажет ететіні белгілі болады.

Көптеген заманауи құралдар мен деректерді іздеу жүйелерінің негізінде жатқан заңдылықтардың негізгі түрлерін, деректерді өндірудің тиісті міндеттерін және сәйкес келетін кейбір алгоритмдерді қарастырыңыз.

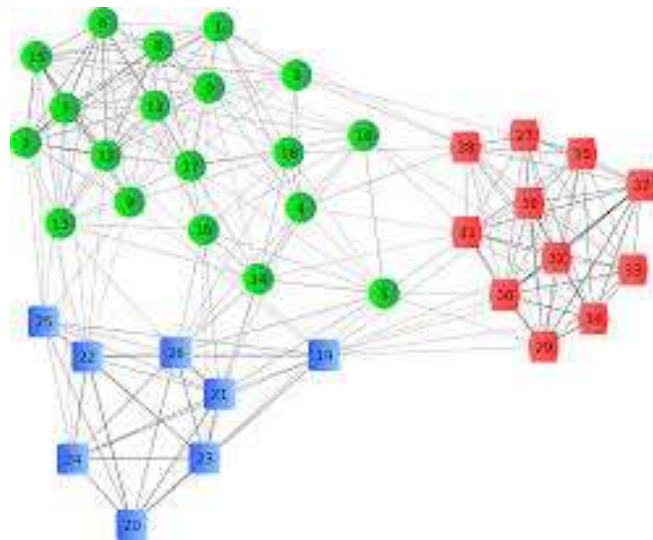


Есеп түрлері және деректерді іздеу алгоритмдерінің түрлері. Деректерді талдау алгоритмдерінің әртүрлі түрлері бар. Олардың ішінде кейбір танымал әдістер мен тәсілдерді бөліп көрсетуге болады — мысалы, статистикалық талдаудың әртүрлі әдістері (мысалы, корреляциялық, факторлық, регрессивті, дисперсиялық және кластерлік талдау), көп өлшемді кестелер, генетикалық Алгоритмдер және нейрондық желілер, анық емес логика, жақын көршінің алгоритмі, дәлелдердің топтық есебі және шешім ағаштары.

Деректерді өңдеу алгоритмін таңдау көптеген факторларға байланысты, алайда, ең алдымен, жүйе шешетін мәселенің түрін және осыған сәйкес жүйенің зерттелетін деректер көлемінде анықталуы керек үлгі түрін ескеру қажет. Дәстүр бойынша, Data Mining – те келесі типтік тапсырмалар бөлінеді (және сәйкесінше типтік үлгілер):

- ▶ қауымдастық;
- ▶ кластерлеу;
- ▶ жіктеу;
- ▶ ерекшелік;
- ▶ болжау.

Қауымдастықтарды табу міндеті объектінің байланысты қасиеттерін іздеуге дейін азаяды, ал мәселе базаның өрістері арасында дұрыс диагноз қою және байланыс орнату болып табылады. Өз әдістері бойынша қауымдастықтарды табу міндеті кластерлеу міндетіне ұқсас, тек айырмашылығы-біріншісі объектінің белгілерімен, ал екіншісі объектілердің өзімен жұмыс істейді. Байланыстарды анықтау міндеттерінің қосалқы класы ретінде уақыт тізбегін анықтау мәселесін ажыратуға болады. Объектінің қасиеттері арасындағы байланысты анықтауға шоғырланған қауымдастықтарды табудың классикалық міндетінен айырмашылығы, тізбекті анықтау міндеті уақытқа байланысты оқиғаларды анықтаумен байланысты. Сонымен қатар, бұл объектінің қасиеттерінің уақытқа байланысты өзгерістерінің тізбегі болуы мүмкін.



Кластерлеу міндеті ұқсас қасиеттері бар объектілер тобын бөлу болып табылады, ал топтардың өздері алдын-ала анықталмаған. Деректерді талдайтын қосымша объектілер топтарын қандай критерийлер мен белгілер бойынша ажырату керектігін дербес анықтауы және сол арқылы оларды кейіннен кластерлеу үшін пайдалануы тиіс. Дәстүр бойынша, бұл мәселені шешу (объектілер тобын анықтау) жіктеу тапсырмасын орындаудың бастапқы нүктесі бола алады.

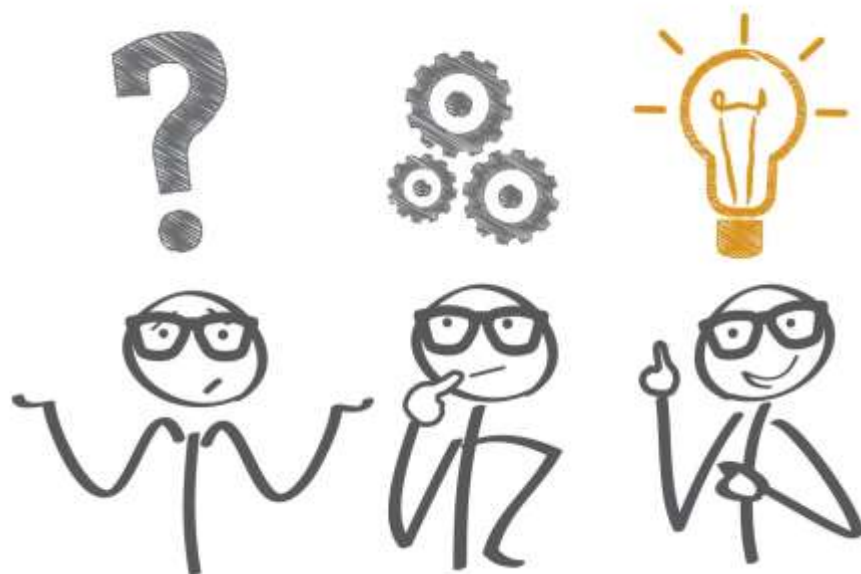
Жіктеу міндеті-топты сипаттайтын белгілерді анықтау және кейіннен тану, яғни. талданатын объектілер массивін олардың атрибуттарын оқу үлгісінің бақылау атрибуттарымен салыстыру негізінде белгілі бір топпен байланыстыру. Сонымен қатар, деректерді жіктейтін қосымшаларға қойылатын маңызды талаптар-бұл ерекшеліктердің ең аз саны, сонымен қатар тану процесін оңтайландыру механизмінің болуы, өйткені объектілер мен олардың қасиеттерінің саны өте үлкен болуы мүмкін. Әрі қарай, жіктеу тапсырмасын орындау нәтижесінде алынған білім негізінде Алып тастау тапсырмасын нақтылауға болады.



Ерекше жағдайларды табу міндеті атрибуттардың мәні басқалардан күрт ерекшеленетін объектілерді табуға бағытталған. Сонымен қатар, бұрын анықталған объектілердің орташа параметрлерінің мәндеріне сүйене отырып, параметрлері осы орташа мәндерден өте ерекшеленетін объектілерге ерекше назар аударылады. Басқа мақсаттардан басқа, бұл талдауды жіктеу тапсырмасының дәлдігін тексеру үшін де жүргізуге болады.

Жоғарыда аталған міндеттерді орындау нәтижелері, өз кезегінде, түпкілікті мәселені шешуде қолданылады — болжау. Бұл тапсырманы орындау кезінде Мақсат объектінің кейбір өрістерінің мәндері бойынша қалған мәндердің мәндерін болжау болып табылады. Сонымен қатар, болжаудың жеке міндеті сонымен қатар белгілі бір өрісте күтілетін нәтижелерге қол жеткізуде шешуші рөл атқаратын өрістер мен олардың оңтайлы мәндерін анықтау болып табылады. Осылайша, деректерді іздеу есептерінің сипаттамасынан олардың ерекшелігі көбінесе тиісті шешім алгоритмін таңдауды анықтайтындығын көруге болады.

Осылайша, дұрыс алгоритмді таңдау шешілетін мәселенің класына, сондай-ақ бастапқы деректердің құрамы мен көлеміне байланысты болатыны анық. Мысалы, гетерогенді деректерді жіктеу мәселелері шешім ағаштарының алгоритмін қолдана отырып, кластерлеу әдісін қолдана отырып, анық емес заңдылықтарды болжау немесе анықтау мәселелерін шешкен дұрыс.



Үлгіні анықтау алгоритмі жүйенің тиімділігін анықтайтын маңызды компонент екенін атап өткен жөн; дегенмен, аналитикалық мәселелерді сәтті шешуді жеңілдететін басқа да факторлар мен құралдар бар. Мысалы, Microsoft SQL Server ДҚБЖ-нің бірдей талдау қызметтерінде T-SQL және Data Mining үшін OLE DB үшін функционалды қондырмалар пайдалы болып табылады. OLE DB-де деректерді қалыптастыру қызметін және мазмұн түрлерінің арнайы бағанын, кесте бағанын қолдана отырып, Data Mining үшін OLE DB оқыту және болжау үшін кірістірілген таңдау мәлімдемелерін (nested cases) және ашылмаған (nested cases) пайдалануға мүмкіндік береді.

