Дәріс 13

Үлкен деректерді талдаудың заманауи бағдарламалық құралдары

Қазіргі уақытта сараптық жүйелерге мұқтаждықтары бар өте көп компаниялар бар, бірақ қажет бағдарламалық қамсыздандырудың қымбаттылығы мен шамадан тыс күрделілігі көп жағдайда жеке сараптық жүйе құру идеясынан бас тартқызып, барлығына белгілі қарапайым Excel құжатын колдануға итермелейді. Сонымен қатар қызметкерлерді оқытуға кетеін қосымша шығындар, деректерді сақтаудың қымбат жүйелерін қолдау және т.б. Мұндай кезде көмекке ыңғайлы, лайықты жүйелер болады, соның бірі – Rapid Miner жүйесі.

Rapid Miner – Data Mining үшін құрылған құрал, оның негізгі идеясы – сарапшы өзінің жұмысын орындау кезінде бағдарлама жазбайды.

Rapid Miner нәтижелері қандай да бір алгоритм немесе алгоритм жиынының «ғажайып мүмкіндіктерінен» емес, көптеген жағдайда деректердің дайындық деңгейіне тәуелді болады. Rapid Miner-дағы жұмыстың шамамен 75% деректерді жинаудан тұрады, ол талдау құралдарын қолдануға дейін орындалады. Құралдарды сауатсыз пайдалану компания әлеуетін мағынасыз шашуға, кейде миллион доллар жоғалтуға әкеледі.

Rapid Miner, Деректер қоймасы және CRM саласындағы әлемдегі танымал сарапшы Херб Эдельштайнның (Herb Edelstein) пікірі: «Two Crows компаниясының жақындағы зерттеуі Rapid Miner әлі де дамудың алғашқы сатысында тұрғанын көрсетті. Көптеген ұйымдар бұл технологияға қызығушылық танытады, бірақ кейбіреуі ғана мұндай жобаларды белсенді түрде енгізіп жатыр. Тағы бір маңызды сәт анықталды: Rapid Miner іске асыру процессі тәжірибе жүзінде күтілгеннен де қиын болып шықты. IT-командалары Rapid Miner құралдары қолдануда қарапайым деген мифпен қызықты. Мұндай құралды терабайтты деректер қорында іске қосу жеткілікті және пайдалы ақпарат бірден пайда болады деп болжанады. Шын мәнінде, сәтті Rapid Miner жобасы қызмет мәнін, деректерді және құралдарды білу, сонымен қатар деректерді талдау процессін түсінуді талап етеді» [31]. Осылайша, Rapid Miner технологиясын қолданудан бұрын, әдістер жүктейтін шектеулерді және онымен байланысты болатын сыни сұрақтарды мұқият талдау, сонымен қатар технологияның мүмкіндіктерін дұрыс бағалау қажет.

Сыни сұрақтарға келесілер жатады:

1. Технология қойылмаған сұрақтарға жауап бере алмайды. Ол сарапшыны алмастыра алмайды, тек қана оның жұмысын жеңілдетіп жақсарту үшін қуатты құрал береді.

2. Rapid Miner бағдарламасын құрудағы және пайдаланудағы қиындық. Бұл технология мультипәндік облыс болғандықтан, Rapid Miner қамтитын бағдарлама құру үшін әртүрлі саладағы мамандарды іске қосу қажет, сонымен қатар олардың арасындағы сапалы қарым-қатынасты қамтамасыз ету керек.

3. Пайдаланушылардың біліктілігі. Rapid Miner әртүрлі құралдары интерфейстің әртүрлі дәрежесіне ие және пайдаланушының белгілі бір біліктілігін талап етеді. Сондықтан бағдарламалық қамсыздандыру пайдаланушының дайындық деңгейіне сай болуы керек. Data Mining-ті қолдану пайдаланушы біліктілігін жоғарылатумен үзіліссіз байланыста болуы қажет. Бірақ бизнес-процесстерді жақсы білетін Data Mining мамандары қазіргі уақытта аз.

4. Деректер мәнін жақсы түсінусіз пайдалы мәліметтерді алу мүмкін емес. Анықталған шаблондар немесе тәуелділік моделі мен түсіндіруін мұқият таңдау қажет. Сондықтан мұндай құралдармен жұмыс істеу пәндік облыстағы сарапшы мен Rapid Miner құралы бойынша маманның тығыз қарым-қатынасын талап етеді.

5. Деректерді дайындау қиындығы. Сәтті талдау деректердің сапалы алдын ала өңделуін талап етеді. Сарапшылар мен деректер қорын пайдаланушылардың айтуы бойынша, алдын ала өңдеу процессі Data Mining-процессінің жалпы 80% пайызын алуы мүмкін. Осылайша, технолгия өзіне-өзі жұмыс істеу үшін, алдын ала деректерді талдау мен модельді таңдап, оны дұрыстауға кететін өте көп уақыт пен күш керек болады.

6. Жалған, сенімді емес немесе пайдасыз нәтижелердің үлкен үлесі. Data Mining технологиясының көмегімен шын мәнінде өте бағалы ақпаратты табуға болады, ол алдағы жоспарлауда, басқаруда, шешім қабылдауда айтарлықтай артықшылық бере алады. Бірақ, Data Mining әдістері көмегімен алынған нәтижелер, жеткілікті жиі жалған және мәні жоқ қорытындылар қамтиды. Көптеген мамандар Rapid Miner құралы статистикалы дұрыс емес нәтижелердің үлкен санын бере алады деп айтады. Осындай нәтижелер пайызын төмендету үшін алынған модельдердің дұрыстығын тестілі деректерде тексеру қажет. Алайда жалған қорытындыларды толық болдырмау мүмкін емес.

7. Жоғарғы құны. Сапалы бағдарламалы өнім әзірлеуші тарапынан бірталай еңбек шығындарының нәтижесі болып табылады. Сондықтан Rapid Miner бағдарламалық қамсыздандыруы дәстүрлі түрде қымбат тұратын бағдарламалық өнімге жатады.

8. Репрезентативті деректердің жеткілікті санының болуы. Rapid Miner құралдары статистикалыдан қарағанда, теория жүзінде ретроспективті деректердің қатаң нақты бір санын талап етпейді. Бұл ерекшелік сенімсіз, жалған модельдерді анықтау себебі, және соның негізінде қате шешім қабылдау нәтижесі бола алады. Табылған білімдердің статистикалы мәнін бақылауды іске асыру қажет.

Rapid Miner жүйесінің Statistica 8 ортасындағы мүмкіндігі

StatSoft компаниясымен STATISTICA Rapid Miner жүйесі әзірленген болған, ол деректерді талдаудың әмбебап және жан-жақты құралы ретінде жобаланған және іске асырылған – әртүрлі деректер қорымен өзара қарым-

қатынасынан бастап, графикалы-бағытталған тәсіл дегенді іске асыратын дайын есептерді құруға дейін. Берілген пакеттің барлық мүмкіндігін сипаттау үшін өте көп уақыт қажет болады, сондықтан бұл пакеттегі негізгі Data Mining құралдарына сипаттама береміз.

– Бағдарламалық қамтамасыз ету нарығындағы Data Mining әдістерінің толыққанды пакеті;

– Дайын шешімдердің үлкен жиыны;

– MS Office-пен толық біріктірілген ыңғайлы қолданушы интерфейсі;

– Барлау талдаудың қуатты құралы;

– Үлкен көлемді ақпаратпен жұмыс істеуге арналған толық оңтайландырылған пакет кешені;

– Икемді басқару механизмі;

– Жүйенің көп міндеттілігі;

– Өте тез және тиімді өрістету;

– Ашық архитектура, автоматтандыру мен қолданушы қосымшаларын қолдаудың шектелмеген мүмкіндіктері (Visual Basic, Java, C/C++ өндірістік стандартын қолдану).

STATISTICA Rapid Miner жүйесінің жүрегі болып, Data Mining тапсырмаларына арнайы оңтайландырылған және 300 негізгі процедура қамтитын, Data Mining процедурасының браузері болып табылады, және олардың арасындағы логикалық байланыс пен деректер ағынын басақару, жеке талдау әдістерін құруға мүмкіндік береді.

STATISTICA Rapid Miner жүйесінің жұмыс кеңістігі төрт негізгі бөлімнен тұрады:

1. Data Acquisition – Деректерді жинау. Бұл бөлімде қолданушы талдауға арналған деректер көзін анықтайды, ол деректер файлы немесе деректер қорынан сұраныс болуы мүмкін.

2. Data Preparation and Cleaning Transformation – Деректерді дайындау, түрлендіру және тазалау. Бұл жерде деректер түрленеді, сүзгіден өтеді, топтастырылады және т.б.

3. Data Analysis, Modeling, Classification, Forecasting – Деректерді талдау, модельдеу, классификациялау, болжау. Бұл жерде қолданушы браузер немесе дайын модель көмегімен, болжау, классификациялау, модельдеу және т.б. сияқты қажетті деректерді талдаудың түрін беруі мүмкін.

4. Reports – Нәтижелер. Бұл бөлімде қолданушы талдау нәтижелерін көру, түрін беру және жөндеуі мүмкін (мысалы, жұмыс кітабы, есеп беру немесе электронды кесте).

RapidMiner - де тізбек операторлары интерактивті граф түрінде ұсынылады және XML тілінде (eXtensible Markup Language, негізгі тіл жүйесі) түрінде көрсетіледі. Осы жүйесі Java тілінде жазылған және AGPL version 3

лицензиясы арқылы қолданылады. Барлық негізгі функцияларына қатынау үшін Java API арқылы мүмкіндік алады.