

Adatbányászat és adatvizualizáció

Dombi József

1. Adatbányászat

1.1 Bevezetés

1.1.1 Az információs túlterhelés

Az információs társadalom létrehozása szinte minden fórumon elsődleges téma. A hálózatok robbanásszerű terjedése miatt az információs korszak kezdetének nevezik korunkat, ugyanis a lokálisan rendelkezésre álló információk a hálózatok segítségével mindenütt elérhetővé válnak. A globális információáramlásnak a felállított biztonsági eljárások sem tudják útját állni. Mindenesetre a hálózaton elérhető információk olyan gazdagságával állunk szemben, ami a kezelhetőség szempontjából új problémákat vet fel. Ha ehhez hozzávesszük a nem ingyenes adatszolgáltatók által nyújtott információ-özünt, tényleg elveszett embernek hihetjük magunkat.

Az információt nemcsak úgy lehet elrejtteni, ha titkosítjuk, hanem ha mértéktelen mennyiségben áll rendelkezésre adat.

Egy történettel megvilágítva a fentieket, tegyük fel, hogy a király titkos üzenetet akar elküldeni, ekkor folyamodhat ahhoz, hogy rébuszokban beszélve küldöncével egy étel elkészítésének módját taglalva adja tudtára tervét: "Ne bárányból készítsd az ebédet, elég, ha kappannal fogadod vendéged, amit kellőképpen fűszerezel és hozzá savanyúság helyett kompótot tálalsz ..." Az ellenfélnek - a küldöncöt elfogva és megtalálva az üzenetet - csak arra kell koncentrálnia, hogy a megfelelő értelmezést megadja. A helyzetet úgy jellemezhetnénk, hogy a küldönc receptek tucatjait esetleg egy szakácskönyvet visz magával és csak azt az információt tartja a fejében, hogy melyik receptet kell átadnia a király bizalmasának. Az ellenség megkaparintva a recepteket, szinte reménytelen helyzetbe kerül a megfejtést illetően.

A mai helyzet szerint a példában szereplő mennyiségeket nagyságrendekkel megnövelhetnénk. A király üzenete esetünkben valamilyen törvényszerűség (társadalmi, pl. fogyasztói szokás), aminek megfejtése a feladat.

Az "adatbányászat" szó azért terjedt el, mivel a valódi bányászat esetén is az érték csak töredéke a megmozgatott anyag mennyiségének. Először is össze kell gyűjteni azokat az anyagokat, amikből kinyerhető az érték. Erre a fázisra jellemző a hálózatok építése, ill. az adatáruházak létrehozása. Ma mindenki az adatok felhalmozásának szükségességéről beszél és csak másodlagosnak tekintik a feltáró folyamatot, aminek az a veszélye, hogy a nem-feladatorintált megközelítés sok felesleges munkát eredményez. A felhalmozás még csak az előkészületi fázis, ettől még nem lesz meg az adatbányászat eredménye: az ásvány ill. a "kincs".

A gazdasági haszon miatt nem egyetlen ember, vagy szervezet indul az információ felderítésének, valamint megszerzésének versenyében.

1.1.2 Az adatelemzés fontossága

Az adatbányászat nem csak a kecsegtető haszon miatt fontos, hanem egyszerűen kényszer is, mert a világban felgyorsult folyamatokra adatbányászati eszközök használata nélkül a szervezetek már nem tudnak megfelelően reagálni. Ismét hasonlattal élve olyan ez, mint amikor az utcán közlekedő autók sebességkorlátozása megszűnik és a gyalogosok közlekedése lehetetlenné válik, vagy mint amikor a számítógépes játékoknál a képernyőn mozgó objektumok sebességét többszörösére növeljük és nem lehetséges hatékony beavatkozás. Valahogy így kell elképzelni a társadalmi folyamatok felgyorsulása közepette való létünket is.

Az adatok közötti tájékozódás szinte minden vállalat és szervezet számára létszükségletté válik. Nemcsak a profit nagysága függ ettől, hanem sokkal több, az életben maradásé. Szükségük van szolgáltatásaik, termékeik, piacuk, pénzügyi helyzetük pontos meghatározására nemcsak lokális környezetükben, hanem sokkal szélesebb körben, a globalizálódó világban is. (Tipikusan ilyen feladat a termékmenedzseré, a vásárlói szokásokat feltáró elemzőé, a piackutatóé, a gazdasági stratégiát meghatározóé, stb.) Nemcsak a törvényszerűségeket kell feltárni, a vezető számára néha sokkal fontosabb a rendellenességek megtalálása a folyamatok megértése.

A nagy szervezetek létben való fenyegetettsége azt is jelenti, hogy statisztikusokat alkalmaznak adatelemzés céljából, akik a statisztikához ugyan értenek, de a vállalat irányításához nem, és alkalmazásuk csak részben oldja meg a problémát. Az eredmények értelmezésében ez mindig gondot okoz, nem beszélve arról, hogy bizonyos fontos diszciplínák nem is tartoznak a statisztika területéhez, így pl. a tanulás, neurális hálók döntési fák stb. alkalmazása.

Az európai szervezetek (de nemcsak az európaiak) eddig képtelenek voltak felvenni a versenyt a kihívásokkal. Új típusú vezetői szemlélet kell ahhoz, hogy a helyzetet meg lehessen menteni és - a következő generáció színrelépéséig - a középvezető rétegnek kell ezt a problémát megoldani. Ők csak közvetve válnak felelőssé a döntésért azzal, hogy elvégzik a döntési lehetőségek feltárását.

Az adatbányászat eszköz az információ-rengetegben való tájékozódáshoz. Segítségével gyorsabb, jobb javaslatok készíthetők elő és segíti az ügyintézését is. A legnagyobb probléma, hogy az alkalmazók nem adatelemző specialisták, ezért a jelenleg rendelkezésre álló eszközök alkalmatlanok számukra. Mint ahogy a rádióhallgatónak sem kell ismernie a rádió működését, ahhoz hogy használja, ezért arra van szükség, hogy elkészüljenek a felhasználót messzemenően figyelembe vevő újszerű programok, lehetőleg azt a fejlődési fázist is kihagyva, amikor az eszköz jóságát annak bonyolultsága igazolta. Jól ismert tény, hogy az egyszerűség marketing szempontból jelentős előny. (Az automata fényképezőgép sokkal nagyobb piaci szegmenst birtokol, mint a professzionális gépek.)

1.2 Mi az adatbányászat?

Az adatbányászat, mint önálló diszciplína a statisztika mellett úgy jöhetett létre, hogy könnyen használható elemző eszközöket kellett biztosítani az üzleti szakértőknek, a piackutatóknak és a gazdasági és stratégiai tervezés érdekelteknek. Ezeknek az eszközöknek a közös lényegi vonása, hogy használóikat segítik az adatelemzésben és az adatelemzés részletei helyett, inkább az üzleti problémákra lehet koncentrálni.

Az adatáruházak manapság az információfeldolgozás összes formájával foglalkoznak, különösen nagy hangsúlyt fektetve az adatbányászat irányzatára.

Az adatbányászatot nem könnyű definiálni; régebben tudás-kezelésnek (knowledge management), vagy tudás-technológiának nevezték. Az adatbányászat általánosan elfogadott definíciója: ismeretlen minták és összefüggések keresési folyamata a teljes adatbázis alapján. A régi keresési módszerek, amelyek az adatbázis alapján bizonyos kritériumnak való megfelelést vizsgálnak (speciális tényre vagy tényekre vonatkoznak), nem azonosak a mai

adatbányászatával. Az adatbányászat eljárása sokkal inkább hasonlít egy törvényszéki nyomozó tevékenységéhez, aki minden lényegesnek tűnő tényt megvizsgál és keresi az összefüggő motívumokat, hogy felderítse a bűncselekményt. Nem használ lekérdező nyelvet a speciális adatok keresésére, hanem inkább megvizsgálja az összes tényezőt, hogy kiderítse, van-e olyan motívum, vagy összefüggés, aminek értelme (jelentése) van.

A bányászat helyett tehát a nyomozás, felderítés legalább olyan jó szó, mert a tevékenységre utal. Az adatbányászat feladata olyan eszkörendszer összeállítása, kifejlesztése, ami segíti a feltárási folyamatot.

Az adatbányászat olyan keresési módszert alkalmaz, amellyel kialakul a minták egy bizonyos sorrendje. Az adatbányászat software-termékei speciális eszközök, amelyekkel a felvetett kérdésekre adhatunk választ azzal, hogy lehetővé tesszük a felhasználóknak, hogy kereső eljárásokat hajtsanak végre. Az adatbányászat így tartalmazza a lekérdező (SQL) nyelvek fejlesztését is.

Az adatbányászat során a nagy adatbázisokból olyan kapcsolatokat, motívumokat és jellegzetességeket keresnek, amelyekről előzően nem tudták, hogy léteznek, vagy nem is voltak láthatóak. A megtalált kapcsolatokat ugyan elfogadhatták a szakemberek vagy értékesítők, de mielőtt alkalmaznák, először ki kellett próbálni, esetleg pontosítani, finomítani kellett őket.

Az adatbányászat eredménye olyan új információ vagy tudás, amely lehetővé teszi a felhasználó közösségeknek, hogy hatékonyabbak legyenek. Az adatbányászat nehézsége, hogy néhány összefüggő tény feltárása miatt hatalmas adatbázist kell feldolgozni. Mint ahogy nincs két egyforma bűnügy, úgy ugyanazt az algoritmust és keresési kritériumot, amit egyszer használtunk, valószínűleg nem lehet már újra pontosan ugyanolyan módon használni.

Összefoglalva: az adatbányászat egy olyan információ feldolgozó eljárás, amely megmutatja a válaszokat azokra a kérdésekre is, amelyeket gyakran még fel sem tudunk tenni. Ahelyett, hogy egy relációs adatbázisnak hagyományos lekérdező nyelven azt mondanánk: "Menj és keresd meg azokat az embereket, akik ebben az évben ablakredőnyt vásároltak és valamivel később ágyneműt is vettek!", az adatbányászatban a kérdés így hangzik: "Találd meg az összefüggő vásárlási mintákat!". A válasz pedig: "van egy minta, ami az idő x százalékában jelenik meg, mégpedig akkor, ha valaki ablakhoz szükséges alkatrészeket vásárol (nemcsak redőnyt) és 1-3 hónapon belül ágyneműt is vásárol, a következő négy hónapon belül még bútort is vesz". Ha egy speciális összefüggésre kérdezzük rá, akkor pontos, de használhatatlan információt kapunk. Ha viszont olyan kapcsolatokra kérdezzük, amelyeknek a létezéséről még nincs tudomásunk, sokkal jelentősebb összefüggésre találhatunk, aminek üzleti értéke van. Számos technológiát kell alkalmazni ahhoz, hogy az adatbányászat működőképes legyen.

Először is, az egyik legfontosabb feladat egy adatáruház létrehozásának vállalása. Az adatbányászatnak képesnek kell lennie az összes adat megvizsgálására, amihez egy adatraktárt kell készíteni. Az adatbázisok adatáruházzá való átalakítása csak az első, kezdő lépésnek lehet tekinteni az adatbányászat megvalósításában.

Másodszor, léteznie kell egy jegyzéknek az adatbázisok tartalmáról, azaz egy metainformációs rendszert is létre kell hozni. Ez azért szükséges, hogy a használók (elemzők) tudják milyen adatok állhatnak rendelkezésükre. Az olyan információs rendszer, ami az üzleti adatoknak csak egy részét teszi elérhetővé, valójában értéktelen. Az adatok hatékony feldolgozásához a legkülönbözőbb forrásból származó adatokat is ismernünk kell. Az adatbányászati eszközöknek ténylegesen képesnek kell lenniük az adatáruházi és bármilyen más, pl. a szerveren szétszórta adat megkeresésére.

Harmadszor, olyan eszközökkel kell rendelkezni, amelyek kivitelezhetővé teszik az adatbányászati technológiát. Számos eszköz áll rendelkezésre, amelyek különböző kategóriákba sorolhatók pl. a legközelebbi szomszéd algoritmus, a döntés-fák és az adatok vizualizációja. Néhány eszköz az ipar sajátos területeire specializálódik, mint pl. a pénzügyre vagy a biztosításra, kihasználva a sajátos üzleti modelleket és speciális, jól meghatározott összefüggéseket keresve.

Az általános adatbányászati eszközök is kezdenek megjelenni, amelyek ugyan nem speciális üzleti ágakra irányulnak, de meghatározott összefüggéseket keresnek. A neurális hálózaton alapuló technikák rendkívül hasznosnak bizonyulnak. A humán kérdésselvetést meghaladó eljárás jött létre alkalmazásukkal, ami csupán a tanulás sikerességére koncentrál. De ebbe az osztályba tartozik a döntési fák konstruálása is táblázat alapján. Az adatbányászatban ezen technikák megvalósítása egy új információfeldolgozási eljárást jelent. A SQL és a statisztika elavultnak tűnik hatékonyságuk tükrében.

1.3 Adatbányászati technológiák

Az adatbányászat szempontjából a relációs adatbázisok jelentik a kiinduló pontot. Az adatbányászat sikere részben a már meglévő kapcsolatoktól függ, újabb kapcsolatok pedig könnyen beépíthetők. Az adatbázisokból visszanyerhető információ azonban nem szükségszerűen az adatbázis szerkezetéből származik, hanem az összefüggések elemzéséből. A tények felderítésére két különböző technikát alkalmaznak, a származtatást és a következtetést (dedukció, indukción).

A dedukciós módszerek általában az összes relációs adatbáziskezelő rendszerben rendelkezésre állnak (DBMS), míg az induktív módszerek nem használhatók fel közvetlenül. A dedukció mindig olyan információkat ad, amelyeknek be lehet bizonyítani a térszerűségét, míg az indukción olyanokat, amelyekről elfogadható, hogy valamilyen valószínűséggel igazak, de nem szükségszerűen bizonyíthatóak.

Az adatbányászatban induktív módszereket kell alkalmazni. Nagy mennyiségű adatot kell megvizsgálni, hogy eljussunk az eredményekhez, amit csak új technológiák alkalmazásával lehet végrehajtani. Ezek a technológiák részben a relációs DBMS rendszerek továbbfejlesztéseként állnak rendelkezésre. Az adatok típusától függően különböző eljárások léteznek. A legfőbb probléma az, hogy a hagyományos SQL módszerek nem használhatók adatbányászatra jelenlegi korlátaik következtében és így teljesen új fejlesztések szükségesek ennek a folyamatnak a végrehajtásához.

Az adatbányászati eszközöket a feldolgozó algoritmusok és eljárások szerint négy fő típusba lehet sorolni:

- egyesítés, vagy összekapcsoló elemzés asszociáció,
- sorrendi minták
- csoportosítás (clustering),
- osztályozás,

1.3.1 Az egyesítés, vagy összekapcsoló elemzés

Az adatbányászati asszociációk általában arra irányulnak, hogy az adatbázisból kikeressék az összes olyan tranzakciót, amelyek nagy valószínűséggel ismétlődnek. A kérdésnek ehhez a típusához egy olyan algoritmus szükséges, amely megtalálja az összes olyan szabályt, ami az események vagy adatok egy készletét összefüggésbe hozza egy másikkal. Az ilyen típusú folyamatra jellemző eredményeket a kiskereskedelmi példák használata mutatja legjobban. Tipikus válasznak tekinthető a következő: Azok közül, akik hordozható számítógépet vásárolnak, 78 % további kiegészítő termékeket is vásárolni fog. Ezt a típusú feldolgozást minden olyan relációs adatkészletknél lehet használni, amelyek standard SQL állító logikát használnak.

Az asszociatív algoritmusok célja, hogy bővítsék az SQL lehetőségeit. Az algoritmusoknak nagyon alkalmazkodóknak és dinamikusoknak kell lenniük. Szabályok szerint kell megtalálni a mintát, miközben változhat a megvizsgált adatkészlet és ennek megfelelően az asszociációs szabályok ill. a bennük előforduló százalékok is változnak. Ezeket a szabályokat aztán rendezni lehet a gyakoriságok szerint, hogy a felhasználó

megtalálja a legjobb lehetséges jelenségeket az üzleti stratégiák és termékelhelyezések szempontjából. Pl. Az élelmiszerárúházakban meghatározó, hogy az italvásárlók nagy százalékban sós süteményt is vásárolnak, ha az üzletben ugyanazon az útvonalon van a két termék. Ma már sokrétű asszociációs algoritmusok és feldolgozó eszközök állnak rendelkezésre, amelyek mind az ipar, mind pedig az üzleti folyamatok területén használatosak. Az általánosított algoritmusok jók ugyan, de legtöbbször nem biztosítják a kívánt pontosságot a versenyképes üzleti gyakorlatban.

1.3.2 Sorrendi minták

A sorrendi minták ugyanazokra az alapadatokra támaszkodnak, mint az asszociatív eljárások, azzal a kiegészítéssel, hogy az adatokat egy meghatározott időtartam szerint kell összegyűjteni, ami a rendszerből kigyűjtött tételek és tranzakciók egy ún. történeti adatkészlete, az eredmény pedig az ismétlődésre legnagyobb valószínűségű mintákat tartalmazza. A sorrendi mintákat az üzleti életben leginkább mintaelemzésre használják. Ez a leggyakoribb példa arra, hogy az adatbányászat hatékonyságát bemutassák.

A sorrendi mintákkal az a probléma, hogy nagyon sok használhatatlan eredmény keletkezhet. Az események sorrendjében nagy valószínűsége van annak, hogy egy bizonyos idő elteltével az első és a második esemény egy olyan mintát mutat, ahol a harmadik esemény sohasem fordul elő. Az üzletelemző fő feladata a szabályok finomítása az ismétlődő végrehajtások során, beállítva a minimum és maximum százalékos küszöbököt.

A sorrendiségnek van még egy másik megközelítési módja is, ami az üzleti életre hatással van. Ezt a technikát hasonló sorozatoknak nevezik. A sorrendi mintáknál az eredmény az események időbeli lefolyása. A hasonló sorozatoknál az időbeli események sorrendje hasonlít egy másik sorrendiséghez. Például a kiskereskedelemben olyan üzleteket keresünk, ahol hasonlóak az eladási árképzési stratégiák, vagy olyan raktárakat, amelyeknél hasonló ármozgással dolgoznak.

1.3.3 Csoportosítás (Clustering)

Az adatbányászat során néha még hipotézisünk sincs arra, hogy miképp tegyük fel a kérdést. Ezekben az esetekben csoportosító algoritmust kell használni, hogy felfedezzük a jellegzetesség egy eddig ismeretlen, vagy csak sejtett formáját. A csoportosító példák hiányelemzések, vagy rokon viszonyban álló csoportelemzések. Vannak eljárások, amelyek olyan csoportot találnak, ami néhány gyakori osztályt bont részekre. A csoportosító folyamatok néhány olyan sajátos eseményen alapulnak, mint pl. amikor egy jó vásárló megszűnteti hitelkártyáját.

Ebben az esetben a vevő típusát meghatározó szabályok ismeretlenek. Ahhoz, hogy ennek a vevőnek a típusát meghatározzák, a csoportosítás (clustering) képes olyan szabályokat létrehozni, amelyek lehetővé teszik a társaságnak, hogy megelőzze a fent említett hitelkártya törlését. A "clustering"-et irányítatlan tanulásnak nevezik.

1.3.4 Osztályozás

Az osztályozás, vagy más néven (irányított) tanuló algoritmus, olyan eljárás, amelynél példák alapján kell az algoritmusnak a szabályokat megtalálni. A szabályok alapján pedig az adatállomány bővítése során létrejövő eseményeket lehet jellemezni, ill. események bekövetkezését lehet megjósolni. Az adatbányászatban legfőképpen az üzleti haszon alapján kell meghatározni a szabályokat. Pl. feladat a nyereség meghatározása a vállalat egyéb jellemzőinek felhasználásával, a megoldás pedig a szabályrendszer megalkotása. Az

osztályozás a relációs adatbázisban nagyon összetett folyamattá válhat, mert gyakran sok táblázatot és tulajdonságot kell megvizsgálni. Így aztán nincs adott nevük vagy tulajdonságtípusuk sem. A leggyakoribb üzleti példa az osztályozásra a hitelkártya jóváhagyó eljárás. A legnevezetesebbek a perceptron, back-propagation és a döntési fákra (ID3, C4.5) vonatkozó algoritmusok.

1.4 A kereső eljárások

Az adatbányászati módszerek típusai után vizsgáljuk meg a kereső algoritmusokat. Hiába végzünk el egy részletes keresést, az adatbázis méreteinek köszönhetően ezt mégsem mindig célszerű megtenni a feladat komplexitása miatt. Hatékony kereső algoritmusokra van szükség. A legtöbb adatbányászati megoldás klasszikus logikai lekérdezést hajt végre, s aztán a lekérdezéseket módosítja iteratív módon, az üzleti célnak megfelelő határérték eléréséig. Ezekben a megközelítésekben számos különböző típusú stratégia létezik. Nincs legjobb megközelítési mód, a feladattól függ, mikor melyiket célszerű alkalmazni.

A keresés kezdeti megközelítésének függvényében beszélhetünk felfelé vagy lefelé irányuló folyamatról. A felfelé irányuló megközelítést adatvezérelt megközelítésnek nevezzük. Mint a szakértői rendszereknél, ez egy adatbázissal és egy kezdeti szabállyal indul, majd finomító eljárást hajt végre, mindaddig míg a határértéknek megfelelően nem tartalmaz olyan adatot, amely megszegi a szabályt. Egyszerűbben kifejezve, megvizsgálja az adatot, megkeresi benne az összes egyedi elemet, amely megfelel a szabályoknak és aztán ezekből az elemekből meghatározza az eredményt.

A lefelé irányuló megközelítés szabály-vezérelt keresés és néhány leíró művelet alkalmazásával kezdődik, ami elindít egy válogatást. Kezdetben, az összes adatból indulunk ki, a finomítási eljárások folyamán végül azokat az elemeket kapjuk, amik nem tesznek eleget a szabályoknak. Az eljárás addig folytatódik, amíg olyan adatmennyiséget kapunk, ami megfelel a kívánt értékhatároknak.

A felfelé irányuló megközelítés csak a kriteriumoknak megfelelőket válogatja ki, hogy végül megtalálja az események mintáját és sorrendjét. A lefelé irányuló megközelítés pedig hatalmas készlettel indul és csak a kriteriumoknak nem megfelelőket válogatja ki.

Mindkét megközelítésnél finomítási és műveleti eljárásokat hajtunk végre, amely új adatbázist, vagy adatbázis leírást eredményez. Aztán újrakezdődhet a folyamat egy másik adatbázisra ugyanazokat az eljárásokat alkalmazva. Ez az eljárás az új szabályok megtalálása mellett az adatbázist szem előtt tartva a keresés grafikonját is felépíti segítve ezzel az előírt határértékek figyelembe vételét. A folyamatot könnyebb megérteni egy valós példán. A szabályalkotás és finomítás bemutatásához próbáljuk meg elképzelni a következő műveleteket:

1. Vizsgáljunk meg minden vásárlást, ami egy élelmiszerüzletben történt és válasszuk ki bármelyik két elemet ebből az egyéni vásárlásból.

2. Vizsgáljuk meg az összes többi vásárlást. Nézzük meg, hogy a két kiválasztott kezdeti elemet ugyanakkor vásárolták-e. Ha ezen vásárlások száma az előírt határértéket eléri, elkezdhetünk egy eredménykészletet építeni.

3. Vegyük az egyik első elemet és párosítsuk össze egy másikkal az első vásárlásból. Ezután indítsuk újra az eljárást.

A kereső eszközök a mesterséges intelligencia kutatásának középpontjában állnak. Különböző megközelítéseket és stratégiákat alkalmaznak, hogy megelőzzék a részletező keresést és gyorsan egy konklúzióhoz érkezenek.

1.5 Néhány jelenlegi eszközzel

A mai adatbányászati eszközök nagy része különféle intelligens mintafelismerési technológiát használ, mint például neuronhálózatok, döntési fák, indukciós szabálykeresés, fuzzy logika.

Ellentétben a tipikus algoritmos adatbányászattal, ami a korábban említett AI technikákon alapul, könnyebben használhatók azok az adatbányász programok, amelyeknek jól meghatározott profiljuk van. Ilyen például az *IVEE Development's Spotfire 1.0* rendszer, amely minden statisztikai változóhoz egy beállító csúszkát rendel, amelyekkel a változók egymásra hatása megvizsgálható és így az érdekes minták felfedezhetők. Arra is lehetőséget ad, hogy valamilyen háttérinformációval (pl. egy térképpel) együtt mutassa be az adatokat, hogy még könnyebb legyen felfedezni a mintákat.

Az ilyen rendszerek mégis korlátozottabban használhatók, mint a hagyományosak – főként ha a neuron hálózatok flexibilitásához hasonlítjuk – mivel a használók a saját különleges adatbázisra épülő tudásukra korlátozottak. Másrészt, ahogy Chris Alberg az IVEE Development-től kimutatta, a legtöbb szakember jobban ismeri a saját üzletét, mint egy neurális hálózat. Valójában a neurális hálózat tanításának eredménye erősen függ attól az egyéntől, aki a tanítást végzi.

A Cognos, az elemző folyamatok eszközeinek egyik vezető képviselője. A *Right Information System* egy neurális hálózaton alapuló software üzleti mintákra és előrejelzésekre.

A SAS Intézet által kifejlesztett *DMINE* nevű termék a neurális hálózatok, a döntési fák és vizuális technikák kombinációja és az üzleti életre irányul. Manapság a döntéskészítők sok időt töltenek az adatbázisok lekérdezésével. Annak érdekében, hogy az adatbányászat elérhetővé tegyük szélesebb körű közönség számára és segítsük a döntéshozókat, hogy ezt a rendszert a saját PC-jükön futtathassák, a SAS Intézet egy olyan módszert támogat, amely öt alapvető lépésből áll: kiválasztás, megvizsgálás, kezelés, mintaelőállítás, becslés. A módszer az adatbázis egy részét használja, hogy csökkentse az eljárások futási idejét. Az adatok vizualizációja segíti a használót, hogy a helyes adatállományrészletet válassza ki. Ha az adat túl összetett a grafikus bemutatásra, akkor a hagyományos statisztikai módszerek használhatók, mint pl. csoportosítás, összefüggés-elemzés. Az ilyenfajta adatvizsgálattal a felhasználó megtisztíthatja és aktuálisra teheti a kiválasztott részt, aztán futtathatja a kereső folyamatot, amely meghatározza azokat a legfontosabb ismertetőjeleket, amelyek a megadott adatállományhoz tartoznak. A folyamat utolsó lépéseként a használó értékeli a mintákat és a valósággal szembeállítva ellenőrzi az értékét.

Az Isoft az *Alice* nevű termékének egy leegyszerűsített változatát forgalmazza sikeresen. Az *Alice* széles körű statisztikai algoritmusokat használ. A program a minta eredményeit döntési fákban jeleníti meg, lehetőséget adva a használónak, hogy megértse az adatbázis belső viszonyait és könnyen ellenőrizhesse a feltevéseket. A döntési fák készítése nagyon vonzóvá teszi az adatbányászati folyamatát, ezzel az adatok nagy tömegét is eredményesen tudja kezelni. A döntési fák segítségével a felhasználó szét tudja választani, vagy egybe tudja olvasztani a csomópontokat, lecsökkentheti, vagy kibővítheti az ágakat és meghatározhatja a paraméterek számát egy fában. Ráadásul a felhasználó könnyen kiválaszthatja azokat az ágakat, amelyek érdeklik.

A vizuális adatbányászati lehetővé teszi az algoritmusok összekapcsolását a személyes tapasztalattal és tudással, így a statisztikai eredményeket könnyebben lehet fordítani életképes üzleti stratégiára. Ezen alapul a Cygron *DataScope* nevű rendszere, ami nem az algoritmusokra, hanem az emberi intuícióra koncentrálna interaktív *játszadózást* tesz lehetővé az adatokkal, mely közben értékes felfedezések szülehetnek. A folyamatok megértése itt fontosabb, mint egy szabályrendszer létrehozása. A következő rész a *DataScope* vizualizációjával foglalkozik.

Az adatbányász rendszerek meg tudják mutatni az időbeli változásokat és azt, hogy néhány változó hogyan befolyásolja a többit. Mégis az a mód, ahogy ezeket az információkat kifejezik gyakran misztikus hatású és néhány üzletember számára nehezen érthető. Szükség van egy olyan nyelvre, amely közvetlenül az üzleti élet képviselői számára transzformálja a gépi tudást.

2. Adatvizualizáció

2.1 Bevezetés

Az adatbányászati eljárásokat két fő szempont szerint lehet értékelni: a törvényszerűségek megtalálása, ill. az eredmények értelmezhetősége szempontjából. Az utóbbi az eredmények kognitív aspektusa, amely a vezetők számára sokkal jelentősebb, mint pl. egy nagy pontosságú statisztikai mintavételre alapuló szabály. A vizualizáció éppen ezért került a középpontba. A DataScope szoftver egyedül állómódon, vizuálisan valósítja meg a lekérdezést. 1997-ben Európai Információ Technológiai díjjal tüntették ki innovatív megoldásért. A továbbiakban a DataScope főbb jellegzetességeit ismertetjük.

2.2 Mi a DataScope?

A DataScope egy Microsoft Windows 3.1 alkalmazás, melynek segítségével adatbázisainkon hatékony elemzéseket végezhetünk. A szoftver grafikusán megjeleníti egy tetszőleges adatbázis tartalmát, és sokféle eszközzel támogatja az egyes adatbázis rekordok közötti viszonyok tanulmányozását, valamint a (pozitív vagy negatív értelemben) kivételes tulajdonságokkal rendelkező alternatívák kiválasztását. A DataScope használható grafikus online lekérdező rendszerként is. Különböző szempontok szerinti szűréseket (lekérdezéseket) is végezhetünk vele és a munkát a leszűrt adatokon folytathatjuk.

A rendszer segítségével az adatok közötti összefüggések sokkal jobban és gyorsabban feltérképezhetők, így hatékonyabb döntések hozhatók. Ezt egy példán keresztül mutatjuk meg. Tekintsünk egy adatbázist, amelynek rekordjai autók néhány adatát tartalmazzák. Az egyes autókról a következő adatok állnak rendelkezésre: név, ár, teljesítmény, fogyasztás, hengerűrtartalom és a használt üzemanyag fajtája. Ezekre a mezőkre a bemutatás során hivatkozunk.

Felsorolunk néhány tipikus kérdést, amivel a vezetők az adatbázishoz fordulnak és eddig csak lekérdező nyelv segítségével voltak megvalósíthatók:

- Melyik a legdrágább/legolcsóbb autó?
- Melyik a legdrágább/legolcsóbb dízel autó?
- Drága-e az autó?
- Melyek azok az autók, amelyeknek a fogyasztása kisebb, mint 7 l/100 km és ára alacsonyabb 20000 márkánál?
- Hogyan lehet néhány autót sok tulajdonság szerint könnyen összehasonlítani?
- Mi az ára egy közepes árú autónak?
- Mi az ára egy közepes árú dízel autónak?

- Igaz-e, hogy a dízel autók általában drágábbak, mint a benzinesek?
- A 20000 márkánál olcsóbb autóknak hány százaléka dízel?
- Hány autó elégít ki egy bizonyos feltételt?
- Hány autónak nincs az áráról adat?
- Mi a dízel és a nem dízel autók aránya az adatbázisban?
- Van-e kapcsolat a fogyasztás és a teljesítmény között? Melyek a kivételek?
- Hogyan található olyan autó, amelynek ára 20000 márka körül van, és teljesítménye minél magasabb?
- Hogyan lehet egyszerűsíteni a munkát, ha csak a dízel autókkal akarunk foglalkozni?
- Hogyan jelölhető ki az összes olyan autó, amelynek áráról nincs adat?
- Hogyan jelölhető ki lokálisan a 10 legnagyobb teljesítményű autó?
- Ha van egy diszkrét mezőnk, és pontozni tudjuk az egyes kategóriák jóságát, hogyan készíthetünk olyan numerikus mezőt, amelyben ezek a jósági értékek szerepelnek?

A fenti kérdések DataScoppal vizuálisan, egér mozgatóval megvalósíthatók és a válasz vizuálisan áll rendelkezésre.

2.3

A program nemcsak az adatbázis egyes rekordjainak elemzésére alkalmas. Lehetőséget biztosít csoportos kiválasztásra, szűrésre is. Bármely ablakban (azaz bármelyik tulajdonság szerint) kiválaszthatunk tetszőleges rekordokat. Ezt lokális kijelölésnek nevezzük. A lokális kijelölések összességéből az unió (*vagy*), vagy a metszet (*és*) operátor segítségével képezhető a globális kijelölés (**eredmény**). Így például könnyen kijelölhetjük az olcsó és kis fogyasztású autókat. A többi ablak is mutatja, hogy melyek a kiválasztott elemek, így ezután megvizsgálhatjuk ezek teljesítménymutatóit, vagy a relációs diagramok alapján megkereshetjük a kivételes tulajdonságokkal rendelkezőket. Készíthetünk egy új projektet is, amely csak a globálisan kijelölt rekordokat tartalmazza, ezzel kiszűrve az érdektelen alternatívákat és áttekinthetőbbé téve adatainkat.

2.4 Kijelölések, szűrések

Mielőtt a mezőket megjelenítő ablakok használatával foglalkoznánk, meg kell ismernünk néhány fogalmat. Mint a bevezetőből tudjuk, a DataScope segítségével nemcsak egyes rekordok tulajdonságait vizsgálhatjuk, hanem rekordcsoportokét is. Ehhez ki kell jelölnünk a vizsgálandó rekordokat. A kijelölésnek két típusa létezik: a lokális (feltételrendszer) és a globális (eredmény) kijelölés.

2.4.1 Lokális kijelölés

Bármelyik mező szerint kiválaszthatunk bizonyos rekordokat. Például, ár szerint kijelölhetjük a 15000 és 20000 márka közötti autókat, vagy az azonosító mező alapján bizonyos autókat (pl. az összes Audit és BMW-t), vagy az Üzemanyag mező szerint az összes dízel autót. Mivel akár minden rekordról egyesével megmondhatjuk, hogy az kijelölt-e (kivéve a diszkrét mezőket, ahol csak egész kategóriákat jelölhetünk ki), ez a kijelölés tetszőlegesen bonyolult kritériumok szerint történhet. Ezt a kijelölési formát a mező szerinti

lokális kijelölésnek nevezzük. Egyidejűleg több mező szerint is elvégezhetjük. Az egyes mezők szerinti lokális kijelölések függetlenek egymástól és bármikor módosíthatók. Minden mezőablak jelzi, hogy az általa képviselt mező(k) szerint milyen lokális kijelöléseket végeztünk. Az azonosító ablakoknak (ld. később) kitüntetett szerepük van, mivel ezek mindig az éppen aktív ablak szerinti lokális kijelölést jelzik.

A státussorban megjelenik az aktív ablakban megjelenített mező lokálisan kijelölt rekordok száma ill. az, hogy ezek az összes rekordnak hány százalékát képviselik. A figyelmünket felkeltő rekordok megjelölhetők egy, legfeljebb kétbetűs azonosítóval. A jelek az összes diagramon megjelennek, így a továbbiakban a rekordok könnyen nyomon követhetők. A jelölés lehetőséget ad a rekordok tulajdonságainak összehasonlítására is (pl. néhány megjelölt autó közül azonnal látszik, hogy melyik a legolcsóbb).

2.4.2 Globális kijelölés vagy eredmény

Globális kijelölést a lokális kijelölések együttes figyelembe vételével az Unió vagy Metszet művelet végrehajtásával hozhatunk létre. Egyidejűleg csak egy globális kijelölés létezhet. A globális kijelölés használatával meghatározhatjuk azokat az adatbázisrekordokat, amelyek egy adott feltételkombinációnak eleget tesznek. A metszetképzés az "és" típusú kapcsolatok megadására szolgál. Ha az előbb említett két lokális kijelöléssel rendelkezünk (15000 márkánál alacsonyabb árú autók és 7 liter alatti üzemanyagfogyasztású autók), akkor ezek metszetét véve megkapjuk az olcsó és kis fogyasztású autókat. (A metszetben szereplő autók mindkét feltételnek eleget tesznek.) Az unióképzés segítségével "vagy" típusú kapcsolatokat állíthatunk fel. Például, ha lokálisan kijelöljük a 15000 márka alatti autókat az árat képviselő ablakban, valamint a 7 liternél kisebb fogyasztásúakat a megfelelő ablakban, akkor az unióképzéssel megkapjuk az olcsó vagy kis fogyasztású autók csoportját. A lokális kijelöléseket bármikor módosíthatjuk és új globális kijelölést hozhatunk létre. Szükség lehet erre például, ha nincs olyan autó, amelyik megfelelne minden megadott feltételnek. Ekkor a metszetképzés üres halmazt eredményez. Ilyen esetekben megpróbálhatjuk bővíteni valamelyik lokális kijelölést, és újraképezni a metszetet. A státussorban megjelenik a globálisan kijelölt rekordok száma, és hogy ezek az összes rekordnak hány százalékát képviselik.

2.5 Adatbázismezők típusai

A DataScope segítségével megjeleníteni, vagy elemezni kívánt adatbázisok különféle mezőket tartalmazhatnak. A megjeleníteni kívánt adatbázis mezői három típusba sorolhatók: Azonosító, numerikus, diszkrét mező.

2.5.1 Azonosító mező

Ennek alapján történik az egyes rekordok azonosítása. (A példában ez az autók megnevezése.) Az azonosító mező az adatbázis rekordjainak egyedi azonosítására szolgál. Fontos, hogy értéke minden rekordnál különböző legyen, vagy legalábbis, csak igen ritka ismétlődések forduljanak elő. Egy személyek adatait tartalmazó adatbázisban a rekordok azonosíthatók a személyek nevével, vagy személyi számával, vagy akár mindkettővel (több azonosító mezőt is megadhatunk). A projektnek nem kötelező azonosító mezőt tartalmaznia, nélküle azonban a DataScope lehetőségei leszűkülnek tendenciák megállapítására. A konkrét rekordok vizsgálata nem lehetséges. Az azonosító mező tartalma listában jelenik meg (lásd az 1. ablakot a képen). A listában a mezőtartalom mellett az éppen aktív ablak által képviselt

mező tartalma is szerepel és a lista ez utóbbi mező szerint rendezett. (A képen az autók ár szerint vannak rendezve, az árak az autók neve mellett láthatók.)

Ha maga az azonosító ablak az aktív, akkor a lista ábécé sorrendbe rendezett; Ha valamely más mezőablak az aktív, akkor a lista az aktív ablak által képviselt mező értékei szerint rendezett. (Mivel a rendezések egy előfeldolgozó lépés során megtörténnek, ez nem igényel időt.) Így az adatokat tetszőleges szempont szerint rendezve vizsgálhatjuk.

Azonosítók	Mezőnév	Aktív ablak által képviselt mező értékei
	1. Név	
Volvo 340		18840
Ford Fiesta CLX 1.8 D		18865
Peugeot 309 GR		19005
Citroen BX 1.4 RE		19070
Subaru Justy 1200		19200
Ford Escort C 1.8 D		19335
Opel Kadett LS 1.3i	Op	19370
VW Golf 1.6 D		19810
VW Jetta CL 1.3		19875
Renault R 19 TD		20050
Citroen BX 1.6 RE		20182
Isuzu Gemini CLD	Is	20390
Nissan Sunny SLX 1.6		20445
Mitsubishi Lancer 1.5 GLXi		20580
Ford Escort CL 1.6		20595

Az éppen kiválasztott rekordot inverz sáv jelzi. Az azonosító ablakban minden rekord neve mellett két négyzet található. Az egyik négyzet a lokális, a másik a globális kijelölések jelzésére szolgál. Ha a jobb oldali (nagyobb) négyzet megjelenik, akkor az adott rekord beletartozik az éppen aktuális globális kijelölésbe.

Hasonló ugyan a lokális kijelöltséget jelző kisebbik négyzet szerepe is, mégis van egy fontos különbség: ez a négyzet mindig az éppen aktív ablak szerinti lokális kijelölés állapotát mutatja. Ha egy másik ablakot aktiválunk, akkor az abban érvényes lokális kijelölés jelenik meg az azonosító ablakban. (Ez azért hasznos, mert így láthatjuk az aktív mezőablak lokális kijelölésébe tartozó rekordokat azonosítójuk szerint is.) Az azonosítók mellett láthatjuk a rekordhoz rendelt betűjelet is, ha vannak ilyenek. Az utolsó oszlopban jelennek meg az aktív ablak által képviselt mező(k) értékei.

Ha egy rekordnak különös figyelmet kívánunk szentelni, és szeretnénk azt a továbbiakban egyszerűen nyomon követni, célszerű hozzárendelni egy azonosító jelet. Az azonosító jel minden ablakban megjelenik, így a rekord könnyen nyomon követhető. Ez nagyon hasznos lehet például akkor, ha néhány rekord viszonyát tanulmányozzuk: mindegyiket megjelölve anélkül láthatjuk elhelyezkedésüket, hogy meg kellene őket keresnünk a különböző ablakokban.

Az azonosító ablak általában a szűrési műveletek végén hasznos. Ha nem csak tendenciákat kívánunk megállapítani, hanem konkrétan kíváncsiak vagyunk, hogy melyek azok a rekordok, amelyek feltételeinknek eleget tesznek, akkor végiglépkedhetünk a globálisan kijelölt rekordokon, és az azonosító ablakban mutatja, hogy melyek ezek.

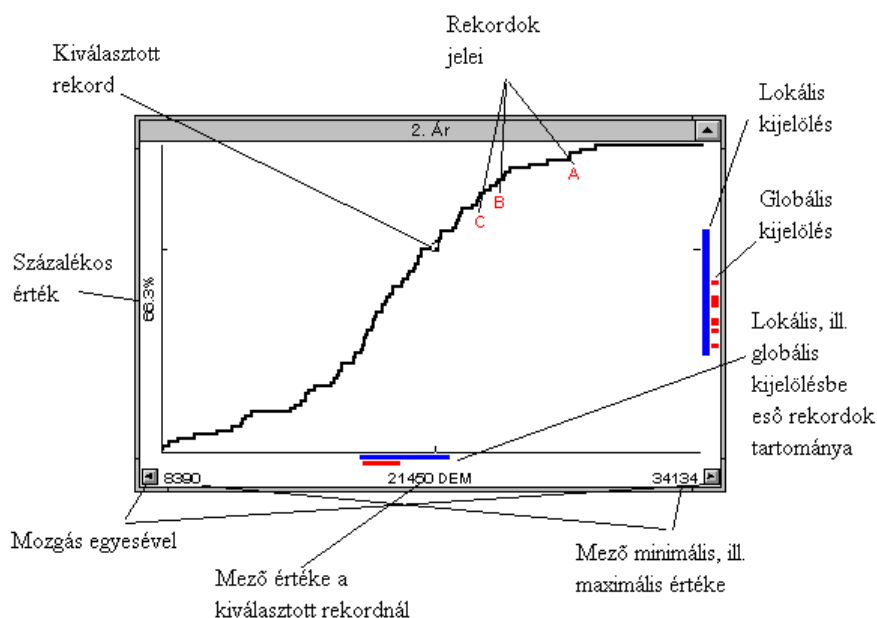
Ha a relációs ablakban (lásd később) egy pontra rákattintunk, akkor szintén az azonosító ablakban állapíthatjuk meg, hogy melyik rekordot képviseli.

2.5.2 Numerikus mező

Értéke tetszőleges számadat lehet. (Ilyen például az ár vagy a fogyasztás.) A DataScope képes feldolgozni olyan adatbázist is, ahol nem áll minden rekordról rendelkezésre az összes adat ("nincs adat" mezőérték). Numerikus adatok esetén a DataScope az adatok

eloszlásfüggvényét ábrázolja, grafikonja lépcsős szerkezetű. (Ilyen például, az autók ára, fogyasztása; személyeket tartalmazó adatbázisnál a bér, az életkor, stb.) A képen ilyen a 2. ablak. Az X tengelyen az adatok a legkisebb értéktől a legnagyobbig vannak ábrázolva. Az Y tengelyen százalékos ábrázolást láthatunk, ahol a 0% képviseli a legkisebb adatot, a 100% pedig a legnagyobbat. A százalékos érték azt mutatja, hogy az adatbázis rekordjainak hány százaléka előzi meg sorrendben a rekordot (a sorrend a rendezettségétől is függ). Például, ha (csökkenő rendezettség esetén) egy autó árához 40% tartozik az Y tengelyen, akkor az adatbázis autói közül 40 százalék ára magasabb (a csökkenő rendezettség miatt a legdrágább autó az első). Más szóval, az Y tengelyről a kiválasztott rekord helyezését olvashatjuk le.

Az ábráról leolvasható az éppen kiválasztott rekord adott mezője értékének viszonya a többihez. (A képen az 1. ablakban az Audi 80 D a kiválasztott és a 2. ablak grafikonján a kis négyzet azt jelzi, hogy ennek az autónak az ára igen magas a többihez képest. Az X tengely alatt leolvasható a konkrét ár (27850 DEM), az Y tengely mellett pedig az, hogy az autók 95.2%-a ennél olcsóbb.) Fordítva, a grafikon tetszőleges pontjára rámutatva a névlista beáll a ponthoz legközelebb eső rekordra, a többi ablak pedig megmutatja a rekord egyéb tulajdonságait.



Az éppen kiválasztott rekord helyét egy kis négyzet jelzi a függvénygörbén. Fontos, hogy több olyan rekord is lehet, amelynek az ablak által képviselt mezője azonos értékű, így egy bizonyos X értékhez több rekord is tartozhat. Így előfordulhat, hogy tovább mozgunk az adatbázisban, de a kis négyzet nem mozdul, ha a következő rekordnál sem változott a mező értéke. Az ablak alsó részén látjuk a kiválasztott rekord mezőjének értékét, valamint a legkisebb és a legnagyobb mezőértéket. A bal szélén a kiválasztott mező értékéhez tartozó százalékos érték (az eloszlásfüggvény értéke a kiválasztott pontban) látható. A kijelöléseket az ablak jobb szélén és alján lévő területeken láthatjuk. Az ablak jobb szélén két egymás melletti oszlop található. A bal oldali a lokális, a jobb oldali pedig a globális kijelölést jelzi..

Az ablak alján lévő vízszintes téglalapok mindig folytonosak, és az első lokálisan (globálisan) kijelölt rekordtól az utolsóig tartanak. Segítségükkel leolvasható, hogy a kijelölt rekordok az adott mező szerint milyen értéktartományba esnek. Például, ha az 'Ár' ablakban a legolcsóbb és a legdrágább autó globálisan kijelölt, a felső vízszintes sáv az ablak bal szélétől a jobb széléig terjed.

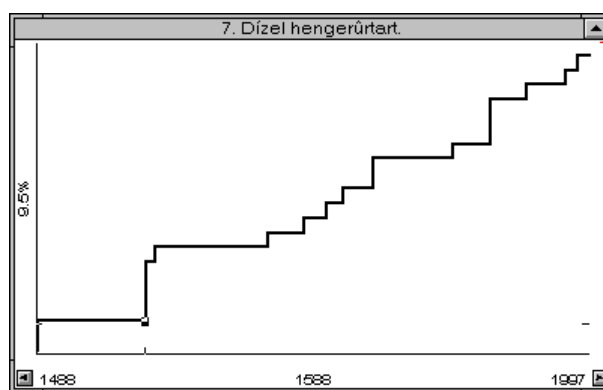
Az eloszlásfüggvény alatt lévő vízszintes sáv megmutatja, hogy az ablak által képviselt mező értékei milyen tartományba tartoznak a legutolsó szűrés szerint. Például, ha kiválasztjuk a dízel, 70 lóerőnél nagyobb teljesítményű autókat, akkor az árat képviselő ablak alján

láthatjuk, hogy a feltételnek megfelelő autók árai milyen intervallumba esnek, és kiválasztható legolcsóbb.

A globálisan kijelölt rekordok helyét mutató függőleges sáv az utoljára megadott feltételt kielégítő rekordok elhelyezkedését mutatja. Ha globálisan kijelöljük a nagy teljesítményű autókat, az 'Ár' ablakban láthatjuk, hogy ezek nagy része magas árú (a függőleges vonal felső részén több lesz a kitöltött terület).

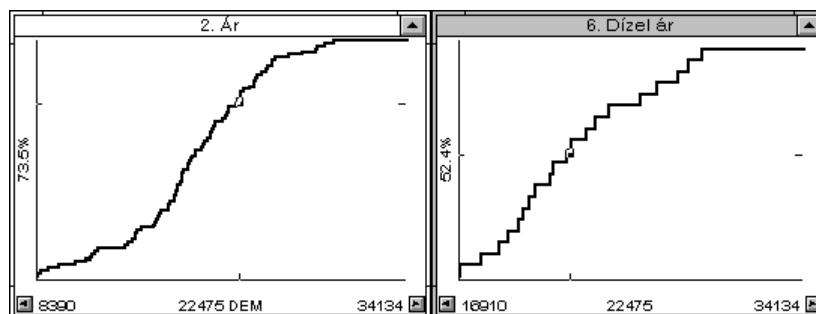
Az eloszlásfüggvényről más információk is leolvashatók. Ha az eloszlásfüggvény szokatlanul hosszú vízszintes vonalat tartalmaz, az azt jelenti, hogy az X tengely ezen intervallumához nem tartoznak rekordok. Például a fenti képen a jobb felső sarokban látható egy hosszú egyenes szakasz, tehát itt egy nagyobb tartományba egyetlen autó ára sem esik.

A hosszú függőleges vonal azt jelzi, hogy az adott X értékhez nagyon sok rekord tartozik. Az alábbi képen például a dízel autók hengerűrtartalmát ábrázoltuk. Láthatjuk, hogy nagyon sok az 1588 köbcéntiméteres autó. Ez egy régebbi (nyugat-európai) korlátozás eredménye, amely az 1600 cm³ feletti dízel autókat megadóztatta. Utána egy hosszú vízszintes vonalat is látunk, amely mutatja, hogy egy nagyobb tartományba nem esnek autók.



A vízszintes, illetve függőleges vonalat természetesen szabadabban is értelmezhetjük (főleg mivel sok rekord esetén az eloszlásfüggvény nem ilyen tagolt). Ha például sok 1580 és 1600 cm³ közötti autó szerepelne, akkor a vonal nem lenne teljesen függőleges, de csak kis vízszintes irányú eltérések lennének. Más szóval, az eloszlásfüggvény azon a helyen meredekebben emelkedne. Ugyanez igaz a vízszintes esetre is: ha az eloszlásfüggvény nem, vagy csak alig emelkedik, akkor kevés rekord esik abba az értéktartományba.

Az eloszlásfüggvény egy másik tulajdonsága: az a kijelentés, hogy közepes árú dízel autók drágábbak, mint a benzinesek, vizuálisan pontosan megmutatja. Válasszuk ki az 50%-hoz legközelebb eső árú autót a Dízel ár ablakban:

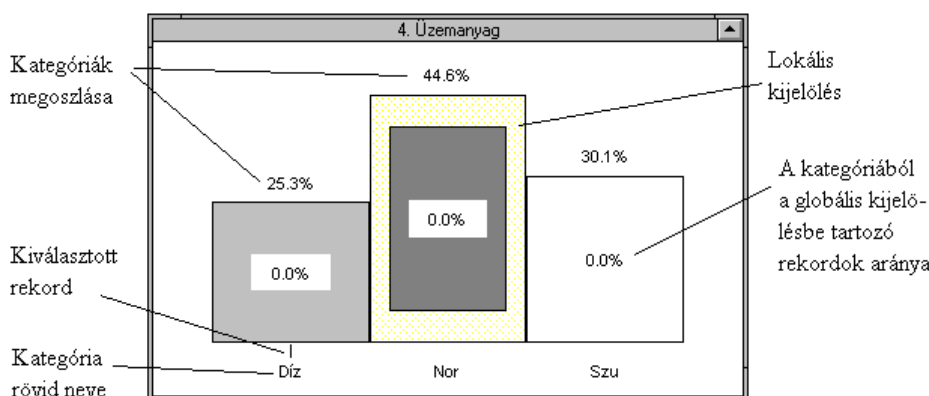


Amint látjuk, a dízel autók esetében az 52.4%-os ár 22475 DEM. A DataScope szinkronitási elve miatt most ugyanez a rekord kiválasztódott az Ár ablakban is, itt azonban százalékos értéke 73.5%. Ez azt jelenti, hogy ami egy dízel autónál közepes árú számítás, az az összes autó között már a magas árkategóriát képviseli. Vagyis a dízel autók drágábbak. Ehhez hasonló következtetések levonására alkalmas a feltételes mező.

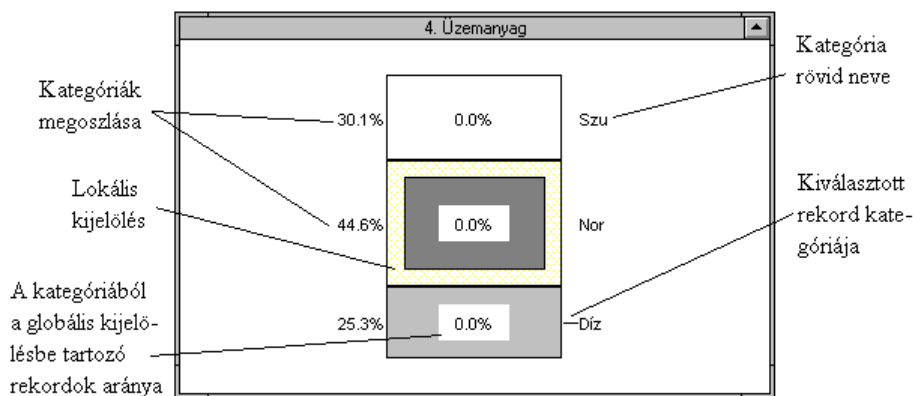
2.5.3 Diszkrét mező

Tartalma általában néhány különböző érték lehet. (A fenti példában a használt üzemanyag fajtáját — dízelolaj('D'), normál benzin('N'), szuperbenzin('S') — tartalmazó mező ilyen.) A diszkrét adatok különféle módokon ábrázolhatók (pl. kördiagram, oszlopdiagram, stb.). A DataScope itt is jelzi, hogy az éppen kiválasztott rekord melyik kategóriába tartozik. (Az 5. ablakban a diagram alatt a "D" (dízel) tartományban látható egy jelzővonalka.) A diszkrét kategóriákat össze is vonhatjuk. (Például a különféle benzintípusoknak (normál, szuper) megfelelő felosztás helyett tekinthetünk egyszerűen dízel- és benzinüzemű autókat.) Ha egy mező lehetséges értékei nem számok, vagy az előforduló értékek számok ugyan, de csak néhány különböző szám, akkor érdemes diszkrét típusúnak definiálni a mezőt (ha az értékek nem számok, akkor kizárólag diszkrét típusú lehet). Ilyenkor a DataScope kategóriákat képez az adatbázis rekordjaiból. Egy kategóriába azok a rekordok tartoznak, amelyeknél az adott mezőben szereplő érték azonos. Az autós adatbázisban diszkrét mező a használt üzemanyag típusa (Üzemanyag mező), amelynek lehetséges értékei a D, N, S betűk. Más adatbázisokban ilyen lehet mondjuk egy olyan mező, amely rekordokat osztályoz (pl. A, B, C osztályon dolgozó személy; I., II. vagy III. osztályú áru, stb.)

A megjelenítésre több lehetőség van. Az oszlopdiagrammon az egyes kategóriák százalékos arányát láthatjuk. A kiválasztott rekord helyét itt is egy vonalka jelzi (a képen az D terület alatt). A lokálisan kijelölt kategóriákat vastagabb keret jelzi (D kategória). Az oszlopok belsejében látható, hogy az egyes kategóriákból mekkora rész tartozik a pillanatnyilag érvényes globális kijelölésbe.



A másik megjelenítési mód az eloszlásdiagram. Ez egy állandó magasságú oszlop, amely az egyes kategóriák arányában van felosztva. A kiválasztott rekord és a lokális kijelölés jelzése az oszlopdiagramnál leírt módon történik. Hasonlóan az oszlopdiagramhoz, itt is látjuk, hogy az egyes kategóriák mekkora része esik az éppen aktuális globális kijelölésbe.



A diszkrét típusú mezőt megjelenítő diagramokat a szokásos módon használhatjuk statisztikai következtetések levonására. A diagramokon kényelmesen jelölhetünk ki kategóriákat (csak egész kategóriákkal dolgozhatunk) és végezhetünk velük szűrési műveleteket. Például egy kattintással kijelölhetjük a dízel autókat, és a Kijelölés menü segítségével módosíthatjuk globális kijelölésünket. Diszkrét mezők esetében a lokális kijelölés kategóriánként történik, vagyis egy egész kategóriát jelölünk ki egyszerre. Ez gyorsítja az egyes kategóriák szerinti szűréseket. Ha létrehoztunk egy globális kijelölést, az oszlopdiagramon láthatjuk, hogy az hogyan oszlik meg az egyes kategóriák között. Így láthatjuk, hogy egy bizonyos feltételrendszerrel kielégítő rekordokból mennyi esik az egyes kategóriákba. Például ha kiválasztjuk a 20000 márkánál olcsóbb autókat és belőlük globális kijelölést képzünk, akkor az Üzemanyag ablakban láthatjuk, hogy ezek az autók nagyrészt a benzines kategóriába esnek.

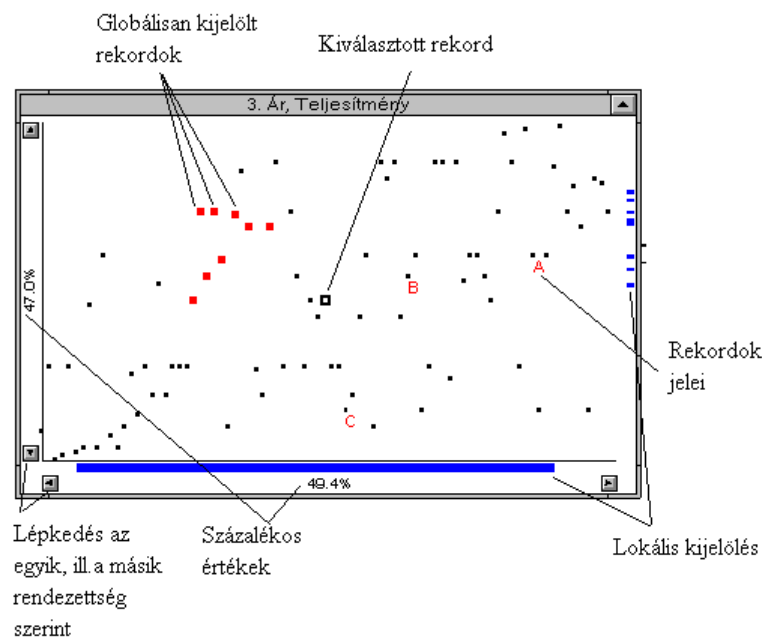
Ha kívánjuk, a diszkrét kategóriánkat össze is vonhatjuk. Ha két kategóriának ugyanazt a rövid nevet adjuk, azok ideiglenesen összeolvadnak. Például megtehetjük, hogy a normál és szuper benzinnel üzemelő autókat egybevonjuk úgy, hogy mindkét kategóriának a 'B' rövid nevet adjuk. Ezután csak D és B jelű kategóriák lesznek.

2.5.4 Relációs diagram

A leghatékonyabb megjelenítési mód az ún. relációs diagram (3. ablak). Ez két numerikus adatmező viszonyának elemzését teszi lehetővé. Az egyik tengelyen az egyik adatmező értékeit ábrázoljuk a minimálisról a maximálisig, a másik tengely ugyanígy képviseli a másik adatmezőt. Az adatbázis rekordjait egy-egy pont jelzi, amely a rekord két mezője eloszlásfüggvény értékének metszéspontjában helyezkedik el.

A relációs diagram lehetőséget ad a kivételek gyors kiszűrésére. Minél távolabb van egy pont az átlótól (az $y = x$ egyenestől), annál inkább eltér az adott rekord az átlagtól. A fenti példában a 3. ablak az ár (vízszintes tengely) és a teljesítmény (függőleges tengely) mezők kapcsolatát ábrázolja. A listából kiválasztott Audi 80 D helyét egy kettős négyzet jelzi. Az ábráról leolvasható, hogy ennél az autónál igen magas árhoz alacsony teljesítmény (14.5%) tartozik, ami kedvezőtlen tulajdonság. Azt is láthatjuk, hogy az autók túlnyomó részénél a két tulajdonság között lineáris a kapcsolat, és főképp negatív kivételek vannak.

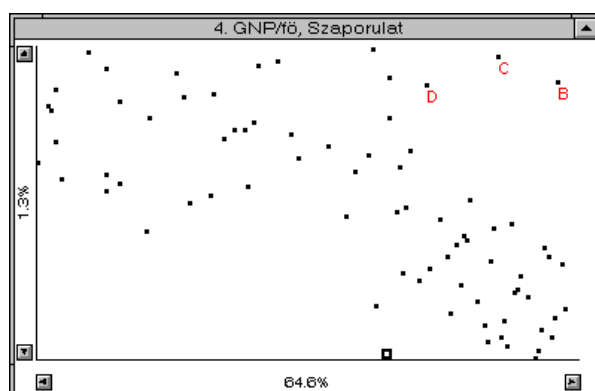
Ha a diagramon kiszemelünk egy számunkra érdekes pontot, rákattinthatunk az egérrel. A többi ablak (köztük a névlista) azonnal megmutatja, hogy melyik rekordról van szó és milyenek az egyéb tulajdonságai. Természetesen több különböző relációs diagramot is megnyithatunk és az egyikben egy kivételes rekordot kiválasztva, megállapíthatjuk, hogy az más szempontok szerint is eltérő-e az átlagtól.



Az ablak átlói fontos választóvonalak, ezek mentén helyezkednek el ugyanis azok a rekordok, amelyeknél a két tulajdonság szerint elért helyezés hasonló (pl. átlagos ár/teljesítmény viszony). Minél távolabb helyezkedik el egy pont ettől az egyenestől, annál inkább kivételes az általa képviselt rekord.

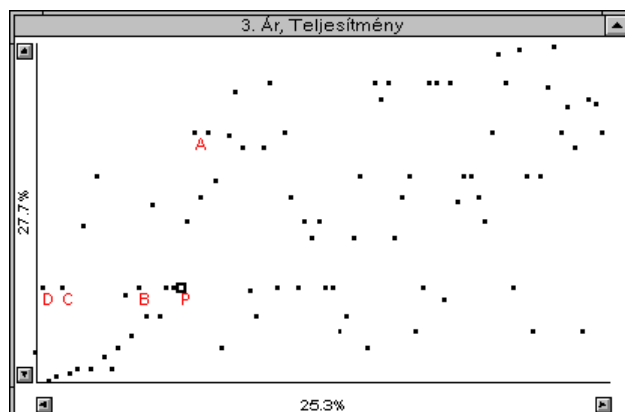
Ez az ablaktípus tehát jó lehetőséget ad a kivételek gyors kiszűrésére. Ha ábrázoljuk két mező viszonyát, akkor azonnal szembeötlenek a kiugró pontok. Egy pontra rákattintva, kiválaszthatjuk a hozzátartozó rekordot, és megnézhetjük az azonosító ablakban, hogy név szerint melyik az. Más ablakokban pedig láthatjuk annak egyéb tulajdonságait. Ha több relációs ablakot is megnyitottunk, akkor ezekben azonnal láthatjuk azt is, hogy más szempontok szerint is kivételes-e az adott rekord.

A relációs ablak a kivételek mellett a tendenciákat is mutatja. Egy pillantással megállapíthatjuk, hogy milyen jellegű kapcsolat van a két tulajdonság között, és hogy ez mennyire "erős", azaz hogy mennyi olyan pont van, amely nem követi ezt a tendenciát.



A fenti kép például a világ országainál mutatja a nemzeti jövedelem és a szaporulat arányát (1988-as adatok). Jól látható a tendencia, amely szerint minél gazdagabb egy ország, annál kevesebb gyermeket vállalnak. Megjelöltünk néhány kivételt is: B - Kuvait, C - Líbia, D - Venezuela. Középen alul a kiválasztott rekord Magyarország. A további következtetések levonását az Olvasóra bízunk.

Azt is könnyen megvizsgálhatjuk, hogy létezik-e egy bizonyos rekordnál jobb választás. Ábrázoljuk az autók ár/teljesítmény viszonyait egy relációs diagramon, mindkettőt növekvő rendezettséggel:



Vizsgáljuk meg a fenti képen a "P" betűvel jelölt autót. Tegyük fel, hogy szeretnénk egy hasonló árú, de nagyobb teljesítményű autót. Mivel az árat a vízszintes tengelyen ábrázoltuk, ezért az azonos árú autók azon a függőleges egyenesen helyezkednek el, amelyik átmegy a "P" ponton. Láthatjuk, hogy bár ezen az egyenesen nincs másik pont, jóval a "P" fölött, egy kicsivel jobbra van egy érdekes pont, amelyet "A"-val jelöltünk meg. Az "A" pont által képviselt autó egy kicsivel drágább ugyan, viszont a teljesítménye jóval nagyobb. Ezután eldönthetjük, hogy megéri-e nekünk az ártöbbletet ez a teljesítménynövekedés, és persze meg kell vizsgálnunk a másik autó egyéb tulajdonságait is.

Azt is láthatjuk a képen, hogy a "P"-vel azonos teljesítményű autót olcsóbban is kapunk (pl. a "B", "C" és "D" jelű autók ilyenek). Természetesen itt is meg kell vizsgálnunk, hogy ezek megfelelnek-e egyéb elvárásainknak is.

2.6 A DataScope főb tulajdonságai és további lehetőségei

2.6.1 Felismerhető adatbázisok

A DataScope az adatokat a Microsoft Open Database Connectivity (ODBC) szabványa segítségével olvassa be az adatbázisokból. Ez a szabvány lehetővé teszi tetszőleges típusú adatbázis kezelését egy meghajtó segítségével. A DataScope-hoz mellékelt meghajtók segítségével beolvashatók adatok szöveges, DBase, Excel, Paradox, Btrieve, MS-Access, FoxPro file-okból, és SQL szerverek adatbázisaiból. További meghajtók a Microsoft-tól szerezhetők be.

2.6.2 Szinkronitás

A program legfontosabb jellemzője a teljes szinkronitás. Az előbbieken láttuk, hogy miközben az adatbázis valamely elemét egy bizonyos szempontból vizsgáljuk, a program automatikusan megjeleníti ugyanazon elem más tulajdonságait is. Kiválaszthatunk meghatározott szempont szerint is bizonyos tulajdonságú elemeket, és láthatjuk ezek egymáshoz való viszonyát más szempontok szerint. Egyidejűleg 16 ablakot nyithatunk meg, azaz egyszerre ennyi tulajdonság (vagy tulajdonságpár) szerint tanulmányozhatjuk az adatokat.

2.6.3 Interaktivitás, grafikus lekérdezés

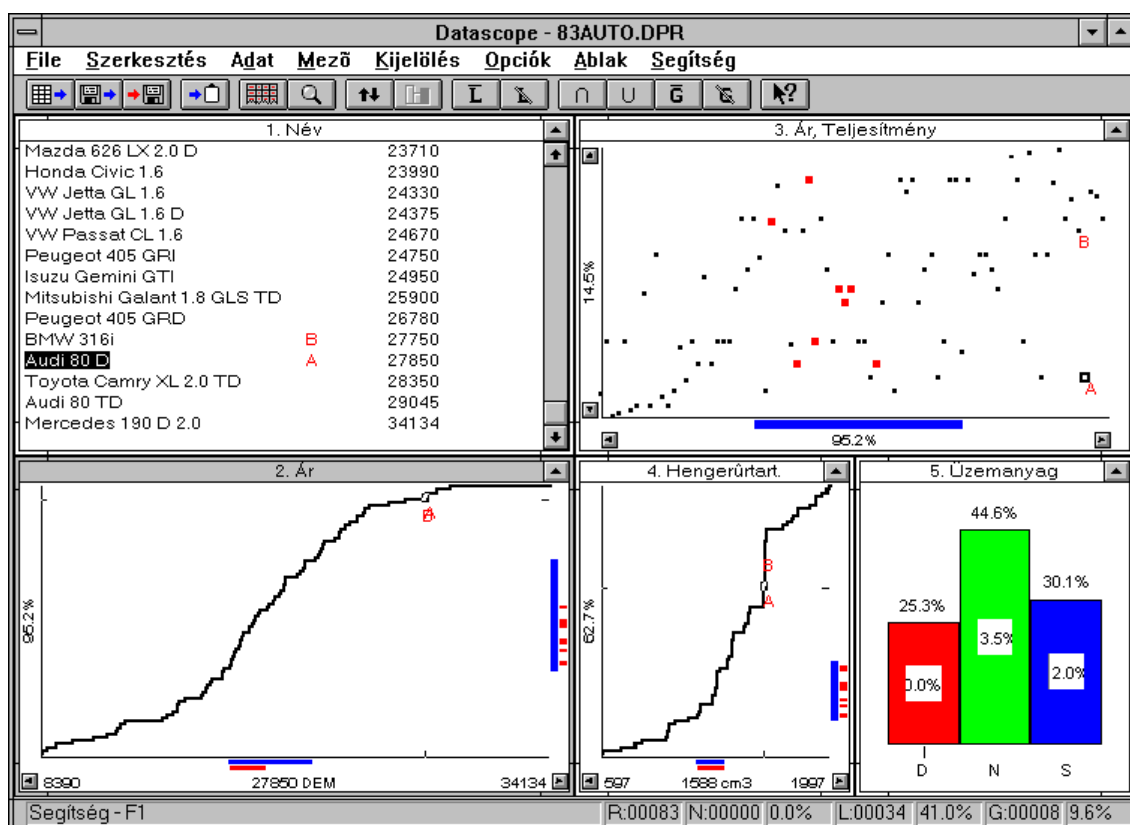
Az adatbázis a megjelenő diagramokból interaktívan lekérdezhető. Ez szükségtelemmé teszi parancsok begépelését. Egyszerűen csak ki kell jelölni egy pontot, vagy intervallumot az egerrel a kívánt információ megjelenítéséhez. A DataScope ezért tekinthető közvetlen vizuális lekérdező rendszernek is.

2.6.4 Kontextfüggő elemzés

A numerikus adatok elemzése sokkal hatékonyabban történik. Eddig nagyon sok időnkbe telt megállapítani egy rekord viszonyát a többiek között, az eloszlásfüggvény értéke azonban ezt azonnal mutatja. Például, ha azt hallottuk, hogy X autó Y litert fogyaszt, nem tudtuk, mennyire jó ez az érték, amíg nem néztünk át egy autókatalógust, vagy nem használtunk statisztikai módszereket, hogy képet kapjunk a jelenlegi helyzetről. Most, a DataScope-pal egyszerűen csak leolvassuk a százalékos értéket az Y tengelyről.

2.6.5 Interaktivitás, grafikus lekérdezés

Az adatbázist interaktív módon, közvetlenül az ábrázolt diagramokból kérdezhetjük le, így szükségtelemmé válik szöveges parancsok kiadása. Egyszerűen csak rá kell mutatnunk a diagramok megfelelő pontjára vagy intervallumára, és azonnal megjelenik a kívánt információ. Így a DataScope on-line lekérdező rendszerként is jól alkalmazható.



2.6.6 Elemzés a DataScope-pal

A DataScope projekt file-okkal dolgozik, amelyek tartalmazzák az adatbázis adatait, és minden beállítást, ami a munka későbbi folytatásához szükséges.

2.6.7 Bonyolultabb műveletek megvalósítása

Az Unió és a Metszet menüpontok mindig törlik az előző globális kijelölést, és újat hoznak létre az összes lokális kijelölésből. A DataScope egy speciális technikával lehetőséget biztosít bonyolult lekérdezések megvalósítására. Miután a lokális kijelölésekből létrehoztunk egy globális kijelölést, amelyen másféle műveletet is szeretnénk végezni, először rögzítenünk kell a globális kijelölést. Majd létrehozhatunk egy új diszkrét mezőt, amelynek két lehetséges értéke van: "Igen", ha a rekord beletartozik a globális kijelölésbe; "Nem", ha nem tartozik bele. Az új mezőnek nevet adva a későbbiek során ugyanúgy használható mint a diszkrét mező.

2.6.8 Numerikus mező létrehozása a globális kijelölésből

Lehetőségünk van arra is, hogy a globális kijelölésből egy új numerikus mezőt készítsünk valamelyik, már létező numerikus mező alapján. Ennek segítségével elérhetjük, hogy csak egy bizonyos feltételnek megfelelő rekordcsoport tulajdonságait vizsgáljuk

2.6.9 Új projekt létrehozása a globális kijelölésből

A globális kijelölésbe tartozó rekordokból új projektet hozhatunk létre a Kijelölés/Globálisból új projekt menüpont használatával. Az újonnan létrehozott projekt adatbázisa csak a kijelölt rekordokat fogja tartalmazni.

2.6.10 Kijelölések törlése, komplementálása

Egy adott mezőablakhoz tartozó lokális kijelölést megszüntethetünk a Kijelölés/Lokális törlése menüpont segítségével. Megtehetjük azt is, hogy a kijelöléseket ellenkezőjére fordítjuk (komplement-képzés). Ezzel a kijelölt rekordokból jelöletlenek, az eddig nem jelöltek pedig kijelöltek lesznek.

Irodalomjegyzék

- Fayyad U.M.,Piatetsky-Shapiro G., Smyth P.,Uthurusamy R. (1996): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Fayyad U.M.,Piatetsky-Shapiro G., Smyth P.: From Data Mining to Knowledge Discovery: An Overview, 1-34
- Fayyad U.M.,Piatetsky-Shapiro G., Smyth P.,Uthurusamy R. (1996): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Brachman R.J., Anand T.: The Process of Knowledge Discovery in Databases: A Human-Centered Approach, 37-57
- Fayyad U.M.,Piatetsky-Shapiro G., Smyth P.,Uthurusamy R. (1996): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Buntine W.: Graphical Models for Discovering Knowledge, 59-82
- Fayyad U.M.,Piatetsky-Shapiro G., Smyth P.,Uthurusamy R. (1996): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Simoudis E., Livezey B., Kerber R.: Integrating Inductive and Deductive Reasoning for Data Mining, 353-373
- Fayyad U.M.,Piatetsky-Shapiro G., Smyth P.,Uthurusamy R. (1996): Advances in Knowledge Discovery and Data Mining. AAAI Press/The MIT Press, Piatetsky-Shapiro G.: Data Mining and Knowledge Discovery Internet Resources, 593-595
- KESDA`98 Conference (1998.04.27-28.): Knowledge Extraction from Statistical Data (Data Mining and Symbolic Data Analysis). Conference Organiser, Luxembourg
- Piatetsky-Shapiro G., Frawley W.J.(1991): Knowledge Discovery in Databases. AAAI Press/The MIT Press,