**1 Paskaita** Duomenų gavybos tikslai, pritaikymai, uždaviniai,

sistemos, metodai

**1.1*. Kas yra duomenų gavyba?***

Sukauptų duomenų analizė leidžia efektyviai pasinaudoti informacija, slypinčia duomenyse. Duomenų gavyba (*Data Mining*) yra šiuolaikinė informacijos analizės sritis, atsiradusi duomenų bazių technologijų, dirbtinio intelekto ir statistinės duomenų analizės sankirtoje. DG yra labai plati sritis, apimanti daug metodų, algoritmų bei taikomųjų programinių sistemų. Jei įprasti duomenų analizės metodai padeda atskleisti tiriamų kintamųjų priklausomumą, tai DG unikali tuo, kad analizės rezultatas yra naujų priklausomybių, apie kurių egzistavimą buvo, ar net nebuvo, įtariama radimas.

Šiuolaikinės DG technologijos pagrindas yra šablonų (*pattern* – angl.), atvaizduojančių įvairia-aspektinius duomenų tarpusavio santykius, koncepcija. Šie šablonai rodo dėsningumus, būdingus duomenų poaibiams, kurie gali būti išreikšti vartotojui suprantamu pavidalu. Svarbi DG savybė yra ieškomų šablonų netrivialumas. Tai reiškia, kad rasti šablonai turi atvaizduoti neakivaizdžius, netikėtus (*unexpected*-angl.) duomenų reguliarumus, kurie sudaro taip vadinamas paslėptas žinias (*hidden knowledge*-angl.). Nemažai žmonių duomenų gavybą labiau laiko metodologinių principų rinkiniu, arba netgi verslo filosofijos ar matematikos sritimi, negu praktinių sprendimų rinkiniu, skirtų verslo problemoms.

Štai pora apibrėžimų.

Pateikiame "Gartner Group" apibrėžimą:

*Duomenų gavyba yra prasmingų šablonų* (patterns)*, dėsningumų, modelių ir tendencijų radimo procesas dideliuose informacijos kiekiuose, pasinaudojant modelių atpažinimo, statistiniais bei matematiniais metodais.*

Dažnai naudojamas toks DG apibrėžimas (G. Piatecki-Shapiro).

*Duomenų gavyba yra “žalių”, neapdorotų duomenų (*raw data*) tyrinėjimo procesas žinioms nustayti, kurios yra:*

• *naujos, prieš tai nežinotos;*

• *netrivialios;*

• *praktiškai naudingos;*

• *interpretuotinos;*

• *būtinos sprendimams priimti pasirinktoje veiklos srityje.*

Dažnai vartojami keli terminai, kuriuos galima laikyti sinonimais: duomenų gavyba

(*Data Mining*) ir žinių paieška duomenų bazėse *(KDD — Knowledge Discovery in Databases*).

Duomenų gavyba versle kartais vadinama verslo intelektika (*Business Intelligence*).

**1.2*. Kur pritaikoma DG ?***

Šiuolaikinė duomenų analizė pasižymi tokia specifika:

• duomenų apimtis yra beveik neaprėžta;

• duomenys yra įvairialyčiai (*kiekybiniai, tekstiniai, video, audio, ..*.);

• analizės išvados turi būti konkrečios ir aiškios;

• analizės priemonės turi būti paprastos naudoti.

DG taikymo sritys apima dvi kryptis:

- verslo pritaikymai,

- unikalūs tyrimai (*bioinžinerija, genetika, socialinės sistemos, ir pan.*).

DG yra aktuali finansų ir draudimo institucijoms, telekomunikacijų įmonėms, medicinos įstaigoms, transporto kompanijoms, vyriausybinėms organizacijoms, didelėms prekybos organizacijoms, ir pan.. Duomenų gavybos užduotys ir rezultatai gali būti ir kasdieniai (*pavyzdžiui, įvertinimas, kad maisto prekių parduotuvės klientai savaitgaliais* *kartu perka stipresnį alų ir traškučius*), ir sensacingi (*vaikams duodamo aspirino ir* *mirtinos ligos* (Reyes Syndrome) *ryšys*), ir intriguojantys (*NBA komandos Toronto Raptors* *treneris, norėdamas pagerinti komandos žaidimą, naudoja duomenų gavybos programas* *varžybų metu*).

Išvardinsime uždavinius, kurie sprendžiami DG priemonėmis įvairiose srityse.

Komercija:

kliento „krepšelio“ tyrimas (*strategijai, prekių planavimui ir pan*.), skirtas prekių derinių, kurias vartotojai linkę įsigyti kartu, paieškai;

laiko šablonų tyrimas padeda prekeiviams priimti sprendimus apie atsargų kaupimą;

prognozavimas leidžia prekeiviams nustatyti įvairių klientų kategorijų poreikius.

Bankai, finansų institucijos:

sukčiavimų nustatymas – analizuojant buvusių sukčiavimų stereotipus;

klientų klasifikacija – taip marketingo politika tampa labiau tiksli ir rezultatyvi;

klientų būklės prognozė leidžia prognozuoti klientų vertę.

Telekomunikacijos:

iškvietimų analizė (*skambučių analizė*) leidžia nustatyti klientų su panašiais poreikiais kategorijas;

klientų lojalumo didinimas – klientų nustatymas, kurie ir toliau naudosis kompanijos paslaugomis;

tinklo apkrovimo analizė (*tinklo pralaidumas*).

Draudimas:

sukčiavimų analizė – analizuojant buvusių sukčiavimų stereotipus;

rizikos analizė – nustatant tam tikrus panašumus tarp skirtingų klientų.

Medicina:

šablonų ieškojimas tam tikrų susirgimų srityje (*vaistų pirkimų tyrimai,….);*

genetiniai tyrimai.

**1.3*. Duomenų gavybos uždaviniai***

Duomenų gavyba apima dvi plačias tyrimų kategorijas:

• Priklausomybių tyrimo duomenų gavyba

• Prognozuojanti duomenų gavyba

***Priklausomybių tyrimo duomenų gavyba***

Priklausomybių suradimo duomenų gavybai priskiriama grupė uždavinių, kuriuose nustatomi šablonai duomenų visumoje be išankstinio žinojimo apie jau egzistuojančius šablonus.

Štai keletas pavyzdžių.

*Grupavimas*.

Grupavimas – sąvoka, naudojama metodams apibūdinti, kurie bando sugrupuoti duomenų įrašus pagal tai, kiek jie yra panašūs. Duomenų įrašas gali apimti, pavyzdžiui, kiekvieno verslo kliento aprašymą. Šiuo atveju grupavimo metodas sugrupuotų visus panašius klientu kartu tuo pačiu maksimizuodamas skirtumus tarp skirtingų klientų grupių, sudarytu remiantis šiuo metodu. Paprastai grupuojama remiantis klasterizavimo metodais.

*Ryšių analizė*.

Ryšių (asociacijų) analizė apibūdina problemų šeimą, kuriose reikia nustatyti ryšius tarp duomenų įrašų. Labiausiai žinoma pardavimo krepšelio analizė. Šiuo atveju, duomenų įrašai yra kliento pirktos prekės vieno pirkimo metu. Pardavimo krepšelio analizė nustato prekių, kurias pirko skirtingi klientai, kombinacijas ir remiantis asociacijomis (arba sąryšiais) galima susidaryti vaizdą, kokie produktai perkami kartu. Jei prekybos krepšelį laikyti duomenų grupavimo metodu, tai toks metodas gali būti pritaikytas bet kurioje situacijoje, kai apdorojamas didelis kiekis grupių su duomenų įrašais.

*Dažnumų analizė*.

Dažnumų analizė apima duomenų gavybos metodikas, nustatančias tam tikrų įvykių dažnumus, kurie yra pritaikomi duomenų rinkinių analizei.

Prognozuojanti duomenų gavyba

Prognozuojanti duomenų gavyba apima problemų grupę, kuriose reikia rasti sąryšius tarp tam tikro kintamojo (vadinamo *kintamasis-taikinys*) ir kito kintamojo duomenyse, kuris leidžia prognozuoti *kintamąjį-taikinį*. Štai keletas pavyzdžių.

*Klasifikavimas*.

Klasifikavimo metu duomenų įrašai yra priskiriami iš anksto nustatytoms klasėms (kategorijoms). Pavyzdžiu galėtų būti klientų priskyrimas prekybos segmentui. Šiuo atveju kintamasis-taikinys yra segmento klasė, o klasifikavimo metodai nustato ryšius su kitais kintamaisiais ir klasėmis. Klasifikuojant naują įrašą klasifikavimo metodas nustato klasę bei tikimybę, kad įrašas priklauso tai klasei. Klasifikavimas dažnai atkiekamas po grupavimo, kurio metu nustatomos klasės.

*Reikšmių prognozavimas*.

Kintamojo-taikinio reikšmė yra prognozuojama pagal kitus kintamuosius duomenų įraše. Pavyzdžiui, tikėtinų klientų išlaidų prognozavimas pagal amžių, lytį ir pajamų grupę.

*Nuoseklumų paieška laike.*

Tai dėsningumų paieška, apimanti prognozavimą laike ir prognostinių scenarijų tyrimą. Čia svarbu ne tik tai, kokiomis paslaugomis naudojasi vartotojai, bet ir kokia eilės tvarka. Šie metodai bei technologijos padeda efektyviau teikti paslaugas. Dažnai DG uždaviniui išspręsti taikomi keli metodai iš eilės ar net sudėtingi jų deriniai. Uždavinių bei metodų įvairovę papildo grupė duomenų gavybos algoritmų. Nė vienas jų nėra universalus ar nepriekaištingas. Parenkant algorimus atsižvelgiama į jų operacinį ir loginį sudėtingumą), sugaištamą analizei kompiuterio laiką bei atmintį, analizės patikimumą.

***1.4. DG sistemos***

Išskirkime tokias DG sistemų klases.

1) *Dalykinės analizės sistemos* (remiasi tos dalykinės srities empiriniais duomenimis – pvz., statybininkai žino viską apie betoną, medikai apie ligas ir vaistus, ir pan.). Šie metodai dažnai naudoja paprastą statistinį aparatą, bet maksimaliai įvertina susiklosčiusios situacijos specifiką. Šios sistemos gali būti labai įvairios. Tokios sistemos dažnai naudojamos finansinių rinkų tyrimuose. Specializuota rinkos analizės sistema yra sudaryta iš finansinių indeksų sekų dinamikos analizės ir investicijų portfolio planavimo metodų.

2) *Statistiniai programų paketai* (SAS, STATISTICA)

Beveik visų žinomų statistinių paketų paskutinėse versijose kartu su tradiciniais statistiniais metodais įtraukiami ir DG metodai. Tačiau pagrindinis dėmesys juose skiriamas klasikiniams metodams – koreliacinei, regresinei, faktorinei analizei ir pan. Tokių sistemų trūkumu laikoma specialaus vartotojo paruošimo būtinybė.

*3) DG paketai*

Daugelis programinės įrangos korporacijų kuria ir platina atskirus produktus,skirti DB duomenų gavybai. Lietuvoje tokią įranga platina Oracle, IBM DM Technology, SONEX, ir kt.

***1.5. Metodai***

Duomenų gavyba atsirado ir vystėsi taikomosios statistikos, vaizdų atpažinimo, dirbtinio intelekto metodų, duomenų bazių teorijos ir t.t., bazėje. Todėl DG turi daugybę realizuotų jos schemose algoritmų ir metodų. Daugelyje tokių schemų integruota iškart keletas metodų bei algoritmų. Trumpai apžvelgsime dažniausiai taikomus metodus.

1. *Neuroniniai tinklai* (neural network)

Tai netiesiniai modeliai, sėkmingai naudojami klasifikavimo ir prognozavimo uždaviniams spręsti. Jų struktūra primena biologinius neuroninius tinklus. Dažnai dėl struktūros ir sudėtingumo neuroniniai tinklai vadinami "juodąja dėže". Tinklas sudarytas iš kelių sluoksnių: įvesties, išvesties, ir vieno ar daugiau paslėptų neuronų. Įvesties sluoksnyje pateikiama elementą aprašanti informacija, o išvesties sluoksnyje gaunamas rezultatas - priklausymas klasėms (klasifikavimo uždaviniuose) ar numanomas rezultatas (prognozavimo uždaviniuose). Paslėpti sluoksniai realizuoja uždavinio logiką. Kiekvienas tinklo sluoksnio mazgas yra susijęs su kiekvienu šalia esančio sluoksnio mazgu. Nagrinėjami elemento duomenys verčiami į rezultatus, atsižvelgiant į tam tikrus didesnius ar mažesnius svorius, turinčius įtakos sprendimui, todėl šių svorių parinkimas ir yra svarbiausias algoritmo uždavinys.

1. *Analogijų paieška* (case-based reasoning)

Klasifikavimo uždavinius galima spręsti ne tik grupuojant elementus pagal jau nustatytas ar analizės metu rastas naujas taisykles, bet ir pagal elemento panašumą į gretimus elementus. Šie algoritmai remiasi daikto, veiksmo ar kitaip apibrėžiamo nagrinėjamo vieneto lyginimu su kitais. Analizė remiasi istorinės informacijos perrinkimu, todėl algoritmo trūkumas — sudaryto modelio sudėtingumas.

1. *Sprendimų medžiai* (decision trees)

Šiuos algoritmų rezultatą galima įsivaizduoti kaip panašią į medį struktūrą. Kiekvienas išsišakojimas reiškia vieną ar kitą atsakymą į klausimą. Tokiu būdu sudaromos taisyklės, kurios nagrinėjamą duomenų aibę klasifikuoja atsižvelgdamos į element savybes. Proceso pradžioje turėta duomenų aibė skaidoma į šakas, kol kiekviena jų tampa homogeniška. Sprendimų medžių privalumas — jų aiškumas, jie suprantami tiek problem formuluojantiems, tiek bandantiems uždavinį realizuoti, tiek analizuojantiems gautus rezultatus.

1. *Evoliucinis programavimas*

PolyAnalyst sistema gerai iliustruoja šį algoritmą. Šioje sistemoje hipotezes (kaip priklausomybės vieno kintamojo nuo kitų) formuluojasi kaip programos tam tikra vidine programavimo kalba. Programų sudarymo procesas panašus į evoliuciją program pasaulyje (šituo algoritmas panašus į genetinius algoritmus). Kai sistema randa programą, labiausiai tinkančią ieškomai priklausomybei atvaizduoti, ji pradeda toje programoje daryti nedideles modifikacijas ir tarp sukurtų dukterinių programų atrenka didinančias tikslumą. Specialus PolyAnalyst modulis verčia rastas priklausomybės iš sistemos vidinės kalbos į vartotojui suprantamą (matematinės formulės, lentelės ir t.t.)

1. *Genetiniai algoritmai*

Tai algoritmai, naudojantys selekcijos, mutacijos ir kryžminimo procesus bei besiremiantys evoliucijos koncepcija.

1. *Riboto perrinkimo algoritmas*

Algoritmas remiasi loginių taisyklių "jei — tai" taikymu. Panašiai kaip sprendimų medis, elementai grupuojami pagal jų duomenis. Jis nėra tobulas, tačiau šiandien laikomas vienu iš geriausiu DG rinkoje (šio algoritmo naudojimo pavyzdį galima rasti pakete WizWhy).

1. *Duomenų vizualizacija*

Grafinio duomenų atvaizdavimo galimybes turi beveik visos DG priemones. Priemonės, skirtos būtent tokiam tikslui užima didelę rinkos dalį. Jose daugiau dėmesio skirta vartotojo sąsajai, kuri leidžia atvaizduoti nagrinėjamus duomenis naudojant skirtingus parametrus (spalva, forma, orientacija savo ašies atžvilgiu, dydis ir kiti grafinių elementų ypatumai).

Duomenų gavybos taikymas nesibaigia teisingo algoritmo parinkimu. Kaip ir daugelyje sričių, čia svarbūs visi etapai: problemos identifikavimas, kokybiški duomenys, kurių pagrindu bus atliekama analizė, modelio parinkimas ir realizavimas, gautų rezultatų interpretavimas. Šioje srityje ypač reikalingas vartotojų išprusimas, nes nė viena sistema negali tapti "stebuklingu mygtuku", išspręsiančiu visas problemas.