

1. Datu-Meatzaritzarako sarrera

Irakasgaia: Konputazio-Zientzien Metodo Matematikoak

Titulazioa: Informatikan Ingeniaria

Konputazio Zientzia eta Adimen Artifiziala saila

Universidad del País Vasco - Euskal Herriko Unibertsitatea

Aurkibidea

- 1 Datu-Meatzaritza
- 2 Knowledge Discovery from Databases (KDD): datubaseetan ezagutza aurkitzea

Aurkibidea

1 Datu-Meatzaritza

2 Knowledge Discovery from Databases (KDD): datubaseetan ezagutza aurkitzea

Definizio batzuk

- **Data mining**. Aurretik ezezaguna den ezagutza erabilgarri eta ulergarria erauzteak formatu desberdinetan gordeta dauden datu-multzo handietatik (Witten eta Frank, 2000)
- **Knowledge discovery in databases**. Datubaseetan ezagutza aurkitzea, datuetatik abiatuz eredu baliagarriak, berriak eta azken batean ulergarriak identifikatuz (Fayyad eta lag. 1996)

Eredu-motak

- **Datuetatik ezagutza** lortzea eredu konputazionalak erabiliz
- **Eredu deskribatzaileak**: datuak deskribatzen edo laburtzen dira
 - **Erregelak**: datuen portaera-ereduak erakusten dira
 - **Clustering**: homogenoak diren kasuak multzokatzen dira
- **Eredu iragarleak**: aldagai iragarleen balioetatik abiatuz iragarri beharreko beste aldagai batzuren balioak estimatzen dira
 - **Erregresioa**: Iragarri beharreko aldagaia jarraia da
 - **Gainbegiraturako sailkapena**: Iragarri beharreko aldagaia diskretua da (ordinala edo nominala)

Datu-motak

● Database erlazionalak

- Erlazio-multzoak (taulak). Ezaugarri-multzo baterako (aldagaiak, zutabeak, eremuak) n -koteak gordetzen dira (kasuak, errenkadak, erregistroak)
- Taula moduko aurkezpena: ezaugarri-balio

● Database espazialak: geografi-datuak, medikuntza-irudiak, garraio-sareak,

● Database tenporalak: denborazko une edo tarte desberdinak

● Database dokumentalak: Objektuak testu-dokumentuak dira, aldagaiek hitzak adierazi ditzakete, edo laburpenak...

● Multimedia datubaseak: irudiak, soinua, bideoa

● World Wide Web: gaurregun dagoen informaziorako gunerik handiena eta anitzena

- Edukiaren analisia: web orrietan ereduak aurkitzea
- Egituraren analisia: hiperestekak eta URL-ak aztertzea
- Erabileraren analisia: nabigazioa aztertzea

Erlazioa beste zenbait arloarekin

- **Estatistika.** Datu-meatzaritzaren " ama "
- **Ikasketa automatikoa.** Konputagailuak adibideetatik ikasten du
- **Ereduen ezagutza.** Clustering. Sailkapen-gainbegiratua
- **Erabakiak hartzeko sistemak.** Zuzendaritzari laguntzeko tresnak eta sistemak
- **Datuen bistaratzea.** Grafiken bidez datuetatik ateratako ereduak ikusten eta ulertzen laguntzea
- **Datubaseak.** Datuen biltegiak. Atzipen eraginkorra
- **Informazioaren erauzketa.** Testuak. Liburutegi digitalak. Bilaketak Interneten
- **Konputazio paraleloa eta banatua.** Datu-meatzaritzak dakarren kostu konputazionala banatu egiten da prozesamendu paraleloa eta banatua erabiliz

Aplikazioak

Finantza-aplikazioak

- Kreditu-txartelen iruzurrezko erabilerak hautematea
- Kreditu-txartelaren bidezko gastuaren iragarpena
- Kreditua ematerakoan arriskua aztertzea
- Datu historikoetatik merkatu-erregelak identifikatzea
- "Fidelak" ez diren bezeroak identifikatzea

Aplikazioak

Merkataritza

- Erosketa-saskiaren analisisia
- Publizitate-kanpainen ebaluazioa
- Bezeroen sailkatzea
- Salgaien, kostuen edota salmenten estimazioak egitea

Aplikazioak

Aseguruak

- Garesti gerta daitezkeen bezeroak identifikatzea
- Poliza berriak kontratatzen dituzten bezero-motak identifikatzea
- Arrisku egoeran egon daitezkeen bezeroen portaera-ereduak identifikatzea
- Iruzurrerako portaerak identifikatzea

Aplikazioak

Hezkuntza

- Ikasleen aukeraketa
- Eskola-porrota hautematea
- Ikasleek ikastegian emango duten denboraren estimazioa egitea

Aplikazioak

Medikuntza

- Gaixotasunak diagnostikatzea
- Gaixotasun jakin bat jasateko arriskuan egon daitezkeen pazienteak identifikatzea
- Ospitale-kudeaketarako. Baliabideen erabilera hobetzeko asmoz osasun-zentroetarako denbora-iragarpenak egitea
- Mediku-irudien tratamendua egitea

Aplikazioak

Bioinformatika, bioingeniaritza

- Geneen bilaketa (genoma kodifikatzen duten eremuak)
- Proteinen bigarren mailako egituraren iragarpena egitea
- Uholdeen iragarpena
- Uren kalitatea neurtzeko ereduak sortzea

Aplikazioak

Beste zenbait arlotan

- Telekomunikabideak: iruzurra identifikatzea
- Posta elektronikoa eta agenda pertsonalak: postaren sailkapena eta banaketa automatikoa, *spam* mezuak hautematea
- Ogasuna: zerga-iruzurrak hautematea
- Web: erabiltzaileen portaeraren analisia egitea
- Kirolak: datu medikuetatik abiatuz lesioak jasateko arriskua identifikatzea

Aurkibidea

1 Datu-Meatzaritza

2 Knowledge Discovery from Databases (KDD): datubaseetan ezagutza aurkitzea

Datubaseetan ezagutza aurkitzea

Prozesu iteratibo eta elkarreragilearen faseak

- 1 Datuak bildu eta bateratu
- 2 Aukeratu, garbitu eta eraldatu
- 3 Datu-meatzaritza egin
- 4 Ebaluatu eta interpretatu
- 5 Zabaldu eta erabili

Datubaseetan ezagutza aurkitzea

1. Datuak bildu eta bateratu

- **Datubaseak eta datuen prozesamendu tradizionala** (On-Line Transaction Processing, OLTP): eguneroko beharrak asetzeko nahikoak dira (fakturazioa, inbentarioen kontrola, ...)
- **Erabaki estrategikoak** analisisan, plangintzan eta iragarpenean oinarrituta: datuak sail desberdinetan egon daitezke
- **Jatorrizko datuak** formatu desberdinetan egoten dira
- Datubaseen bateratzea: **datuen biltegiak** (data warehousing)
- Datuen biltegiak gomendagarriak dira informazio-kantitatea handia denean. Hala ez denean ez dira beharrezkoak (**testu-fitxategiak, kalkulu-orriak, ...**)

Datubaseetan ezagutza aurkitzea

2. Aukeratu, garbitu eta eraldatu

- Aurkitutako ezagutzaren kalitatea datu-meatzaritzan erabilitako algoritmoaren mende egoteaz gain **aztertutako datuen kalitatearen mende** dago
- Datuen portaera orokorrera egokitzen ez den zenbait daturen presentzia egon daiteke (**outliers**)
- Balioak falta dituzten datuak egon daitezke (**missing values**)
- Aldagai esanguratsuak aukera daitezke (**feature subset selection**)
- Datubase oso handietan **kasuen auzazko aukeraketa** egin daiteke
- **Aldagai berriak** modu automatikoan eraiki daitezke
- Aldagai jarraiak **diskretiza** daitezke

Datubaseetan ezagutza aurkitzea

3. Datu-Meatzaritza

- **Eredu deskribatzaileak**
 - **Erregelak**
 - **Multzokatzea** (clustering): partizionala, probabilitikoa, ierarkikoa, kontzeptuala
- **Eredu iragarleak**
 - **Erregresioa**: regresio lineala...
 - **Gainbegiraturako sailkapena**: sailkapen-zuhaitzak, K-NN, sailkatzaile Bayestarrak, erregelen indukzioa, erregresio logistikoa, sare neuronalak, sailkatzaileen konbinaketa

Datubaseetan ezagutza aurkitzea

4. Ebaluatu eta interpretatu

- Ebaluazio teknikak: **balidazio sinplea** (entrenamendua + testa), ***k*-geruzatako balidazio gurutzatua**, ***bootstrapping***
- Erregelak: **estaltzea**, **konfidantza**
- Multzokatzea (Clustering): **multzo barruko elementuen eta multzoen arteko distantzia**
- Erregresioa: **batazbesteko errore kuadratikoa**
- Gainbegiraturako sailkapena: **ondo sailkatutakoen portzentaia**, **erroreen matrizea**, **ROC analisia**
- Eredu **ulerterrazak** eta **interesgarriak** (erabilgarriak eta berritzaileak)

Datubaseetan ezagutza aurkitzea

5. Zabalduek eta erabili

- **Zabalduek**: eraikitako ereduak erabiltzaileen artean zabalduek eta erabili erabakiak hartzeko
- **Ereduaren garapena** neurtu behar da denboran zehar:
 - Berrebaluatu
 - Berrentrenatu
 - Berreraiki

Oinarrizko bibliografia

- Liburua: **Introducción a la Minería de Datos**
 - Capítulo 1: ¿Qué es la minería de datos?
 - Capítulo 2: El proceso de extracción de conocimiento
- Egileak: José Hernández Orallo, M^a José Ramírez Quintana, César Ferri Ramírez
- Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia
- Argitaratzailea: Pearson Prentice Hall, 2004
- ISBN: 84-205-4091-9