

UNIVERSITATEA TITU MAIORESCU

Facultatea de Informatică

Conf. univ. dr. DANIELA JOIȚA

TEHNICI DE DATA MINING

Curs pentru învățământul la distanță



 editura
universității
Titu Maiorescu

BUCUREȘTI – 2014

Introducere

Acest material este destinat studenților anului III, învățământ la distanță, specializarea Informatică. Modul de prezentare are în vedere particularitățile învățământului la distanță, la care studiul individual este determinant. În timp ce **profesorul** sprijină studentul prin coordonarea învățării și prin feedback periodic asupra acumulării cunoștințelor și a deprinderilor, **studentul** alege locul, momentul și ritmul pentru studiu, dispune de capacitatea de a studia independent și totodată își asumă responsabilitatea pentru inițierea și continuarea procesului educațional.

Disciplina **Tehnici de data mining** utilizează noțiunile predate la disciplinele *Sisteme de gestiune a bazelor de date*, *Baze de date*, *Probabilități și statistică matematică* și *Inteligență artificială*, discipline studiate în anii I, II sau III.

Competențele dobândite de către studenți prin însușirea conținutului cursului **Tehnici de data mining** sunt des folosite la disciplinele de specialitate precum *Programare orientată pe obiecte*, *Tehnici avansate de programare*, *Proiectarea interfețelor grafice*, *Sisteme de gestiune a bazelor de date*, etc. O neînțelegere a noțiunilor fundamentale prezentate în acest curs poate genera dificultăți în asimilarea conceptelor mai complexe ce vor fi introduse în aceste cursuri de specialitate.

Principalele obiective ale disciplinei **Tehnici de data mining** sunt:

- Înțelegerea principalelor concepte, algoritmi și tehnici data mining
- Însușirea metodologiei data mining
- Selectarea unui criteriu corespunzător de evaluare a modelului data mining
- Cunoașterea a cel puțin unui pachet software specializat în data mining
- Implementarea într-un limbaj de programare a algoritmi specifici disciplinei

Competențele specifice disciplinei **Tehnici de data mining** se pot clasifica după cum urmează:

1. Cunoaștere și înțelegere

- Înțelegerea principalelor concepte, algoritmi și tehnici data mining
- Înțelegerea principalelor atuuri dar și limitări ale algoritmilor folosiți în data mining
- Cunoașterea a cel puțin unui pachet software specializat în data mining
- Cunoașterea principalelor operații data mining
- Însușirea metodologiei data mining

2. Explicare și interpretare <ul style="list-style-type: none"> • Explicarea și interpretarea conceptului de data mining • Interpretarea metodelor data mining de analiză a bazelor de date • Explicarea modalităților de funcționare a algoritmilor specifici disciplinei • Cunoașterea modului de alegere a tehnicilor data mining în funcție de problema de rezolvat
3. Instrumental - aplicative <ul style="list-style-type: none"> • Implementarea într-un limbaj de programare a algoritmi specifici disciplinei • Proiectarea aplicațiilor pentru rezolvarea unor probleme utilizând instrumente specifice disciplinei • Selectarea unui criteriu corespunzător de evaluare a modelului data mining • Corelarea cunoștințelor teoretice cu abilitatea de a le aplica în practică • Elaborarea unui proiect care să scoată în evidență importanța algoritmilor și tehnicilor specifice disciplinei precum și înțelegerea modalităților de aplicare a acestor tehnici
4. Atitudinale <ul style="list-style-type: none"> • Manifestarea unor atitudini favorabile față de știință și de cunoaștere în general • Formarea obișnuințelor de a recurge la concepte și metode informatice de tip algoritmic specifice în abordarea unei varietăți de probleme • Exprimarea unui mod de gândire creativ în structurarea și rezolvarea problemelor

Structura cursului este următoarea:

Unitatea de învățare 1. Introducere în data mining

Unitatea de învățare 2. Metode de clasificare

Unitatea de învățare 3. Asocieri

Unitatea de învățare 4. Clustering

Unitatea de învățare 5. Tanagra

Este foarte important ca parcurgerea materialului să se facă în ordinea unităților de învățare 1 – 4. Unitatea de învățare 5 este dedicată software-ului specializat în aplicarea tehnicilor de data mining și anume Tanagra, folosit la orele de laborator pentru acest curs. Această unitate recomandăm să fie parcursă în paralel cu celelalte unități de învățare. Fiecare UI (unitate de învățare) conține, pe lângă prezentarea notiunilor teoretice, exerciții rezolvate, activități de lucru individual la care sunt prezentate și indicații de rezolvare, exemple iar la finalul fiecărei lecții, un test de autoevaluare. În plus, la sfârșitul fiecărei UI sunt incluse probleme propuse care testează cunoașterea notiunilor teoretice de către student.

Pachet software recomandat:

TANAGRA este un open source software pentru data mining ce poate fi folosit gratuit pentru învățare și cercetare. Acesta va fi folosit la orele de laborator. În UI 5 este prezentat în detaliu acest software.

Bibliografia recomandată

1. M. Berry, G. S. Linoff, *Data Mining Techniques*, Wiley Publishing, 2004
2. I. Witten, F. Eibe, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999
3. A. Berson, S. Smith, K. Thearling, *Building Data Mining Applications for CRM*
4. D. Enachescu, *Tehnici statistice de data mining*, Editura Universitatii din Bucuresti, 2003
5. Ricco Rakotomalala, "TANAGRA : un logiciel gratuit pour l'enseignement et la recherche", in Actes de EGC'2005, RNTI-E-3, vol. 2, pp.697-702, 2005.
6. <http://freestatistics.altervista.org/reviews/tanagra.php>
7. <http://www.cs.ualberta.ca/~zaiane/courses/cmput695-04/work/A2-reports/tanagra.pdf>

Vă precizăm de asemenea că, din punct de vedere al verificărilor și al notării, cu adevărat importantă este capacitatea pe care trebuie să o dobândiți și să o probați de a rezolva toată tipologia de probleme aplicative aferente materialului teoretic prezentat în continuare. Prezentăm în continuare criteriile de evaluare și ponderea fiecărei activități de evaluare la stabilirea notei finale.

La stabilirea notei finale se iau în considerare	Ponderea în notare, exprimată în % {Total = 100%}
- răspunsurile la examen (evaluarea finală)	50%
- răspunsurile finale la lucrările practice de laborator	20%
- testarea periodică prin teme pentru acasă	10%
- testarea continuă pe parcursul semestrului	10%
- activitățile gen proiecte	10%
Modalitatea de evaluare finală: lucrare scrisă descriptivă și/sau probleme	
Cerințe minime pentru nota 5	Cerințe pentru nota 10
<ul style="list-style-type: none"> • Însușirea cunoștințelor de bază • Obținerea unui procent de cel puțin 45% din procentul maxim alocat fiecărei activități care se considera în stabilirea notei finale. • Activitatea în timpul semestrului 	<ul style="list-style-type: none"> • Rezolvarea corectă și completă a subiectelor de examen • Efectuarea corectă și completă a temelor pentru acasă • Participarea activă la curs și laborator • Elaborarea unui proiect corect, complet și bine documentat

În dorința de ridicare continuă a standardelor desfășurării activităților dumneavoastră, după fiecare unitate de învățare vă rugăm să completați un formular de feedback și să-l transmiteți îndrumătorului de an. Acest formular se găsește la sfârșitul acestui material

În speranța că organizarea și prezentarea materialului va fi pe placul dumneavoastră, vă urăm
MULȚ SUCCES!

Coordonator disciplină: Conf. univ. dr. Daniela Joița

Tutore: Conf. univ. dr. Daniela Joița

Cuprins

Introducere.....	1
UNITATEA DE ÎNVĂȚARE 1 Introducere în data mining.....	6
1.1 Lecția 1. Introducere	7
1.1.1 Motivație	7
1.1.2 Ce este data mining?	7
1.1.3 Metodologia data mining	8
1.1.4 Exemple de aplicare a metodologiei data mining	8
1.1.5 Legătura dintre data mining și alte domenii	10
1.1.6 Operații specifice data mining (task-uri)	10
UNITATEA DE ÎNVĂȚARE 2 Metode de clasificare.....	14
2.1 Lecția 1. Clasificare Bayesiană	15
2.2 Lecția 2. Arbori de decizie	20
2.2.1 Definiție	20
2.2.2 Construirea arborelui de decizie	22
2.2.3 Indicele Gini	23
2.2.4 Entropia (sau informația dobândită)	29
2.2.5 Testul Chi-patrat.....	34
UNITATEA DE ÎNVĂȚARE 3 Asocieri.....	40
3.1 Lecția 1. Reguli de asociere	41
3.1.1 Definiție	41
3.1.2 Mulțimi de obiecte frecvente	42

3.1.3	Generarea regulilor de asociere. Algoritmul Apriori	44
UNITATEA DE ÎNVĂȚARE 4 Clustering		49
4.1	Lecția 1. Algoritmul k-means	50
4.1.1	Definiție	50
4.1.2	Alg K-means (MacQueen 1967)	51
4.2	Lecția 2. Clusterizarea ierarhică. Metode aglomerative	54
4.2.1	Măsurarea distanței dintre clustere	55
4.3	Lecția 3. Clusterizarea ierarhică. Metode divizive	62
UNITATEA DE ÎNVĂȚARE 5 Tanagra		68
5.1	Lecția 1. Tanagra	69
5.1.1	Instalare	69
5.1.2	Interfața	69
5.1.3	Tutoriale	69
5.1.4	Baze de date	71
5.1.5	Importul datelor în TANAGRA	72
5.1.6	Precizări privind formatele diagramelor	72
5.1.7	Vizualizarea bazei de date	72
5.1.8	Folosirea operatorului Define status	73
5.1.9	Operatori	74
5.1.10	Alte soft-uri pentru data mining	74
FORMULAR DE FEEDBACK		75

UNITATEA DE ÎNVĂȚARE 1

Introducere în data mining

Obiective urmărite:

La sfârșitul parcurgerii acestei UI, studenții

- vor înțelege conceptul de data mining și de metodologie data mining
- vor cunoaște legătura dintre data mining și alte domenii
- vor ști să motiveze folosirea tehnicilor de data mining în aplicații
- vor cunoaște câteva aplicații ale metodelor și tehnicilor de data mining
- vor fi familiarizați cu principalele operații data mining.

Ghid de abordare a studiului:

Timpul mediu necesar pentru parcurgerea și asimilarea unității de învățare: 2h.

Lecțiile se vor parcurge în ordinea sugerată de diagramă.

Lecția

1

Rezumat:

În această UI sunt introduse conceptele de data mining și metodologie data mining. Se analizează legătura dintre data mining și alte domenii. Sunt prezentate principalele operații data mining. Sunt descrise câteva aplicații practice ale metodelor și tehnicilor de data mining.

Cuvinte cheie:

Data mining, clasificare, modelare, estimare, prezicere, prognoza, asociere, clustering, descriere, profil

1.1 Lecția 1. Introducere

1.1.1 Motivație

Odata cu explozia tehnologica, cu introducerea calculatoarelor in toate domeniile de activitate, s-a creat un urias bagaj de informatii si o retea de transfer de informatii.

Numai cand ridicam receptorul si formam un numar pentru a suna pe cineva, un calculator inregistreaza, nu convorbirea, ci momentul inceperii convorbirii, durata, numarul de telefon apelat, precum si numarul nostru de telefon, deci cel de unde a pornit apelul telefonic. Aceasta informatie, impreuna cu informatia abonatului postului telefonic si bineinteles impreuna cu informatiile despre alte convorbiri efectuate intr-o perioada data de timp, va genera factura telefonica a abonatului.

Din ce in ce mai multe date sunt generate prin tranzactii bancare, telefonice si mai ales prin tranzactiile de afaceri.

In concluzie, magaziiile de date (*data warehouses*) devin tot mai mari. Cercetarile firmei Winter Corporation, unul din cele mai importante centre de expertiza specializate in performanta si scalabilitatea celor mai mari sisteme de management al datelor, indica faptul ca dimensiunile bazelor de date cresc acum mai mult ca niciodata. Din 2003 si pana in 2005, dimensiunile celor mai mari magazii de date s-au triplat, depasind 100TB (terabytes). Va reamintesc ca 1 TB = 1024GB) Astfel, in 2005 companiile din Top Ten (cu cele mai mari baze de date comerciale) sunt Yahoo! (100TB) si AT&T (94TB) (AT&T una din cele mai mari companii de long distance telefonice din US). Iar aceste baze de date sunt mici in comparatie cu cele detinute de Google sau San Diego SuperComputer Centre, care sunt de ordinul de marime a petabytes-ilor (Va reamintesc ca 1PB = 1024 TB \cong 1 milion GB. Mai mult decât atât, într-un studiu realizat de **International Data Corporation (IDC)**, una din cele mai mari firme de furnizare de servicii de consulting, inteligență financiară și organizare de evenimente pentru piața de tehnologii informaționale și telecomunicații, se menționează că lumea a văzut creșterea volumului de date de la 5 hexaocteti = (5000 PB = 5 miliarde GB) in 2003 to 161 hexaocteti in 2006.

Cu aceasta crestere mare a dimensiunilor bazelor de date, este clar ca analiza unor astfel de date nu mai poate fi facuta manual ci trebuie sa existe o modalitate de analiza automata a datelor. Data mining ofera algoritmi si tehnicile necesare interpretarii bazelor de date de dimensiuni mari.

1.1.2 Ce este data mining?

Data mining este unul din cele mai noi si interesante domenii de cercetare. Data mining este de fapt aceasta analiza automata a datelor, in general a bazelor de date de dimensiuni mari, cu scopul de a descoperi tendinte, sabloane, tipare netriviiale, necunoscute anterior, uneori neasteptate, in date si care ar putea oferi informatii utile.

Tehnicile data mining pot face preziceri despre viitoare comportamente si trend-uri, permitand afacerilor sa ia decizii bazate pe cunostinte. Ele pot raspunde unor intrebari ca de exemplu: "Care sunt clientii firmei care vor raspunde, aproape cu siguranta, noii promotii si de ce?" sau "Unde ar trebui localizata noua sucursala a bancii?"

1.1.3 Metodologia data mining

Metodologia folosita se numeste modelare. Modelarea este, spus simplificat, procesul de creare a unui model pentru o situatie in care se cunoaste raspunsul si de aplicare a acelui model intr-o situatie in care nu se cunoaste raspunsul. Acest proces de modelare este bine cunoscut de sute de ani, inainte de aparitia calculatoarelor. Modul in care se face modelarea folosind calculatoarele, modul in care metodologia data mining creaza modele nu este chiar atat de diferit de modul in care oamenii creaza modele. Calculatoarele sunt incarcate cu informatii despre situatii variate pentru care se cunoaste raspunsul si soft-ul data mining cauta caracteristicile datelor care trebuie sa intre in model. Odata construit modelul poate fi aplicat in situatii in care nu se cunoaste raspunsul.

De exemplu, sa spunem ca sunteti directorul de marketing al unei companii de telecomunicatii si vreti sa racolati clienti noi. Puteti foarte simplu sa va faceti cunoscuti prin pliante pe care sa le distribuiti prin posta sau in magazinele specializate, intregii populatii sau puteti sa folositi experienta de afaceri inmagazinata deja in baza dvs de date pentru a crea un model. Ca director de marketing aveti acces la informatii despre aproape toti clientii firmei, precum: varsta, sex, adresa, modul de folosire a serviciului (daca isi platesc factura la timp, daca folosesc serviciul des, adica daca vorbesc mult la telefon si cat profit aduc firmei). Sigur ca ati prefera sa racolati noi clienti care sa fie cat mai profitabili (sa zicem, cu cat mai multe convorbiri telefonice). Folosind datele despre clientii actuali, puteti crea un model pe care il puteti testa folosind tot baza de date pe care o aveti despre clientii actuali. Pentru a face asta, in procesul de modelare data mining, anumite date se pun deoparte, deci se foloseste doar o parte a magaziei de date, parte care trebuie sa fie suficient de cuprinzatoare, totusi, se creaza modelul si se testeaza pe datele care au fost puse deoparte, pentru a-l valida. (De exemplu, se folosesc datele din baza de date de acum 5 ani pentru construirea modelului si apoi se testeaza pe datele din baza de date de anul trecut). Abia apoi se aplica pentru luarea de noi decizii privind promotia ce se doreste. De exemplu, un foarte simplu model ar spune ca 75% din clientii care au un venit $\geq x$ milioane au o factura lunara ≥ 2 milioane. Folosind acest model directorul de marketing poate decide sa se adreseze promotional numai unei parti a populatiei.

1.1.4 Exemple de aplicare a metodologiei data mining

Sunt numeroase domeniile de activitate in care se pot aplica tehnicile data mining:

- in stiinta: astronomie, medicina
- in domeniul afacerilor (comercial): managementul relatiei cu clientii (CRM –customer relationship management), comertul on-line, telefonie, sport si entertainment, marketing, investitii

- Web: motoare de cautare, text si web mining

Exemple:

1. Churn (customer attrition) – renuntarea unui client la serviciile oferite de bussines-ul respectiv
(un client al unei companii renunta la serviciile de telefonie ale acesteia – paraseste compania- nu stim de ce si nici daca o va face)

Sa presupunem urmatoarea situatie:

O companie telefonica mobila observa ca anual 25-30% din clientii sai renunta la serviciile companiei. Ce isi propune compania? Avand informatiile despre clientii sai din ultimele n luni, sa prezica ce clienti vor renunta la servicii in luna (lunile) urmatoare si cunoscand lucrul acesta, sa estimeze asa numita valoare a clientului (prin valoare intelegand, de exemplu, client bun (foloseste des serviciile, plateste la timp, factura este mare etc.) sau in caz contrar, client cu valoare mica) si in functie de valoarea estimata sa pregateasca o oferta sau o modalitate de a-l pastra pe client, sa faca ceva.

2. credit risc

Sa consideram urmatoarea situatie: O persoana aplica pentru un credit bancar. Ce risc isi asuma banca? Ar trebui sa ii aprobe sau nu creditul? Bancile folosesc pentru aceasta o serie de programe pentru estimarea riscului si luarea unei decizii.

3. comertul on-line

Amazon.com este una din cele mai mari companii de vanzari online. La inceput vindea carti, apoi s-a extins la CD de muzica, electronice si alte produse. Amazon.com are un grup care aplica tehnici data mining pentru o mai buna functionare a firmei.

Situatie: Cineva vrea sa cumpere o carte. Atunci i se ofera sa cumpere si alte carti, pe care probabil le va cumpara, mai exact, carti cu cea mai mare probabilitate de cumparare de catre acel client. S-a dovedit ca acest program prin care i se recomanda cumparatorului si alte posibile optiuni de cumparare este de succes si chiar se fac cercetari de gasire a unor programe si mai avansate.

4. medicina

Printr-un pateneriat, IBM si Mayo Clinic, una din cele mai renumite clinici din US, si din lume de altfel, cu spitale in trei state din US, lucreaza din 2004 la un proiect de data mining cu scopul de a gasi metode de tratament individualizate pentru fiecare pacient. Proiectul isi propune sa gaseasca sabloane, tendinte legate de modul in care pacientii, in functie de varsta, bolile pe care le-au avut, precum si alti factori, raspund la diverse tratamente. Aceasta va permite doctorilor sa aleaga tratamentul cel mai bun pentru un pacient, tratamentul care a avut cel mai mare succes pentru pacienti cu aceleasi similaritati ca si pacientul in cauza. Se spera ca pana in 2014, medicii clinicii Mayo sa poata folosi aplicatiile acestui proiect.

Deci ideea este de fapt de a aplica cunostintele despre mai multi pacienti, pentru ca un pacient sa beneficieze de cel mai bun tratament. Se propune de asemenea imbunatatirea bazei de date cu informatii genetice ale pacientilor.

5. detectarea unor infractiuni, ca de exemplu, folosirea unor carti de credit in mod fraudulos, sau efectuarea unor convorbiri telefonice frauduloase.

Aproape toate tranzactiile efectuate cu cartile de credit sunt scanate cu ajutorul unor algoritmi specifici, pentru a identifica tranzactiile suspicioase. De exemplu, cineva din US si care a folosit cartea numai in US, are o tranzactie ce este facuta in Romania. Sigur ca acest lucru devine suspicios. Marile companii de telefonie folosesc software specializate pentru identificarea convorbirilor frauduloase.

1.1.5 Legatura dintre data mining si alte domenii

Unii spun ca data mining nu este decat un nume dat statisticii. Este adevarat ca multe din tehnicile folosite in data mining provin din alte domenii: statistica, informatica, inteligenta artificiala. Cu toate acestea, exista diferente intre aceste domenii si data mining. Statistica ofera un suport teoretic pentru studiul evenimentelor intamplatoare si metode pentru testarea ipotezelor, dar nu studiaza preprocesarea datelor sau vizualizarea rezultatelor, ce fac parte din data mining. Inteligenta artificiala are o alta metoda de aprofundare, mai euristica si se concentreaza pe imbunatatirea performantei agentilor de invatare. Data mining se concentreaza pe intregul proces de descoperire de cunostinte, de la organizarea datelor si eliminarea celor incomplete, invatare si cunoastere prin descoperire si pana la vizualizarea rezultatelor.

1.1.6 Operatii specifice data mining (task-uri)

In data mining exista doua orientari: metodologie data mining directionata si nedirectionata. In cadrul aplicarii metodologiei directionate, se propune explicarea sau divizarea pe categorii a unui atribut al bazei de date. In cea nedirectionata se propune descoperirea de sabloane, relatii sau similaritati intre grupuri de inregistrari fara ca sa se foloseasca o colectie de clase predefinite sau un camp tinta.

Principalele operatii (numite de cele mai multe ori task-uri) care pot fi efectuate cu data mining sunt:

- a. Clasificare
- b. Estimare
- c. Prezicere (proгноza)
- d. Asociere
- e. Clustering
- f. Descriere si profil

Primele trei sunt exemple de metodologie data mining directionata, in care se propune gasirea valorii unui atribut tinta (target variable). Asocierea si clustering sunt operatii nedirectionate, in care se propune descoperirea unor structuri, sabloane in baza de date. Descrierea este o operatie descriptiva ce poate fi si directionata si nedirectionata.

a. Clasificare

A clasifica inseamna a examina trasaturile si caracteristicile unui obiect si a-l repartiza unui set de clase predefinite. Obiectele care sunt caracterizate in general sunt reprezentate de inregistrarile unei baze de date sau fisier, iar clasificarea inseamna a adauga o noua coloana (un nou atribut) cu un cod al unei clase de un anumit tip si a determina pentru fiecare inregistrare care este clasa careia ii apartine. Clasificarea se evidentiaza printr-o caracterizare foarte bine definita a claselor si o multime de training ce consta in exemple preclasificate. Clasificarea consta in construirea unui model care sa poata fi aplicat unor date neclasificate inca tocmai pentru a putea fi clasificate. Tehnicile data mining folosite pentru clasificare sunt arborii de decizie si tehnicile de tipul cel mai apropiat vecin.

b. Estimarea

In timp ce clasificarea lucreaza cu rezultate de tip discret, estimarea lucreaza cu rezultate cu valori continue. Date anumite date de intrare, estimarea determina o valoare necunoscuta inca pentru o variabila de tip continuu. Estimarea se foloseste foarte des pentru a clasifica inregistrarile. De exemplu, sa presupunem ca o firma de telefonie doreste sa vanda spatiu pentru reclama in plicurile pe care le trimite lunar clientilor sai impreuna cu factura, unei firme care comercializeaza CD-uri cu diferite soft-uri. oferte postale promotionale Firma ce comercializeaza CD-uri poate sa cumpere spatiu publicitar pentru 50000 de plicuri. Cea mai proasta solutie ar fi sa trimita aceste plicuri unor clienti alesi la intamplare, de calculator de exemplu. O alta solutie ar fi sa construiasca un model de clasificare in care clientii vor fi clasificati daca folosesc sau nu serviciul de internet oferit de firma de telefonie, apoi sa trimita plicurile cu publicitate unor clienti alesi tot la intamplare dar din clasa celor care folosesc internet-ul. O alta posibilitate ar fi sa se construiasca un model care, bazat pe mult mai multe date, sa poata estima pentru fiecare client care ar fi probabilitatea de a detine un calculator si apoi clientii sa fie ordonati descrescator in functie de aceasta probabilitate, iar primii 50000 clienti cu cele mai mari probabilitati de a detine un calculator sa fie cei alesi pentru a primi oferta publicitara. Tehnicile data mining folosite pentru estimare sunt: regresie si rețele neurale.

c. Prezicere sau prognoza

Prezicerea este la fel ca si clasificarea si estimarea, doar ca inregistrarile sunt clasificate in raport cu o comportare viitoare sau estimate in raport cu o valoare viitoare. De exemplu, prezicerea acelor clienti care sunt pasibili sa raspunda unui produs nou, prezicerea venitului unei persoane bazandu-ne pe detalii personale.

d. Asociere

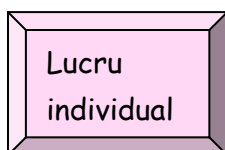
Asocierea este operatia de a determina ce lucruri pot fi grupate impreuna. De exemplu, determinarea produselor care sunt cumparate impreuna intr-un supermarket. Asocierea este o modalitate de a genera anumite reguli de forma "daca se intampla acest lucru atunci se va intampla si acest lucru (cu probabilitatea x)". De exemplu, "Persoanele care cumpara mancare pentru pisici vor cumpara si un vas de mancare pt pisici cu probabilitatea p1" sau "Persoanele care cumpara un vas de mancare pt pisici vor cumpara si mancare pentru pisici cu probabilitatea p2" sau "Persoanele care cumpara o planta vor cumpara si pamant pentru flori in 60% din cazuri iar, ambele produse au fost cumparate impreuna in 5% din cazuri" sau "Persoanele care cumpara o papusa Barbie vor cumpara si o ciocolata cu probabilitatea 60%".

e. Clustering

Clustering este operația de segmentare a unei multimi heterogene într-un număr de subgrupuri mai omogene numite clustere. Clustering se deosebește de clasificare prin aceea că nu se bazează pe clase predefinite. În cazul operației de clustering nu există multe de training sau exemple predefinite. Înregistrările sunt grupate după similarități. Clustering este de obicei o operație preludiv la o altă operație de data mining. De exemplu, clustering este prima operație care se execută în dezvoltarea de strategii de segmentare a pieței. În loc să ne punem întrebarea "Care este tipul de promoție la care clienții răspund cel mai bine?", întâi împărțim clienții în clustere cu obiceiuri de a cumpăra produse asemănătoare și apoi adresăm întrebarea "Care este tipul de promoție la care clienții dintr-un anumit cluster răspund cel mai bine?"

f. Descriere și profil

Uneori scopul metodologiei data mining este de a descrie ce se întâmplă, care sunt tendințele în baza de date, pentru o mai bună înțelegere a persoanelor, produselor și proceselor care au dus la producerea datelor din baza de date. Printre tehnicile folosite se numără arborii de decizie, regulile de asociere și clustering.



În România domeniile în care se aplică tehnicile de data mining sunt din ce în ce mai multe. Prezentați două astfel de aplicații.

Indicații de rezolvare: Căutați pe internet "data mining în România" și veți găsi articole interesante.

Test de autoevaluare

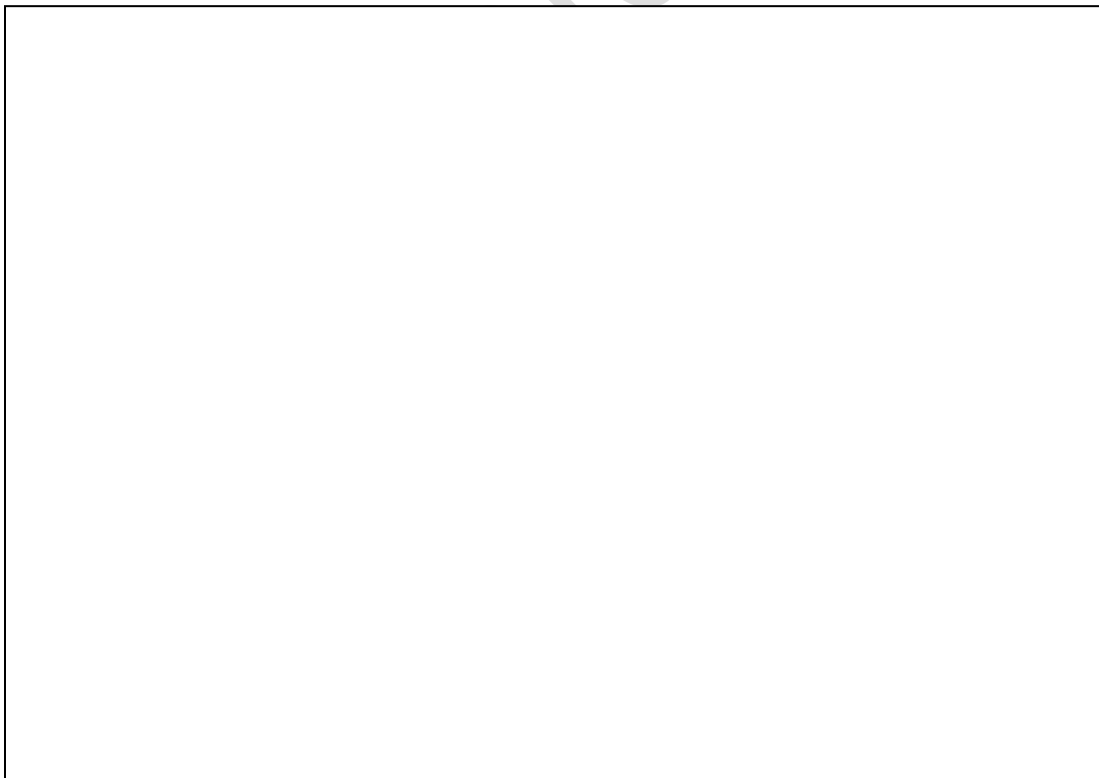


1. Ce este data mining?

2. Dați două exemple de aplicare a metodologiei data mining.



3. Care este relația dintre data mining și statistică?



UNITATEA DE ÎNVĂȚARE 2

Metode de clasificare

Obiective urmărite:

La sfârșitul parcurgerii acestei UI, studenții

- vor înțelege conceptul de clasificare
- vor cunoaște modul de aplicare a clasificării bayesiene unor probleme reale
- vor înțelege noțiunea de arbore de decizie
- vor putea să construiască arbori de decizie folosind unul din criteriile: indicele Gini, informația dobândită, testul chi-pătrat.

Ghid de abordare a studiului:

Timpul mediu necesar pentru parcurgerea și asimilarea unității de învățare: 10h.

Lecțiile se vor parcurge în ordinea sugerată de diagramă.

Lecția
1

Lecția
2

Rezumat:

În această UI este introdus conceptul de clasificare și sunt prezentate două metodologii de clasificare: clasificare bayesiană și clasificarea prin arbori de decizie. Clasificarea prin arbori de decizie este prezentată în detaliu prin trei criterii de construire a lor: indicele Gini, informația dobândită, testul chi-pătrat.

Cuvinte cheie:

clasificare, clasificare bayesiană, Teorema lui Bayes, variabile dependente, variabile independente, ipoteze, mulțime de training, atribut țintă, arbore de decizie, indicele Gini, entropie, informația dobândită, diviziunea unui nod, cel mai bun atribut

2.1 Lectia 1. Clasificare Bayesiană

Problema: Data o multime de exemple preclasificate numita multime de training, trebuie sa construim un model pentru a clasifica cazuri noi.

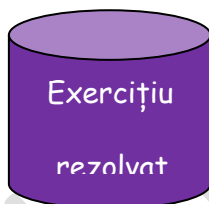
Presupunem ca toate atributele sunt la fel de importante si ca sunt independente.

Date o multime de valori: $X = x_1, x_2, \dots, x_n$ si ipotezele $H = h_1, h_2, \dots, h_m$, presupunem ca o singura ipoteza poate sa apara in acelasi timp si ca x_i este un eveniment observabil, regula Bayes este urmatoarea:

Probab ca o ipoteza h sa fie adevarata dat fiind un eveniment observabil x este:

$$P(h/x) = \frac{P(x/h)P(h)}{P(x)}$$

unde $P(x)$ = probab de aparitie a evenimentului x , $P(x/h)$ = probab ca data ipoteza h , evenimentul x sa apara. $P(h/x)$ se numeste probab aposteriori, $P(h)$ se numeste probab apriori.



Dată mulțimea de training de mai jos pe baza căreia se poate decide dacă se acordă sau nu credit la cumpărarea unui produs unui client în funcție de venitul lunar al acestuia (atribut continuu numit Venit lunar) și de indicele de creditare (atribut discret numit Credit), să se aplice clasificarea bayesiană pentru a determina valoarea atributului țintă Clasa.

Avem 4 ipoteze: $H = h_1, h_2, h_3$ unde h_1 = se acorda creditul, h_2 = se acorda creditul dar cu anumite restrictii, h_3 = nu se acorda creditul.

Multimea de training este urmatoarea:

ID	Venit lunar	Credit	Clasa
1	4800	Excelent	h1
2	2800	Bun	h1
3	1200	Excelent	h1
4	2400	Bun	h1
5	4400	Bun	h1
6	1600	Excelent	h1
7	3200	Nesatisfacator	h2
8	1600	Nesatisfacator	h3
9	2400	Nesatisfacator	h2
10	200	Nesatisfacator	h3

Transformam atributul Venit lunar din variabila continua in variabila discreta prin introducerea urmatorului atribut: Interval_venit:

- 1 corespunde intervalului [0, 400)
- 2 corespunde intervalului [400, 2000)
- 3 corespunde intervalului [2000, 4000]
- 4 corespunde intervalului [4000, infinit).

Tabelul de mai sus devine:

ID	Venit lunar	Credit	Clasa	Interval_venit
1	4800	Excelent	h1	4
2	2800	Bun	h1	3
3	1200	Excelent	h1	2
4	2400	Bun	h1	3
5	4400	Bun	h1	4
6	1600	Excelent	h1	2
7	3200	Nesatisfacator	h2	3
8	1600	Nesatisfacator	h3	2
9	2400	Nesatisfacator	h2	3
10	200	Nesatisfacator	h3	1

Atunci multimea evenimentelor observabile este:

$$X = \{(1, Excelent), (1, Bun), (1, Nesatisf), (2, Excelent), \dots, (2, Nesatisf), \dots, (4, Excelent), \dots, (4, Nesatisf)\}$$

Pentru o data de intrare noua, sa zicem: $\{Venit = 5200, Credit = Excelent\}$ ne punem intrebarea

carei clase ii apartine. Cum Venit = 5200 apartine intervalului de venit 4, vom calcula

$$P(h_i / \{Interval_venit = 4, Credit = Excelent\}) \text{ pentru fiecare } i=1,2,3 \text{ si ii vom atribui clasa pentru}$$

care probab este cea mai mare, adica: h_j astfel incat

$$P(h_j / \{Interval_venit = 4, Credit = Excelent\}) = \max_{1 \leq i \leq 3} P(h_i / \{Interval_venit = 4, Credit = Excelent\})$$

Pentru calculul $P(h_i / \{Interval_venit = 4, Credit = Excelent\})$ folosim regula Bayes:

$$P(h_i / \{Interval_venit = 4, Credit = Excelent\}) = \frac{P(\{Interval_venit = 4, Credit = Excelent\} / h_i) P(h_i)}{P(\{Interval_venit = 4, Credit = Excelent\})}$$

Cum am presupus ca attributele sunt independente:

$$P(\{Interval_venit = 4, Credit = Excelent\} / h_i) = P(\{Interval_venit = 4\} / h_i) \cdot P(\{Credit = Excelent\} / h_i)$$

si

$$P(\{Interval_venit = 4, Credit = Excelent\}) = P(\{Interval_venit = 4\}) \cdot P(\{Credit = Excelent\})$$

Deci trebuie sa determinam

$$\max \left\{ \frac{P(\{Interval_venit = 4\} / h_i) \cdot P(\{Credit = Excelent\} / h_i) \cdot P(h_i)}{P(\{Interval_venit = 4\}) \cdot P(\{Credit = Excelent\})} / 1 \leq i \leq 3 \right\}$$

sau, pentru ca fractiile au acelasi numitor,

$$\max \{P(\{Interval_venit = 4\} / h_i) \cdot P(\{Credit = Excelent\} / h_i) \cdot P(h_i) / 1 \leq i \leq 3\}$$

Calculam probabilitatile:

H	P(H)
h_1	$6/10=0.6$
h_2	$2/10=0.2$
h_3	$2/10=0.2$

Pentru fiecare atribut:

H	h_1	h_2	h_3
Interval_venit			
1	$0/6=0$	$0/2=0$	$1/2$
2	$2/6$	$0/2=0$	$1/2$
3	$2/6$	$2/2=1$	$0/2=0$
4	$2/6$	$0/2=0$	$0/2=0$

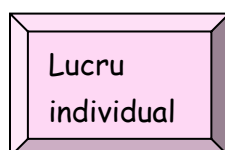
H	h_1	h_2	h_3
Credit			
Excelent	$3/6=1/2$	$0/2=0$	$0/2=0$
Bun	$3/6=1/2$	$0/2=0$	$0/2=0$
Nesatisfacator	$0/6=0$	$2/2=1$	$2/2=1$

$$P(\{Interval_venit = 4, Credit = Excelent\} / h_1) = \frac{2}{6} \cdot \frac{1}{2} = \frac{1}{6} \quad \text{trebuia inmultit si cu } p(h_1) \text{ adica } (1/6) \cdot (1/2) \cdot (6/10) = 1/10$$

$$P(\{Interval_venit = 4, Credit = Excelent\} / h_2) = 0 \cdot 0 = 0$$

$$P(\{Interval_venit = 4, Credit = Excelent\} / h_3) = 0 \cdot 0 = 0$$

Rezulta ca clasa care se atribuie datei noi este h_1 .



Aplicați metoda de clasificare bayesiană bazei de date următoare, ce corespunde datelor meteorologice ale unei zile. Atributul tinta este "Play" și corespunde deciziei dacă într-o zi dată sunt condiții favorabile unui joc de baseball.

Outlook	Temp	Humidity	Windy	Play
sunny	hot	high	F	N
sunny	hot	high	T	N
overcast	hot	high	F	Y
rain	mild	high	F	Y
rain	cool	normal	F	Y
rain	cool	normal	T	N
overcast	cool	normal	T	Y
sunny	mild	high	F	N
sunny	cool	normal	F	Y
rain	mild	normal	F	Y
sunny	mild	normal	T	Y
overcast	mild	high	T	Y
overcast	hot	normal	F	Y
rain	mild	high	T	N

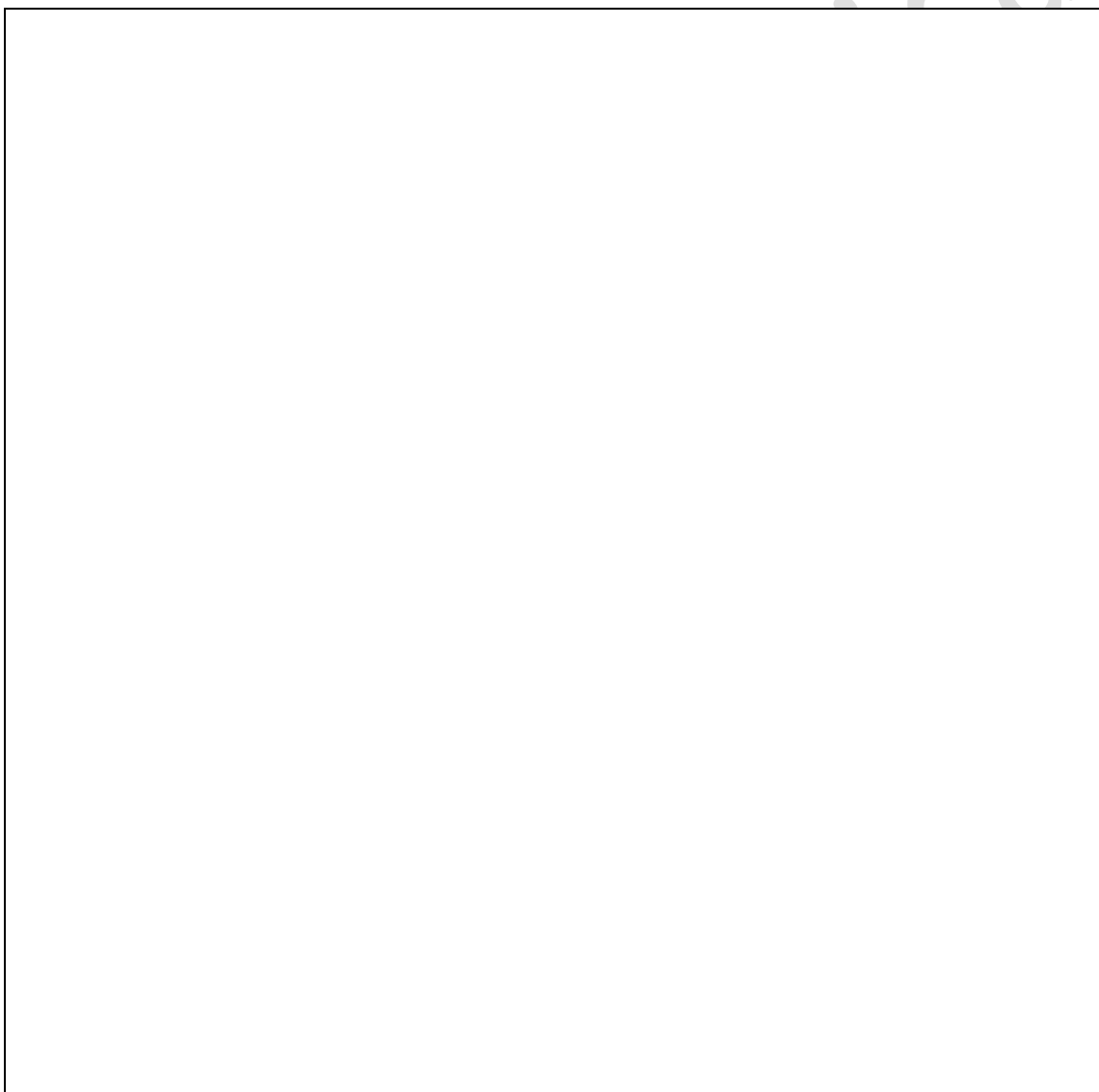
Test de autoevaluare

2

- Presupunem că la medic vine un pacient care strănută. Medicul știe că pacientul poate avea o simplă răceală sau o rinită alergică. Cunoscând următoarele probabilități: $P(\text{răceală}) = 0.02$, $P(\text{rinită alergică}) = 0.04$, $P(\text{strănută}) = 0.08$, $P(\text{strănută}|\text{răceală}) = 0.98$, $P(\text{strănută}|\text{rinită alergică}) = 0.75$ ajutați-l pe medic să ia o decizie folosind clasificarea bayesiană.

- Se da următoarea baza de date, cu attributele a_1 , a_2 și atributul tinta Clasa. Folosind clasificarea Bayesiană, să se estimeze carei clase (Y sau N) ar aparține următoarea înregistrare: $\{a_1 = \text{mare}, a_2 = \text{gri}\}$.

a1	a2	Clasa
mare	alb	Y
mic	negru	Y
mic	alb	N
mare	negru	Y
mijlociu	gri	N
mijlociu	negru	Y



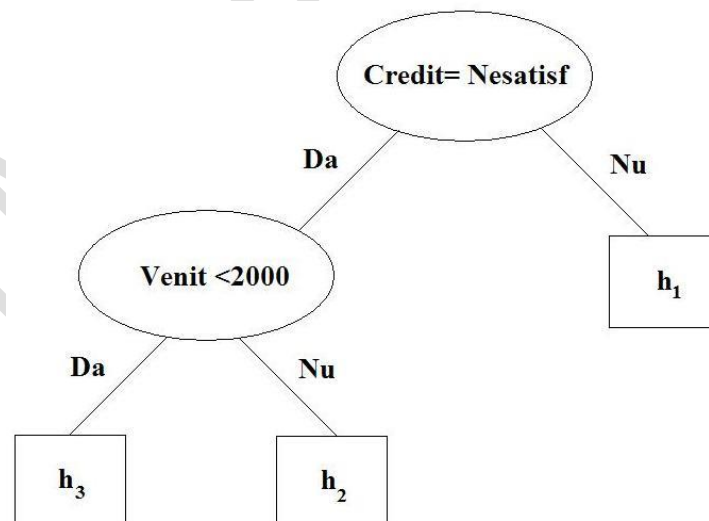
2.2 Lectia 2. Arbori de decizie

2.2.1 Definiție

Un **arbore de decizie** este o structura care poate fi folosita pentru a imparti o colectie mare si eterogena de inregistrari intr-un sir de colectii din ce in ce mai mici si mai omogene in raport cu un atribut tinta. Divizarea colectiei se face prin aplicarea unui sir de reguli de decizie simple.

Ca orice arbore in teoria grafurilor, arborele de decizie are drept componente noduri, ramuri, frunze si de reprezinta cu ramurile in jos, plecand de la radacina. Nodurile interne reprezinta un test asupra unui atribut, ramurile reprezinta o valoare posibila a testului, iar frunzele reprezinta modul de clasificare (clasa careia ii apartine colectia de inregistrari din nodul respective) sau modul de distributie a clasei in nodul respective.

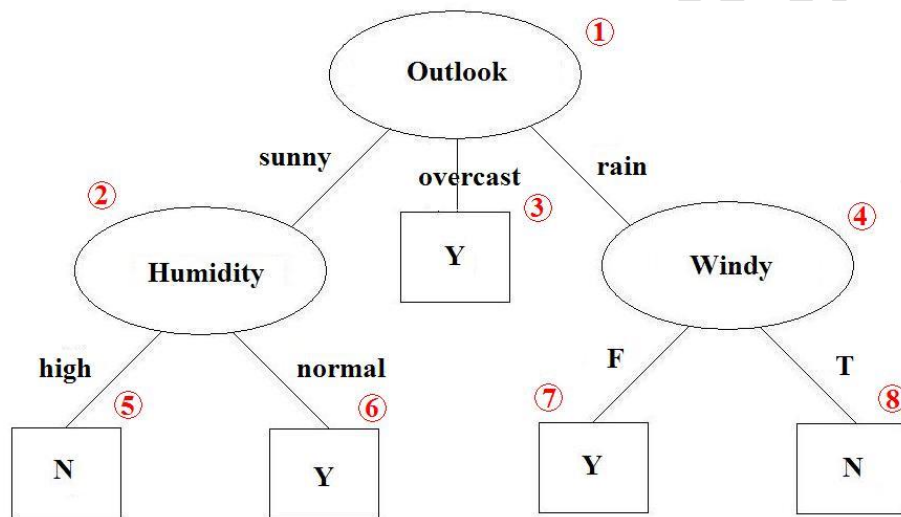
Exemplu 1: Pentru baza de date de mai sus pentru determinarea acordarii unui credit, un arbore de decizie asociat este:



Exemplu 2: Un arbore de decizie pentru baza de date urmatoare, in care atributul tinta este "Play", este:

Outlook	Temp	Humidity	Windy	Play
---------	------	----------	-------	------

sunny	hot	high	F	N
sunny	hot	high	T	N
overcast	hot	high	F	Y
rain	mild	high	F	Y
rain	cool	normal	F	Y
rain	cool	normal	T	N
overcast	cool	normal	T	Y
sunny	mild	high	F	N
sunny	cool	normal	F	Y
rain	mild	normal	F	Y
sunny	mild	normal	T	Y
overcast	mild	high	T	Y
overcast	hot	normal	F	Y
rain	mild	high	T	N



Se testeaza intai valoarea atributului Outlook. In functie de rezultatul obtinut se imparte arborele in trei ramuri, cate una pentru fiecare raspuns posibil.

Observam ca nodul 1 contine toate cele 14 inregistrari, nodul 2 contine inregistrarile pentru care Outlook= sunny, nodul 3 contine inregistrarile pentru care Outlook= overcast si observam ca nodul 3 este nod terminal, toate inregistrarile fiind clasificate cu Y adica apartinand clasei Play = Y, nodul 5 contine inregistrarile pentru care Outlook= sunny si Humidity=high.

Un nod poate avea doua sau mai multe ramuri.

In procesul decizional, o inregistrare intra in arbore la radacina si va iesi printr-un nod terminal clasificata. Un drum de la radacina la o frunza este o expresie care exprima regula folosita pt clasificarea inregistrarii. De exemplu, inregistrarea: {sunny, cool, high, T} parcurge in ordine nodurile 1, 2, 5 si iese cu clasificarea Play=N, regula folosita fiind: {Outlook= sunny & Humidity=high}. Frunze diferite pot fi clasificate la fel, chiar daca din motive diferite. De exemplu, nodurile 3 si 6 sunt clasificate la fel.

Daca atributul tinta este o var discreta atunci arb de decizie se numeste arbore de clasificare, daca este o var continua atunci arb de decizie se numeste arbore de regresie.

2.2.2 Construirea arborelui de decizie

La inceputul procesului de construire exista o multime de training constand din inregistrarile preclasificate. Scopul este de a construi un arbore de decizie care atribuie o clasa(sau probab de a apartine unei clase) unei inreg noi, bzanu-ne pe valorile atributelor de intrare.

Algoritmul general este de tip Greedy.

Arborele este construit de sus in jos recursive, in maniera Divide et Impera. Atributele de intrare sunt discretizate in prealabil.

- La inceput, in radacina arborelui se afla toate inreg multimii de training.
- Se selecteaza atributul care da cea mai buna impartire a nodului radacina.
- Se partitioneaza multimea datelor conform valorilor testului efectuat asupra atributului selectat.
- Pt. fiecare partitie se repeat pasii de mai sus.

Conditii de oprire a divizarii unui nod:

- Toate inreg nodului apartin aceleasi clase
- Nu mai sunt atribute pt a putea face divizarea (se aege clasa cu cele mai multe inregistrari)
- Nu mai exista inregistrari.

Criterii pentru alegerea atributului care da cea mai buna divizare a unui nod:

Se masoara puritatea divizarii = o masura cu valori intre 0 (daca orice 2 inreg nu sunt in aceeasi clasa) si 1 (toate inreg sunt in aceeasi clasa).

Alegerea atributului de divizare se face in functie de rtipul atributului. Daca atributul tinta este discret:

- Indicele Gini
- Entropia (sau informatia dobandita)
- Testul Chi-patrat

Daca atributul tinta este continuu:

- Reducere fluctuatiei
- Testul F

2.2.3 Indicele Gini

Daca T este o multime de date ce contine inreg din n clase (deci atributul tinta are n valori discrete posibile), indicele Gini se defineste:

$$gini(T) = 1 - \sum_{j=1}^n p_j^2$$

unde p_j = frecventa relative a clasei j in T , $p_j = \frac{nr\ aparitiiale\ clasei\ j}{nr\ inreg\ ale\ colectiei\ T}$

Daca T se imparte in submultimile T_1, T_2, \dots, T_k astfel incat T are N elemente, T_1 are N_1 elemente, ..., T_k are N_k elemente, atunci indicele *Gini* al diviziunii se defineste

$$gini_{diviziune}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2) + \dots + \frac{N_k}{N} gini(T_k)$$

Se enumara toate posibilitatile de divizare pentru fiecare atribut si se alege atributul cu cel mai mic indice $gini_{diviziune}(T)$.

Indicele Gini ne da probabilitatea ca doua elemente alese la intamplare din colectia data sa nu fie in aceeasi clasa.

Observati ca daca T contine numai inreg dintr-o clasa atunci

$$gini(T) = 1 - \left[\left(\frac{0}{N} \right)^2 + \dots + \left(\frac{0}{N} \right)^2 + \left(\frac{N}{N} \right)^2 \right] = 0$$

Daca $n=2$ si T contine acelasi numar de inreg din clasa 1 ca si din clasa 2 atunci

$$gini(T) = 1 - \left[\left(\frac{N/2}{N} \right)^2 + \left(\frac{N/2}{N} \right)^2 \right] = 1 - 1/2 = 1/2$$



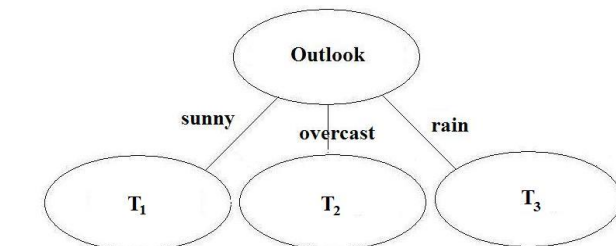
Sa construim arborele de decizie asociat exemplului 2 folosind indicele Gini.

Atribute de intrare: Outlook, Temp, Humidity, Windy

Atribut tinta: Play cu valori posibile {Y, N}.

Cautam atributul cel mai bun pentru divizarea radacinii.

1. Calculam gini (diviziune dupa Outlook):



	Play	
Outlook	Y	N
Sunny	2	3
Overcast	4	0
Rain	3	2

$$gini_{diviziune \text{ Outlook}}(T) = \frac{5}{14} gini(T_1) + \frac{4}{14} gini(T_2) + \frac{5}{14} gini(T_3)$$

$$gini(T_1) = 1 - \left[\left(\frac{2}{5} \right)^2 + \left(\frac{3}{5} \right)^2 \right] = 12/25$$

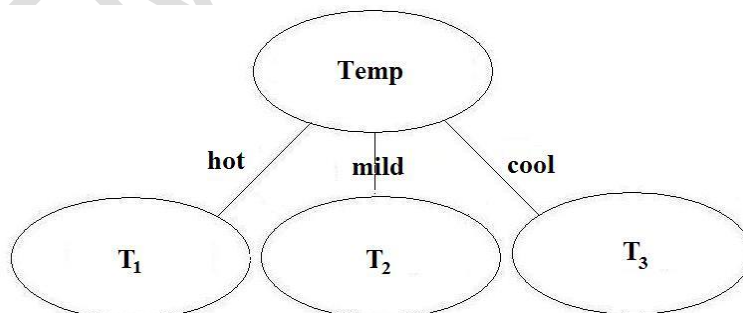
$$gini(T_2) = 1 - \left[\left(\frac{4}{4} \right)^2 + \left(\frac{0}{4} \right)^2 \right] = 0$$

$$gini(T_3) = 1 - \left[\left(\frac{3}{5} \right)^2 + \left(\frac{2}{5} \right)^2 \right] = 12/25$$

Deci:

$$gini_{diviziune \text{ Outlook}}(T) = \frac{5}{14} \cdot \frac{12}{25} + \frac{4}{14} \cdot 0 + \frac{5}{14} \cdot \frac{12}{25} = 0.34$$

2. Calculam gini (diviziune dupa Temp):



	Play	
Temp	Y	N
hot	2	2
mild	4	2
cool	3	1

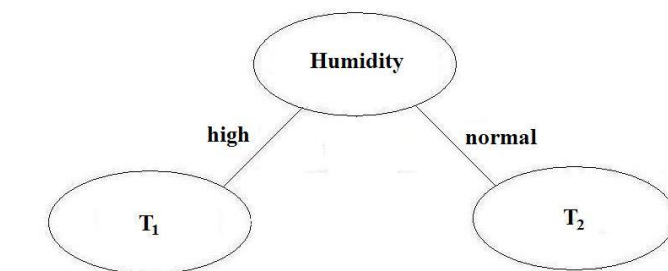
$$gini_{diviziune \text{ Temp}}(T) = \frac{4}{14} gini(T_1) + \frac{6}{14} gini(T_2) + \frac{4}{14} gini(T_3)$$

$$gini(T_1) = 1 - \left[\left(\frac{2}{4} \right)^2 + \left(\frac{2}{4} \right)^2 \right] = 1/2 \quad gini(T_2) = 1 - \left[\left(\frac{4}{6} \right)^2 + \left(\frac{2}{6} \right)^2 \right] = 4/9$$

$$gini(T_3) = 1 - \left[\left(\frac{3}{4} \right)^2 + \left(\frac{1}{4} \right)^2 \right] = 6/16$$

$$gini_{diviziune \ Temp}(T) = \frac{4}{14} \cdot \frac{1}{2} + \frac{6}{14} \cdot \frac{4}{9} + \frac{4}{14} \cdot \frac{6}{16} = 0.44$$

3. Calculam gini (diviziune dupa Humidity):



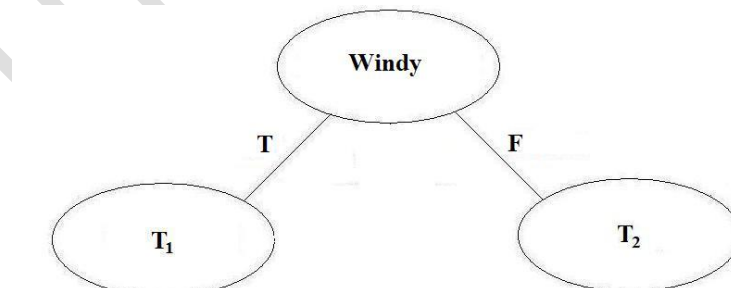
Humidity \ Play	Y	N
high	3	4
normal	6	1

$$gini_{diviziune \ Humidity}(T) = \frac{7}{14} gini(T_1) + \frac{7}{14} gini(T_2)$$

$$gini(T_1) = 1 - \left[\left(\frac{3}{7} \right)^2 + \left(\frac{4}{7} \right)^2 \right] = 24/49 \quad gini(T_2) = 1 - \left[\left(\frac{6}{7} \right)^2 + \left(\frac{1}{7} \right)^2 \right] = 12/49$$

$$gini_{diviziune \ Humidity}(T) = \frac{1}{2} \cdot \frac{24}{49} + \frac{1}{2} \cdot \frac{12}{49} = 0.37$$

4. Calculam gini (diviziune dupa Windy):



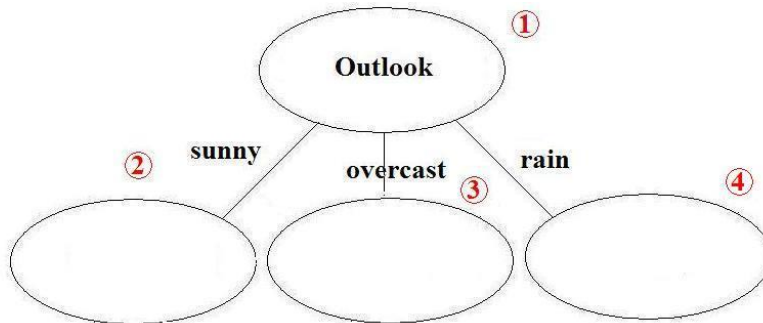
Windy \ Play	Y	N
T	3	3
F	6	2

$$gini_{diviziune\ Windy}(T) = \frac{6}{14} gini(T_1) + \frac{8}{14} gini(T_2)$$

$$gini(T_1) = 1 - \left[\left(\frac{3}{6} \right)^2 + \left(\frac{3}{6} \right)^2 \right] = 1/2 \quad gini(T_2) = 1 - \left[\left(\frac{6}{8} \right)^2 + \left(\frac{2}{8} \right)^2 \right] = 6/16$$

$$gini_{diviziune\ Windy}(T) = \frac{6}{14} \cdot \frac{1}{2} + \frac{8}{14} \cdot \frac{6}{16} = 0.43$$

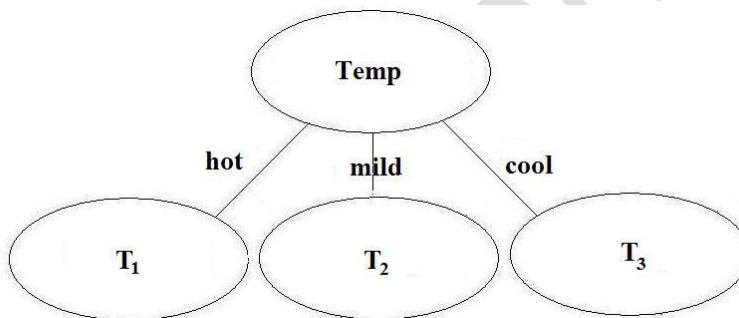
Se alege atributul cu cel mai mic indice $gini_{diviziune}(T)$ deci se va alege atributul Outlook:



Urmeaza **diviziunea nodului 2**.

Alegerea atributului:

1. Calculam gini (diviziune dupa Temp):



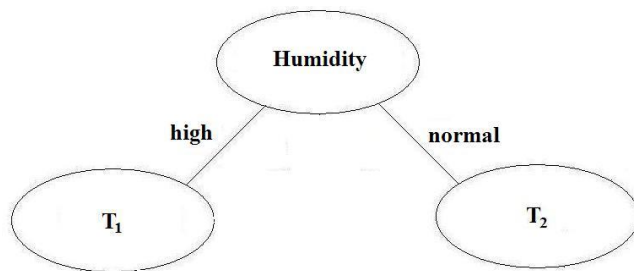
Temp \ Play	Y	N
hot	0	2
mild	1	1
cool	1	0

$$gini_{diviziune\ Temp}(T) = \frac{2}{5} gini(T_1) + \frac{2}{5} gini(T_2) + \frac{1}{5} gini(T_3)$$

$$gini(T_1) = 0 \quad gini(T_2) = 1/2 \quad gini(T_3) = 0$$

$$gini_{diviziune\ Temp}(T) = \frac{2}{5} \cdot \frac{1}{2} = 0.2$$

2. Calculam gini (diviziune dupa Humidity):

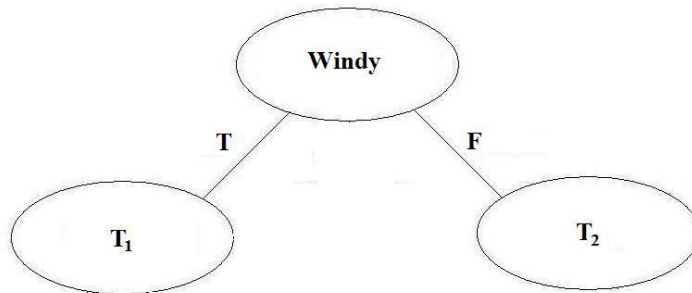


Humidity \ Play	Y	N
high	0	3
normal	2	0

$$gini(T_1)=0 \quad gini(T_2)=0$$

$$\text{Deci } gini_{\text{diviziune Humidity}}(T)=0$$

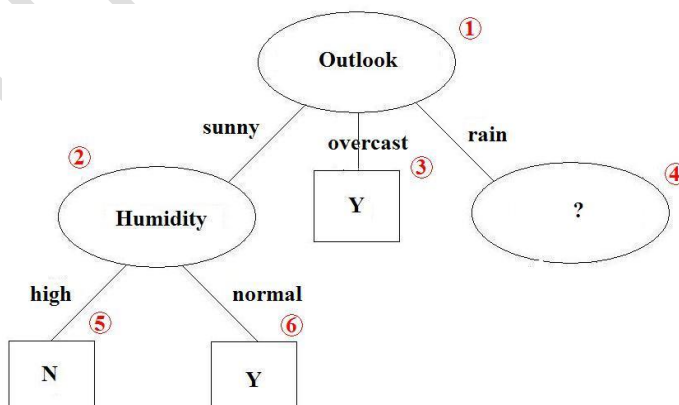
3. Calculam gini (diviziune dupa Windy):



Windy \ Play	Y	N
T	1	1
F	1	2

Cum $gini(T_1)=1/2$ rezulta ca $gini_{\text{diviziune Windy}}(T)>0$

Deci atributul cu indicele Gini cel mai mic este Humidity:

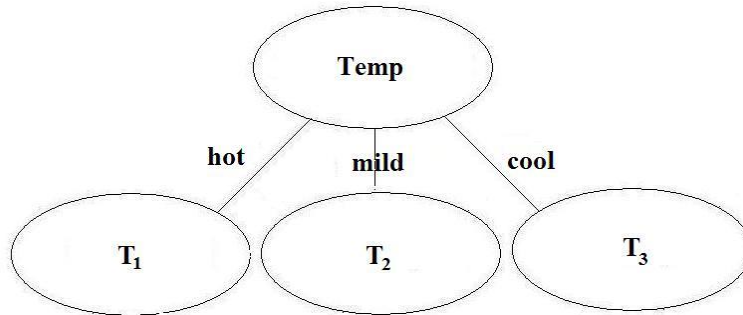


Nodurile 5, 6 si 3 sunt noduri terminale pt ca toate inreg apartin aceleasi clase.

Urmeaza **diviziunea nodului 4.**

Alegerea atributului:

1. Calculam gini (diviziune dupa Temp):



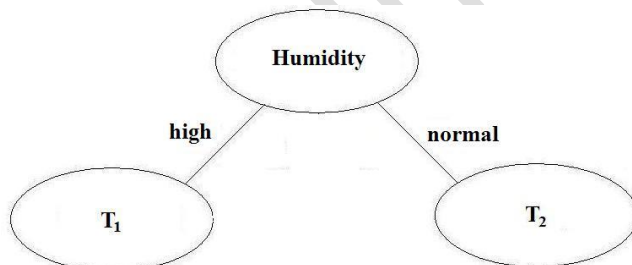
Temp \ Play	Y	N
hot	0	0
mild	2	1
cool	1	1

$$gini_{diviziune \ Temp}(T) = \frac{0}{5} gini(T_1) + \frac{3}{5} gini(T_2) + \frac{2}{5} gini(T_3)$$

$$gini(T_1) = 0 \quad gini(T_2) = 1 - \left[\left(\frac{2}{3} \right)^2 + \left(\frac{1}{3} \right)^2 \right] = 4/9 \quad gini(T_3) = 1/2$$

$$gini_{diviziune \ Temp}(T) = \frac{3}{5} \cdot \frac{4}{9} + \frac{2}{5} \cdot \frac{1}{2} = 0.47$$

3. Calculam gini (diviziune dupa Humidity):



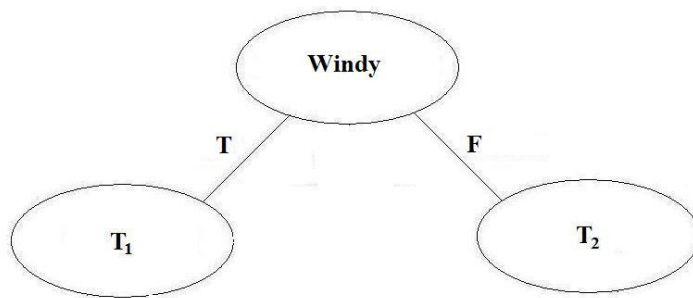
Humidity \ Play	Y	N
high	1	1
normal	2	1

$$gini_{diviziune \ Humidity}(T) = \frac{2}{5} gini(T_1) + \frac{3}{5} gini(T_2)$$

$$gini(T_1) = 1/2 \quad gini(T_2) = 1 - \left[\left(\frac{2}{3} \right)^2 + \left(\frac{1}{3} \right)^2 \right] = 4/9$$

$$Deci \quad gini_{diviziune \ Humidity}(T) = \frac{2}{5} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{4}{9} = 0.46$$

4. Calculam gini (diviziune dupa Windy):



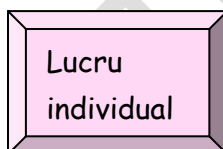
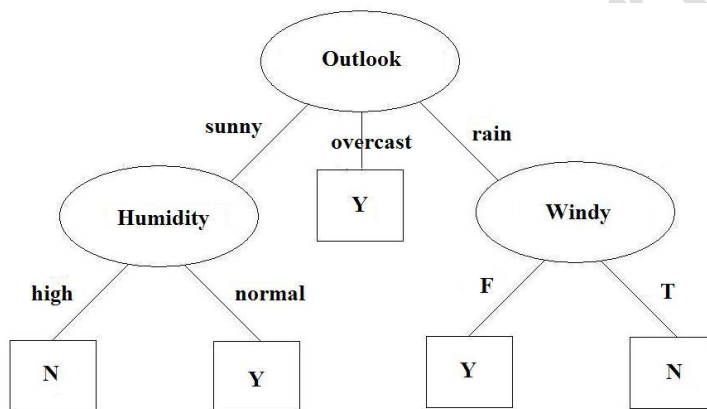
Windy \ Play	Y	N
	0	2
T	0	2
F	3	0

$$gini(T_1)=0 \quad gini(T_2)=0$$

$$\text{Deci } gini_{\text{diviziune Windy}}(T)=0$$

Deci atributul cu indicele Gini cel mai mic este Windy.

Arborele de decizie este:



Sa se construiască arborele de decizie asociat mulțimii de training din Lecția 1. Clasificarea bayesiană pe baza căreia se poate decide dacă se acordă sau nu credit la cumpărarea unui produs unui client în funcție de venitul lunar al acestuia și de indicele de creditare folosind criteriul indicele Gini.

2.2.4 Entropia (sau informatia dobandita)

În teoria informației entropia este o măsură a modului de dezorganizare a unui sistem.

Dacă T este o mulțime de date ce conține înreg din n clase (decă atributul tinta are n valori discrete posibile), entropia se definește:

$$\text{entropie}(T) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \dots - p_n \log_2 p_n$$

unde p_j = frecventa relative a clasei j in T , $p_j = \frac{nr\ aparitii\ ale\ clasei\ j}{nr\ inreg\ ale\ colectiei\ T}$

Daca T se imparte in submultimile T_1, T_2, \dots, T_k astfel incat T are N elemente, T_1 are N_1 elemente, ..., T_k are N_k elemente, atunci

$$entropie_{diviziune}(T) = \frac{N_1}{N} entropie(T_1) + \frac{N_2}{N} entropie(T_2) + \dots + \frac{N_k}{N} entropie(T_k)$$

Definim informatia dobandita a diviziunii:

$$INFO_{divizare} = entropie(T) - entropie_{divizare}(T) = entropie(T) - \sum_{j=1}^k \frac{N_j}{N} entropie(T_j)$$

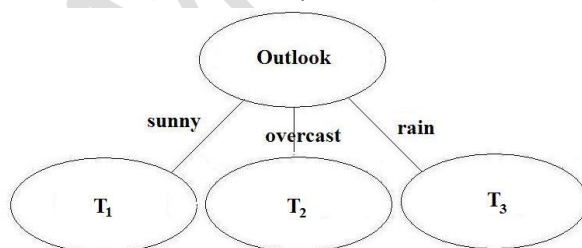
La construirea arborelui de decizie se alege atributul cu informatia dobandita cea mai mare.



Sa construim arborele de decizie asociat exemplului 2 folosind criteriul informatia dobandita.

Cautam atributul cel mai bun pentru divizarea radacinii.

5. Calculam $INFO(\text{diviziune dupa Outlook})$:



	Play	
Outlook \ Play	Y	N
Sunny	2	3
Overcast	4	0
Rain	3	2

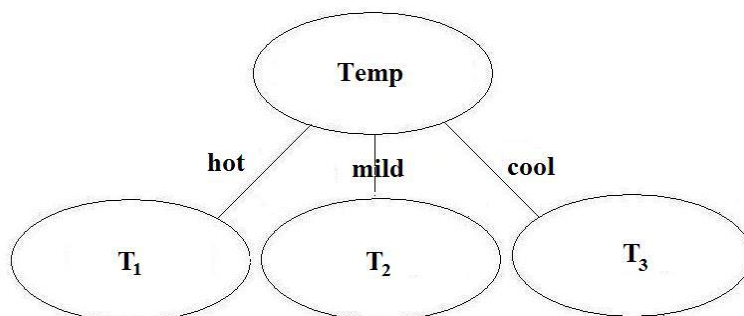
$$entropie(T) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

$$entropie(T_1) = entropie(T_3) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} = 0.97$$

$$\text{entropia}(T_2) = 0$$

Deci

6. Calculam INFO (diviziune dupa Temp):



Temp \ Play	Y	N
hot	2	2
mild	4	2
cool	3	1

$$\text{entropia}(T) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.94$$

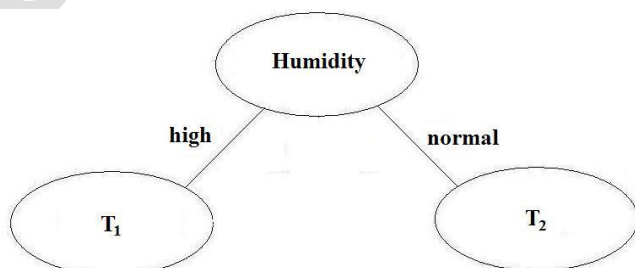
$$\text{entropia}(T_1) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} = 1$$

$$\text{entropia}(T_2) = -\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 0.92$$

$$\text{entropia}(T_3) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.81$$

$$\text{INFO}_{\text{divizareTemp}} = \text{entropia}(T) - \left[\frac{4}{14} \text{entropia}(T_1) + \frac{6}{14} \text{entropia}(T_2) + \frac{4}{14} \text{entropia}(T_3) \right] = 0.028$$

7. Calculam INFO (diviziune dupa Humidity):



Humidity \ Play	Y	N
high	3	4

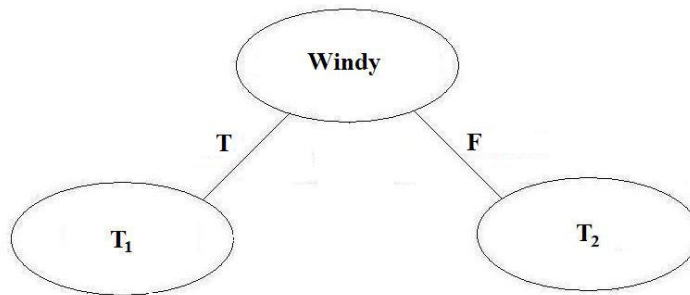
normal	6	1
--------	---	---

$$\text{entropie}(T_1) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = 0.98$$

$$\text{entropie}(T_2) = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.59$$

$$INFO_{\text{divizareHumidity}} = \text{entropia}(T) - \left[\frac{1}{2} \text{entropia}(T_1) + \frac{1}{2} \text{entropia}(T_2) \right] = 0.152$$

8. Calculam INFO (diviziune dupa Windy):



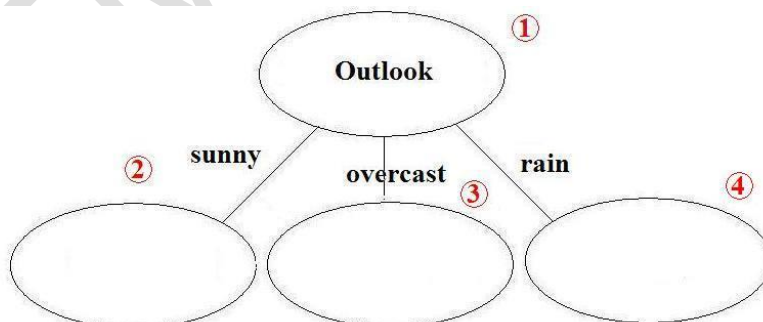
	Play	
Windy	Y	N
T	3	3
F	6	2

$$\text{entropie}(T_1) = 1$$

$$\text{entropie}(T_2) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} = 0.81$$

$$INFO_{\text{divizareHumidity}} = \text{entropia}(T) - \left[\frac{6}{14} \text{entropia}(T_1) + \frac{8}{14} \text{entropia}(T_2) \right] = 0.048$$

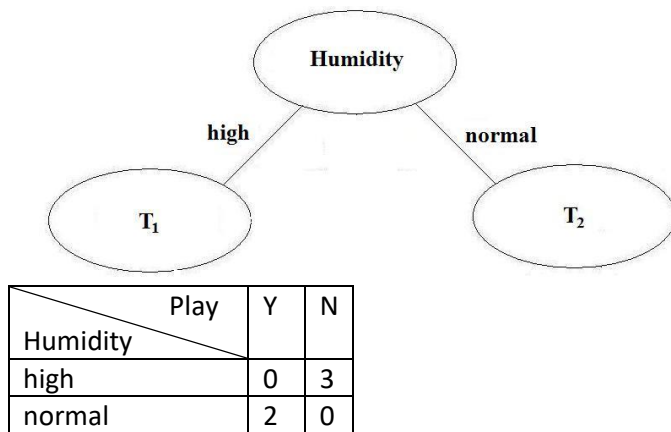
Se alege atributul cu informatia dobandita cea mai mare, deci alegem atributul Outlook:



Urmeaza **diviziunea nodului 2**.

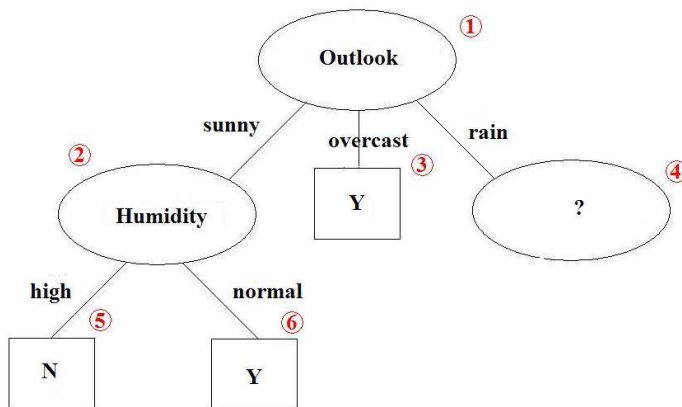
Alegerea atributului:

1. Calculam INFO(diviziune dupa Humidity):



Cum $entropie(T_1) = entropie(T_2) = 0$ rezulta ca $INFO_{divizareHumidity} = entropie(nod2)$ si este maxima pentru acest nod.

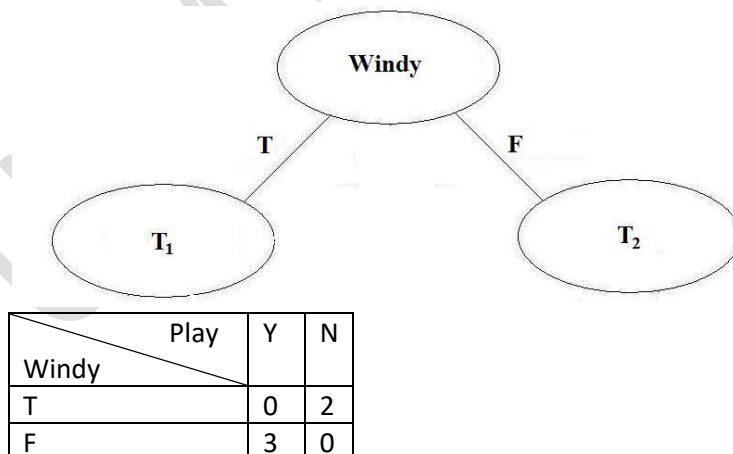
Deci se alege Humidity ca atribut de divizare.



Urmeaza **diviziunea nodului 4.**

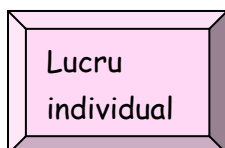
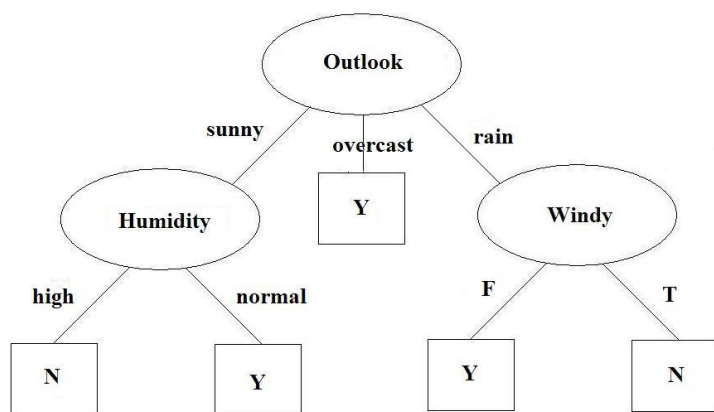
Alegerea atributului:

1. Calculam INFO (diviziune dupa Windy):



Cum $entropie(T_1) = entropie(T_2) = 0$ rezulta ca $INFO_{divizareWindy} = entropie(nod4)$ si este maxima pentru acest nod. Deci se alege Windy ca atribut de divizare.

Va rezulta arborele de decizie:



Sa se construiască arborele de decizie asociat mulțimii de training din Lecția 1. Clasificarea bayesiană pe baza căreia se poate decide dacă se acordă sau nu credit la cumpărarea unui produs unui client în funcție de venitul lunar al acestuia și de indicele de creditare folosind criteriul informația dobândita.

2.2.5 Testul Chi-patrat

Un alt criteriu care se aplica variabilelor de intrare discrete este testul Chi-patrat. Acesta se folosește pentru a determina care sunt variabilele cu maxima importanta pentru o clasificare.

Testul Chi –patrat se folosește la testarea independenței a doua variabile discrete, măsurând corelația dintre cele două variabile.

Dacă X, Y sunt două variabile:

$X = \{x_1, x_2, \dots, x_n\}$ și $Y = \{y_1, y_2, \dots, y_m\}$ cu tabela de corelație asociată lor: $F = (f_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}}$

se definește:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} \text{ observat} - f_{ij} \text{ estimat})^2}{f_{ij} \text{ estimat}}$$

unde $f_{ij} \text{ observat}$ = frecvența de apariție a evenimentului $\{X = x_i, Y = y_j\}$

$$f_{ij}^{estimat} = \frac{total(linia\ i) \cdot total(coloana\ j)}{nr.\ total\ observatii} = \frac{\sum_{j=1}^m f_{ij} \cdot \sum_{i=1}^n f_{ij}}{\sum_{i=1}^n \sum_{j=1}^m f_{ij}}$$

Pentru a testa ipoteza: X si Y sunt independente, adica $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$ se calculeaza χ^2 .

Daca ipoteza este adevarata, χ^2 va avea o distributie Chi-patrat cu $(n-1)(m-1)$ grade de libertate.

Pentru aceasta se va calcula p-valoarea distributiei, prin cautare in tabele statistice sau se poate folosi calculatorul de Chi-patrat ce se gaseste in pagina de web: www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

Data valoare lui χ^2 pentru un set de experimente cu gradul de libertate d, p valoare este probabilitatea ca ipoteza sa fie adevarata numai datorita sansei. Cu cat p este mai mic cu atat sansele ca ipoteza sa fie adevarata sunt mai mici si deci cu cat p este mai mic cu atat X si Y sunt mai strans legate.

Dacă $p < 0.05$ se respinge ipoteza nulă (nonexistența legăturii dintre X si Y) și se consideră ca fiind adevărată ipoteza alternativă (existența legăturii).



Sa construim arborele de decizie asociat exemplului 2 folosind testul chi-pătrat.

Pentru baza de date din exemplul 2 de mai sus, daca alegem X= Windy si Y=atributul tinta= Play, atunci avem urmatoarea tabela de corelatie:

	Play	Y	N	Total
Windy				
T		3	3	6
F		6	2	8
Total		9	5	14

$$Deci \quad f_{observat} = \begin{pmatrix} 3 & 3 \\ 6 & 2 \end{pmatrix}, \quad f_{estimat} = \begin{pmatrix} \frac{9 \cdot 6}{14} & \frac{5 \cdot 6}{14} \\ \frac{9 \cdot 8}{14} & \frac{5 \cdot 8}{14} \end{pmatrix} = \begin{pmatrix} 3.857 & 2.143 \\ 5.142 & 2.857 \end{pmatrix}$$

$$Deci \quad \frac{(f_{observat} - f_{estimat})^2}{f_{estimat}} = \begin{pmatrix} 0.19 & 0.343 \\ 0.143 & 0.257 \end{pmatrix} \text{ rezulta ca } \chi^2 = 0.933$$

Pentru Windy si Play: $\chi^2 = 0.933$ rezulta $p=0.334$
La fel se poate determina corelatia dintre celelalte variabilele si atributul tinta:

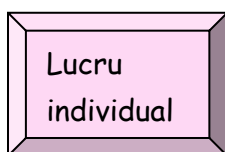
Pentru Outlook si Play: $\chi^2 = 3.547$ rezulta $p=0.1697$

Pentru Temp si Play: $\chi^2 = 0.57$ rezulta $p=0.752$

Pentru Humidity si Play: $\chi^2 = 2.8$ rezulta $p=0.0942$

Se alege atributul pentru care p este cel mai mic adica Humidity.

Se continua ca si pentru celelalte metode de construire a arborelui de decizie.



Sa se construiască arborele de decizie asociat mulțimii de training din Lecția 1. Clasificarea bayesiană pe baza căreia se poate decide dacă se acordă sau nu credit la cumpărarea unui produs unui client în funcție de venitul lunar al acestuia și de indicele de creditare folosind criteriul testul chi-pătrat.

Test de autoevaluare



- 1) Se dă o colecție de 100 de persoane care au participat la un dineu. Dintre aceștia 30 s-au îmbolnăvit de o viroză respiratorie iar restul nu.
 - a) Care este indicele Gini asociat colecției de persoane?
 - b) Care este entropia colecției de persoane?

- 2) Se considera problema divizarii unui nod ce contine 20 de inregistrari, distribuite in mod egal din punctul de vedere al atributului tinta, atribut ce poate lua doua valori distincte: T sau F. Divizarea nodului se poate face dupa doua atribute:

Atrib 1 cu doua valori posibile distincte: x_1 , x_2

Atrib 2 cu doua valori posibile distincte: y_1 , y_2

Tabelele de corelatie dintre fiecare atribut si atributul tinta sunt urmatoarele:

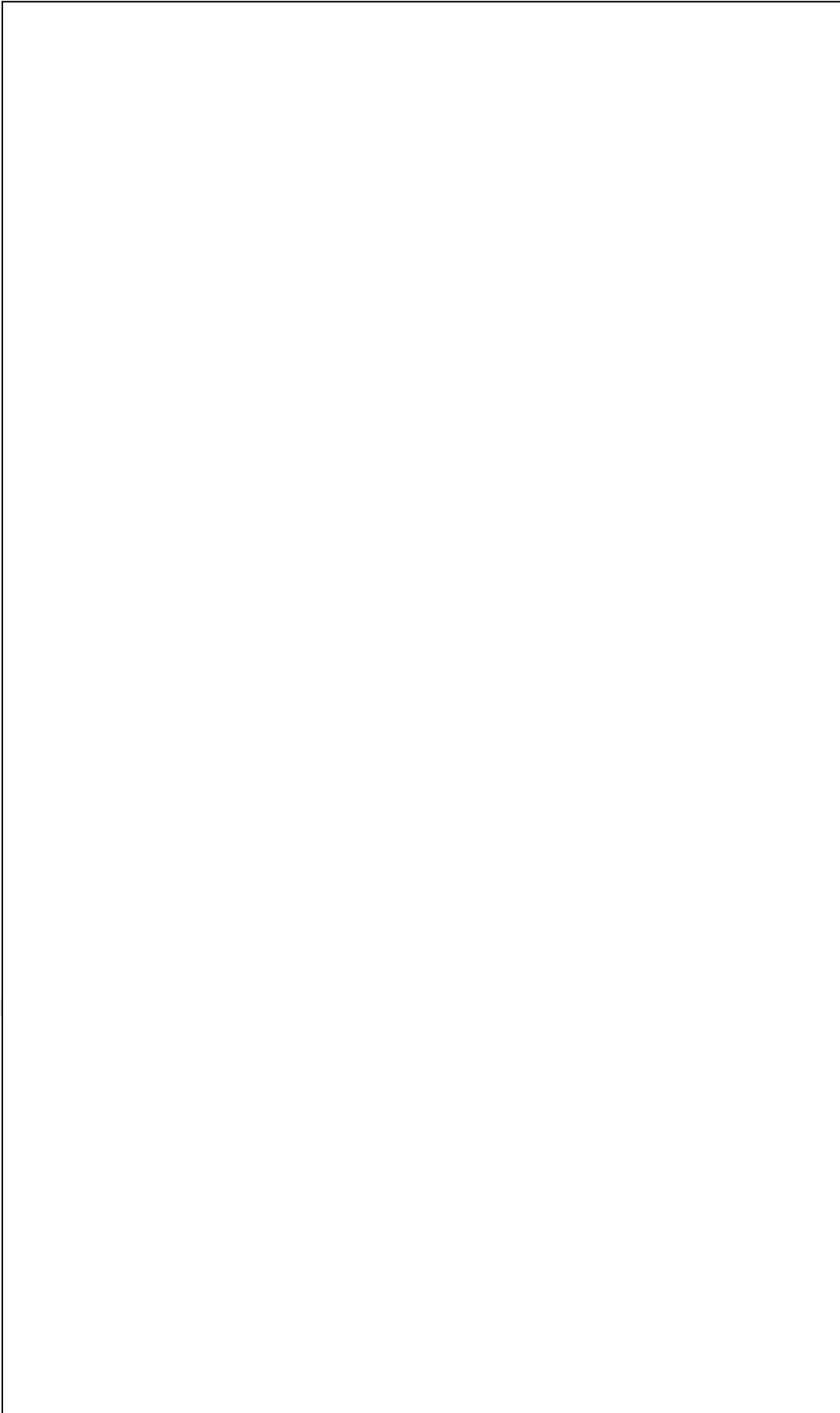
Atribut tinta \ Atrib 1	T	F
x_1	3	7
x_2	7	3

Atribut tinta \ Atrib 2	T	F
y_1	10	5
y_2	0	5

Sa se raspunda la intrebarea:

Care din cele doua divizari vor fi preferate si de ce, daca arborele de decizie se construieste folosind:

- indicele Gini
- informatia dobandita sau entropia
- testul chi-pătrat.



Probleme

proapse

1. Dată mulțimea de training de mai jos pe baza căreia se poate decide dacă un client va cumpăra un calculator, clasificați un nou client cu informațiile următoare: $\text{varsta} \leq 30$, $\text{venit} = \text{mediu}$, $\text{student} = \text{da}$, $\text{credit} = \text{satisfăcător}$, folosind clasificarea bayesiană.

ID	vârsta	venit	student	credit	Cumpără calculator
1	≤ 30	ridicat	NU	satisfăcător	NU
2	≤ 30	ridicat	NU	excelent	NU
3	31..40	ridicat	NU	satisfăcător	DA
4	> 40	mediu	NU	satisfăcător	DA
5	> 40	scăzut	DA	satisfăcător	DA
6	> 40	scăzut	DA	excelent	NU
7	31..40	scăzut	DA	excelent	DA
8	≤ 30	mediu	NU	satisfăcător	NU
9	≤ 30	scăzut	DA	satisfăcător	DA
10	> 40	mediu	DA	satisfăcător	DA
11	≤ 30	mediu	DA	excelent	DA
12	31..40	mediu	NU	excelent	DA
13	31..40	ridicat	DA	satisfăcător	DA
14	> 40	mediu	NU	excelent	NU

2. Dată mulțimea de training de la exercițul 1, sa se construiască arborele de decizie asociat folosind indicele Gini.
3. Dată mulțimea de training de la exercițul 1, sa se construiască arborele de decizie asociat folosind criteriul informației dobândite.
4. Dată mulțimea de training de la exercițul 1, sa se construiască arborele de decizie asociat folosind testul chi-pătrat.

UNITATEA DE ÎNVĂȚARE 3

Asocieri

Obiective urmărite:

La sfârșitul parcurgerii acestei UI, studenții

- vor înțelege conceptele de mulțime frecventă, regulă de asociere
- vor cunoaște algoritmul Apriori
- vor cunoaște modul de aplicare a algoritmului Apriori unor probleme reale
- vor putea să identifice regulile de asociere tari.

Ghid de abordare a studiului:

Timpul mediu necesar pentru parcurgerea și asimilarea unității de învățare: 6h.

Lecțiile se vor parcurge în ordinea sugerată de diagramă.

Lecția

1

Rezumat:

În această UI sunt introduse conceptele de mulțime frecventă, regulă de asociere. De asemenea este prezentat algoritmul Apriori de generare a multimilor frecvente și a regulilor de asociere.

Cuvinte cheie:

mulțime frecventă, regulă de asociere, algoritmul Apriori, suport, mulțime de produse, suportul unei mulțimi de produse, suport minim, nivel de încredere, nivel de încredere minim

3.1 Lecția 1. Reguli de asociere

3.1.1 Definiție

Asocierea este operația de determinare a lucrurilor de pot fi grupate împreună.

De exemplu, determinarea produselor cumparate împreună într-un supermarket. Asocierea este o modalitate de a genera reguli de asociere de forma: "daca X atunci Y (cu probabilitatea p)".

Exemple:

- "Persoanele care cumpara o planta vor cumpara si pamant de flori cu probabilitatea 0.6",
- "Persoanele care cumpara mancare pentru pisici vor cumpara si un vas de mancare pt pisici cu probabilitatea p_1 "
- "Persoanele care cumpara un vas de mancare pt pisici vor cumpara si mancare pentru pisici cu probabilitatea p_2 "

Pentru determinarea regulilor de asociere exista multi algoritmi interesanti. In continuare vom prezenta unul dintre acestia si anume, algoritmul Apriori.

Problema: Se considera urmatoarea lista de tranzactii:

COD TRANZACTIE	CONTINUT (PRODUSE CUMPARATE)
1	LAPTE, PAINE, OUA
2	PAINE, ZAHAR
3	PAINE, CEREALE
4	LAPTE, PAINE, ZAHAR
5	LAPTE, CEREALE
6	PAINE, CEREALE
7	LAPTE, CEREALE
8	LAPTE, PAINE, CEREALE, OUA
9	LAPTE, PAINE, CEREALE

Care sunt perechile de produse care se cumpara frecvent împreună?

Pentru a simplifica tabelul, vom inlocui numele produselor cu litere:

A = LAPTE
B = PAINE
C = CEREALE
D = ZAHAR
E = OUA

Tabelul de mai sus devine:

COD TRANZACTIE	CONTINUT (PRODUSE CUMPARATE)
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

Pentru inmagazinarea fiecărei tranzacții vom introduce pentru fiecare produs posibil o valoare binară care va fi 1 dacă produsul aparține tranzacției și 0 dacă nu. Datele pot fi reprezentate astfel:

COD TRANZACTIE	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

3.1.2 Mulțimi de obiecte frecvente

Definiii:

Produs: unul din A, B, C, D, E

Multime de produse I: submultime de produse posibile; de ex, $I = \{A, B, E\}$

Tranzactie: perechea (COD TRANZACTIE, submultimea tuturor produselor pt care val binară este 1); de ex, (2, {B,D})

Supportul lui I: $\text{sup}(I) = \text{nr. Tranzactii care contin toate produsele din } I$; de ex, $\text{sup}(\{A, B, E\}) = 2$, $\text{sup}(\{B, C\}) = 4$, $\text{sup}(\{B\}) = 7$.

Multime frecventa de produse: o multime de produse I pentru care $\text{sup}(I) \geq \text{minsup}$, unde minsup este o valoare dată numită support minim.

O problema care se pune este determinarea tuturor multimilor frecvente.

Exemplu, pentru exemplul de mai sus,

daca minsup =3, atunci toate multimile frecvente sunt: {A}, {B}, {C}, {A,B}, {A,C}, {B,C},

daca minsup =2, atunci toate multimile frecvente sunt: {A}, {B}, {C}, {D}, {E}, {A,B}, {A,C}, {A,E}, {B,C}, {B,D}, {B,E}, {A,B,C}, {A,B,E}.

La baza alg de determinare a mt frecvente sta urmatoarea

Propozitie: Orice submultime a unei mt frecvente este o mt frecventa.

Demonstratia se poate face prin reducere la absurd.

Definitii:

- R: $I_1 \Rightarrow I_2$ se numeste regula de asociere daca sunt adevarate urmatoarele conditii
 1. I_1 si I_2 sunt multimi frecvente disjuncte si $I_2 \neq \emptyset$ si
 2. Daca $I_1 \in T$, unde T este tranzactie atunci $I_2 \in T$ (cu o anumita probab).

De ex, din {A,B,E} prin partitionare obtinem urmatoarele reguli posibile:

$A \Rightarrow \{B, E\}$, $B \Rightarrow \{A, E\}$, $E \Rightarrow \{A, B\}$, $\{A, B\} \Rightarrow E$, $\{A, E\} \Rightarrow B$, $\{B, E\} \Rightarrow A$,
 $\emptyset \Rightarrow \{A, B, E\}$ (sau $true \Rightarrow \{A, B, E\}$).

- Data o regula de asociere R: $I \Rightarrow J$ se numeste:
 - suportul lui R= $\text{sup}(R) = \text{sup}(I \cup J)$
 - nivelul de incredere al lui R:
$$\text{conf}(R) = \frac{\text{sup}(I \cup J)}{\text{sup}(I)} = \frac{\text{nr. tranzactii ce contin si } I \text{ si } J}{\text{nr. tranzactii ce contin } I}$$

Pentru regulile de mai sus:

R	sup(R)	Conf(R)
$A \Rightarrow \{B, E\}$	2	2/6
$B \Rightarrow \{A, E\}$	2	2/7
$E \Rightarrow \{A, B\}$	2	2/2=1
$\{A, B\} \Rightarrow E$	2	2/4=1/2
$\{A, E\} \Rightarrow B$	2	2/2=1
$\{B, E\} \Rightarrow A$	2	2/2=1
$\emptyset \Rightarrow \{A, B, E\}$	2	2/9

- Date doua valori minsup si minconf numite support minim si nivel de incredere minim, o regula de asociere R se numeste tare daca $\text{sup}(R) \geq \text{minsup}$ si $\text{conf}(R) \geq \text{minconf}$. Se mai spune ca R are parametrii minsup si minconf.

De ex, pentru minsup =2 si minconf =50%=0.5 atunci $E \Rightarrow \{A, B\}$, $\{A, B\} \Rightarrow E$, $\{A, E\} \Rightarrow B$,
 $\{B, E\} \Rightarrow A$ sunt reguli tari. Desigur acestea nu sunt singurele.

3.1.3 Generarea regulilor de asociere. Algoritmul Apriori

Problema: Date valorile minsup si minconf, gasiti toate regulile de asociere cu parametrii minsup si minconf.

Rezolvare: **Algoritmul Apriori**(Agrawal si Srikant)

Pas1. Determinarea tuturor mt frecvente.

1.1. Se determina toate mt frecvente cu 1 element prin examinarea tabelului de apartenenta a unui produs la o tranzactie (cati de 1 exista pe coloana unui produs). Fie $l_1^1, l_2^1, \dots, l_n^1$ aceste mt cu cate un element.

1.2 Se determina toate mt frecvente cu 2 elemente prin interclasarea mt frecvente cu 1 element si verificarea daca mt rezultata este frecventa.

In general, determinarea tuturor mt frecvente cu k elemente, $k > 2$, se bazeaza pe Prop mentionata si deci o mt frecventa cu k elemente este o combinatie obtinuta prin reuniunea a doua mt frecvente cu k-1 elemente.

Pentru eficienta mt frecv vor fi ordonate (lexicografic daca produsele sunt identificate prin litere sau numeric daca produsele sunt cuantificate prin numere).

De ex, {A,B}, {A,C}, {B,C} vor fi mt frecv cu 2 elemente pentru minsup=3 (sau {1,2},{1,3},{2,3}).

Deoarece mt de tipul {1,2} si {3,5} genereaza prin interclasare mt {1,2,3,5}, cu 4 elem, vom face numai combinatii intre mt frecvente de forma: $\{x_1, x_2, \dots, x_{k-2}, a\}$ si $\{x_1, x_2, \dots, x_{k-2}, b\}$. Mt rezultata are k elem: $\{x_1, x_2, \dots, x_{k-2}, a, b\}$.

Verificarea daca mt rezultata este frecventa se poate face prin numarare, folosind tabelul de apartenenta a unui produs la o tranz.

1.3. k=3

while (exista mt frecv cu k-1 elem)

Combina mt frec de tipul $\{x_1, x_2, \dots, x_{k-2}, a\}$ si $\{x_1, x_2, \dots, x_{k-2}, b\}$.

Daca mt rezultata este frecventa atunci ea se adauga la setul mt frecvente de k elem.

k=k+1

endwhile

Pas 2. Genereaza toate regulile de asociere:

Pt fiecare mt frecventa I

Pt fiecare submt J, $J \subseteq I$

Determina regulile de asociere R de forma: $R: I - J \Rightarrow J$ pentru care $\text{conf}(R) \geq \text{minconf}$.



Sa se genereze toate multimile frecvente și toate regulile de asociere tari corespunzătoare mulțimii de tranzacții din exemplul de mai sus pt minsup =3 si minconf =50%.

Mt frecv cu 1 elem: {A},{B},{C},

2 elem: {A,B},{A,C},{B,C}

3 elem: posibil {A,B,C} dar nu este frecventă pt ca sup=2<3

Reguli de asociere:

Regula de asociere	Conf (R)	LIFT(R)
$true \Rightarrow \{A\}$	$6/9 > 1/2$	$6/9 < 1$
$true \Rightarrow \{B\}$	$7/9 > 1/2$	$7/9 < 1$
$true \Rightarrow \{C\}$	$6/9 > 1/2$	$6/9 < 1$
$true \Rightarrow \{A, B\}$	$4/9 < 1/2$	$4/9 < 1$
$A \Rightarrow B$	$4/6 > 1/2$	$(4/6) * (9/7) < 1$
$B \Rightarrow A$	$4/7 > 1/2$	$(4/7) * (9/6) < 1$
$A \Rightarrow C$	$4/6 > 1/2$	$(4/6) * (9/6) = 1$
$C \Rightarrow A$	$4/6 > 1/2$	$(4/6) * (9/6) = 1$
$B \Rightarrow C$	$4/7 > 1/2$	$(4/7) * (9/6) < 1$
$C \Rightarrow B$	$4/6 > 1/2$	$(4/6) * (9/7) < 1$
$true \Rightarrow \{A, C\}$	$4/9 < 1/2$	$(4/9) * (9/4) = 1$
$true \Rightarrow \{B, C\}$	$4/9 < 1/2$	$(4/9) * (9/4) = 1$

Deci reguli tari sunt:

$true \Rightarrow \{A\}$
$true \Rightarrow \{B\}$
$true \Rightarrow \{C\}$
$A \Rightarrow B$
$B \Rightarrow A$
$A \Rightarrow C$
$C \Rightarrow A$
$B \Rightarrow C$
$C \Rightarrow B$

De multe ori baze de date mari pot duce la un nr mare de reguli de asociere, chiar pt minsup si minconf suficient de mari. De aceea sunt necesare reguli de filtrare suplimentare:

Definitie: $LIFT(R) = \frac{conf(R)}{P(J)}$ unde R este regula de asociere $R: I \Rightarrow J$ iar

$P(J) = \frac{sup(J)}{nr. total\ tranzactii}$ se numeste nivelul de incredere asteptat = nivelul de incredere

presupunand ca J este indep de I.

Daca

- $LIFT = 1$ inseamna ca I si J sunt indep
- $LIFT < 1$ inseamna ca I si J sunt correlate negativ
- $LIFT > 1$ inseamna ca I si J sunt correlate pozitiv

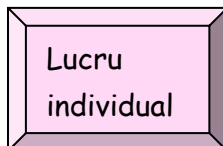
Cu cat LIFT este mai mare cu atat regula de asociere este mai tare. $LIFT > 1$ inseamna ca regula estimeaza mai bine valoarea rezultatului decat daca se ia in calcul numai frecventa de aparitie a rezultatului.

Aplicatii:

1. Determinarea produselor cumparate cel mai des impreuna intr-un supermarket.
2. Determinarea cuvintelor care apar impreuna in mai multe documente poate duce la concluzia ca acele documente sunt despre acelasi subiect.
3. Pagini web care au referinte commune pot fi despre acelasi subiect
4. Prea multe fraze identice in doua documente diferite ne poate indica un plagiat.

Probleme ce pot apare cand se folosesc regulile de asociere in aplicatii:

- Unele reguli obtinute sunt triviale: "Daca X are cablu digital atunci are si cablu basic", "Daca X a cumparat asig pt masina atunci X are masina", "X a cumparat vopsea implica X a cumparat si o pensula de vopsit"
- Unele reguli obtinute sunt inexplicabile: Un studiu facut la inceputul anilor 90 a gasit urmatoarea regula: "Daca este vineri seara si X cumpara scutece atunci X cumpara si bere cu probab 60%". Scopul studiului a fost acela de a descoperi ce produse se cumpara frecvent impreuna pentru ca acestea sa fie asezate in magazin pe rafturi apropiate. Despre acest studiu s-a discutat mult, chiar intr-un articol in Forbes.



Sa se genereze toate multimile frecvente și toate regulile de asociere tari corespunzătoare mulțimii de tranzacții din exemplul de mai sus pt $minsup = 2$ si $minconf = 75\%$.

Test de autoevaluare



- 1) Date tranzactiile din tabelul de mai jos, folositi Algoritmul Apriori pentru a genera toate multimile frecvente si toate regulile de asociere pentru parametrii $minsup = 3$ si $minconf = 0.75$.

Cod tranzactie	Produse cumparate
t1	{A,B,C}
t2	{A}
t3	{B,C}

t4	{B,C,D}
t5	{A,C,D}
t6	{C,D}

Probleme propușe

- 1) Folosiți Algoritmul Apriori pentru a genera toate multimile frecvente și toate regulile de asociere corespunzătoare următoarelor tranzacții, pentru parametrii minsup = 33% și minconf = 60%.

COD TRANZACTIE	CONTINUT (PRODUSE CUMPARATE)
1	{A, B, C, D, E, F}
2	{B, C, D, E, F, G}
3	{A, D, E, H}
4	{A, D, F, I, J}
5	{B, D, E, K}

- 2) Folosiți Algoritmul Apriori pentru a genera toate multimile frecvente și toate regulile de asociere corespunzătoare următoarelor tranzacții, pentru parametrii minsup = 4 și minconf = 0.5.

COD TRANZACTIE	CONTINUT (PRODUSE CUMPARATE)
1	bluza
2	pantofi, fusta, tricou
3	blugi, tricou
4	blugi, pantofi, tricou
5	blugi, pantaloni
6	pantofi, tricou
7	blugi, fusta
8	blugi, pantofi, pantaloni, tricou
9	blugi
10	blugi, pantofi, tricou
11	tricou
12	bluza, blugi, pantofi, fusta, tricou
13	blugi, pantofi, pantaloni, tricou
14	pantofi, fusta, tricou
15	blugi, tricou
16	fusta, tricou
17	bluza, blugi, fusta
18	blugi, pantofi, pantaloni, tricou

19	blugi
20	blugi, pantofi, pantaloni, tricou

UNITATEA DE ÎNVĂȚARE 4

Clustering

Obiective urmărite:

La sfârșitul parcurgerii acestei UI, studenții

- vor înțelege conceptul de clustering (clusterizare)
- vor putea identifica trei tipuri de algoritmi de clustering
- vor cunoaște algoritmul k-means
- vor cunoaște algoritmi de clusterizare aglomerativă și divizivă
- vor ști să construiască dendograme asociate celor două tipuri de clusterizare ierarhică clusterizare: aglomerativă și divizivă.

Ghid de abordare a studiului:

Timpul mediu necesar pentru parcurgerea și asimilarea unității de învățare: 10h.

Lecțiile se vor parcurge în ordinea sugerată de diagramă.

Lecția
1

Lecția
2

Lecția
3

Rezumat:

În această UI este introdus conceptul de clustering și sunt prezentate trei tipuri de algoritmi de clustering: algoritmul k-means și algoritmi de clusterizare aglomerativă și divizivă. Se pun în evidență diferitele metode de calcul al distanței dintre două clustere și modalități de definire a similarității între două înregistrări ale unei baze de date sau între două clustere.

Cuvinte cheie:

Clusterizare, algoritmi de clusterizare aglomerativă și divizivă, algoritmul k-means,

4.1 Lecția 1. Algoritmul k-means

4.1.1 Definiție

Clusterizarea este procesul de divizare a unei baze de date în grupe de înregistrări similare astfel încât membrii aceleiași grupe să fie cât se poate de apropiați unul de altul, iar grupurile sunt cât se poate de departate unele de celelalte.

În clusterizare nu există nici o mît de date preclasificate și nu se face nici o distincție între variabilele independente și cele dependente. Var (atributele) în funcție de care se face operația de clusterizare se numesc var de intrare, iar dimensiunea problemei este dată de nr var de intrare.

Pp ca avem n variabile de intrare: X_1, X_2, \dots, X_n . Atunci fiecare înregistrare conține câte o valoare pentru cele n var ($X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$) reprezintă un punct (x_1, x_2, \dots, x_n) în spațiu n -dimensional.

Cel mai des folosit alg de clusterizare este alg K-means (al clusterizării cu k clustere folosind media aritmetică).

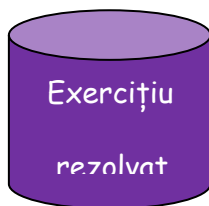
4.1.2 Alg K-means (MacQueen 1967)

Fie k fixat, $k = \text{nr}$ cluster.

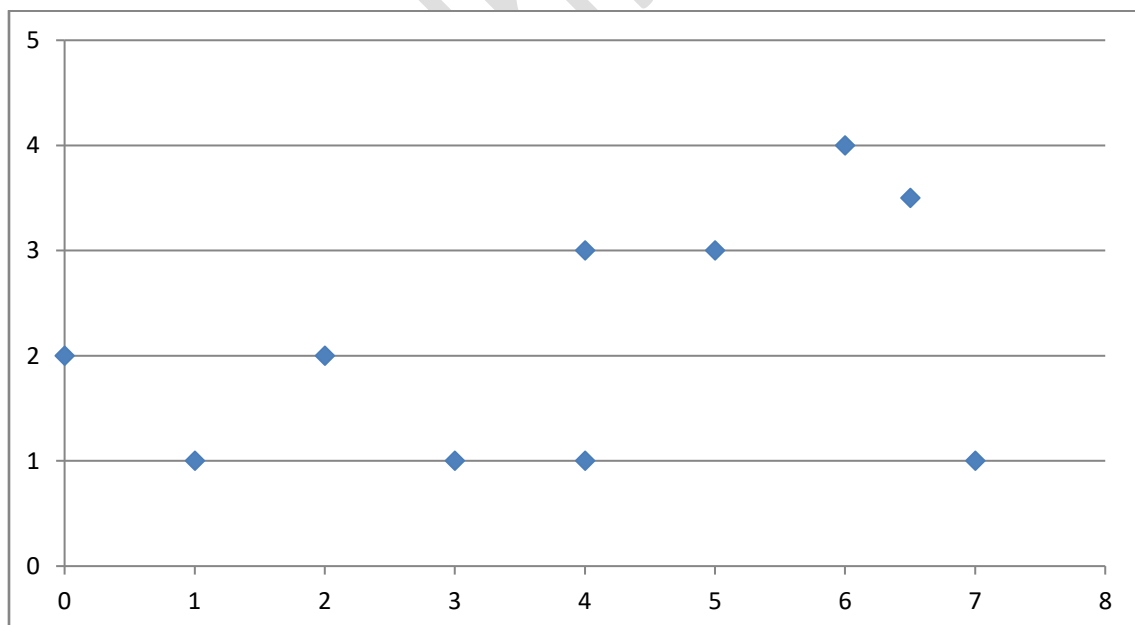
1. Se aleg la intamplare k puncte (inregistrari) ca fiind centrele initiale ale celor k cluster. (MacQueen propune alegerea primelor k inreg)
2. Pt fiecare inreg determina cel mai apropiat centru si atribuie inregistrarii clusterul asociat centrului.
3. Pt fiecare cluster, calculeaza media inreg din cluster. Muta central clusterului in pct coresp mediei.
4. Repeta pasii 2 si 3 pana cand se obtine convergenta adica pana cand nr de reatribuiri ale clusterelor este mai mic decat o valoare ϵ data.

Obs. Alg functioneaza numai pt atribuite cu valori numerice. O posibila fct distant care sa descrie notiunea de cel mai apropiat este fct distanta euclidiană între două pct: $X = (x_1, x_2, \dots, x_n)$ si

$$Y = (y_1, y_2, \dots, y_n): d(X, Y) = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}.$$



Sa se aplice algoritmul Apriori următoarelor punctelor din graficul de mai jos pentru $k=3$.



Alegem la intamplare centrele clusterelor punctele $A(1,1)$, $B(3,1)$, $C(7,1)$. Grupam punctele in cluster astfel:

cluster 1: punctele $(1,1)$, $(0,2)$, $(2,2)$

cluster 2: punctele $(3,1)$, $(4,1)$, $(4,3)$, $(5,3)$

cluster 3: punctele (6,4), (6.5,3.5), (7,1).

Calculam mediile clusterelor si recentram.

cluster 1: $x=(1+0+2)/3=1$, $y=(1+2+2)/3=5/3$

cluster 2: $x=(3+4+4+5)/4=4$, $y=(1+1+3+3)/4=2$

cluster 3: $x=(7+6+6.5)/3=6.5$, $y=(1+4+3.5)/3$

Se recalc distantele de la fiecare punct la noile centre. Cum dist minime nu duc la reatribuiri ale clusterelor, se obtine convergenta.

De asemenea, alegand cele 3 centre initiale altfel, clusterelor rezultate vor fi altele. De exemplu,

Cluster 1: acelasi centru

Cluster 2: acelasi centru

Cluster 3: $x=(2+3+4+7+4+5+6+6.5)/8=4.68$, $y=(2+1+1+3+3+4+3.5+1)/8=2.31$

$$d(1,2) = \sqrt{(3-4.68)^2 + (2.31-1)^2} < 2$$

Deci singura realocare este a nodului (2,2) la clusterul 2.

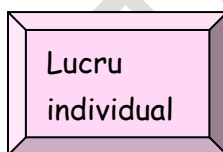
Cluster 1: acelasi centru

Cluster 2: $x=1.5$, $y=1.5$

Cluster 3: $x=(3+4+7+4+5+6+6.5)/7=5.7$, $y=(1+1+3+3+4+3.5+1)/7=2.35$

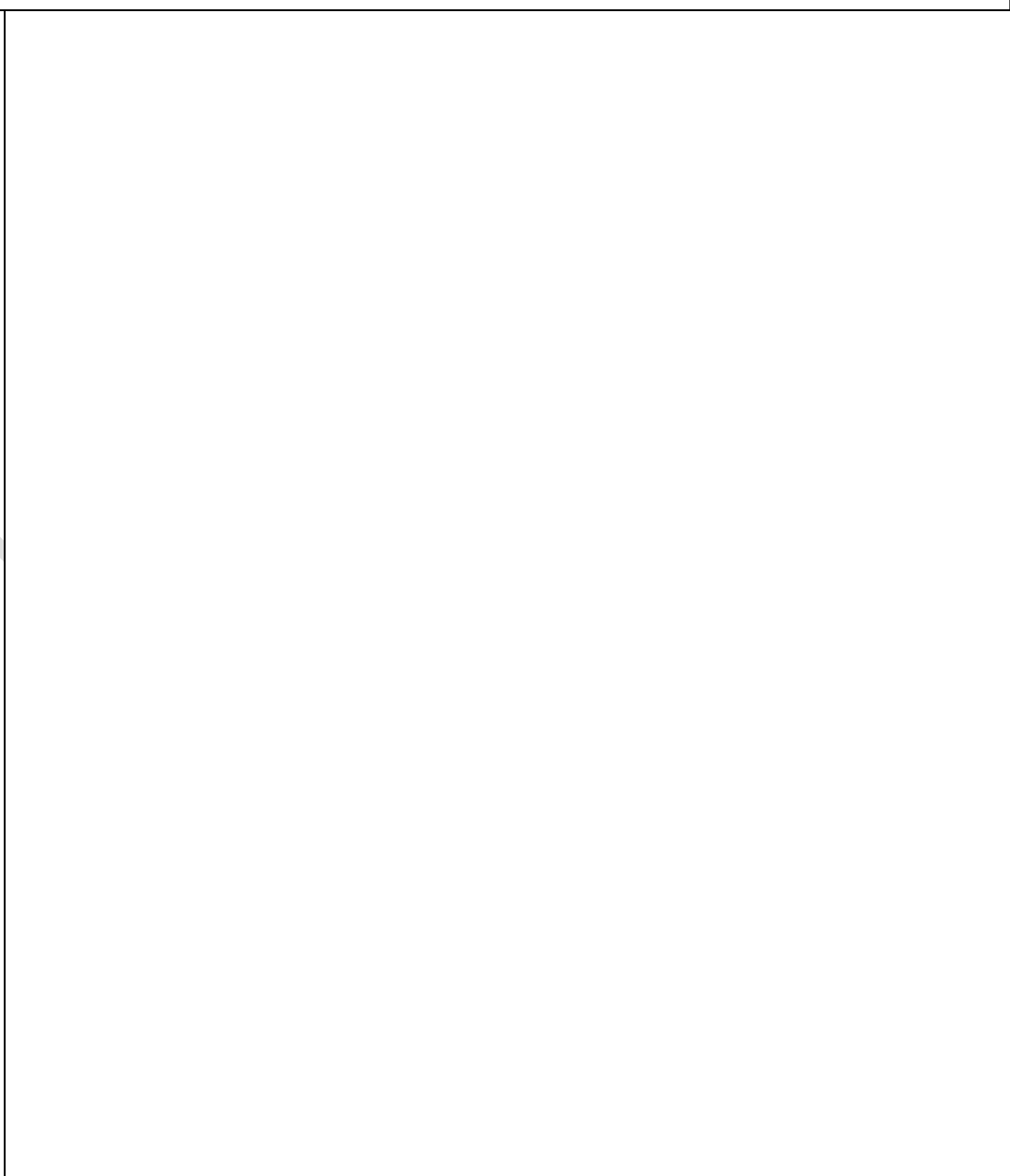
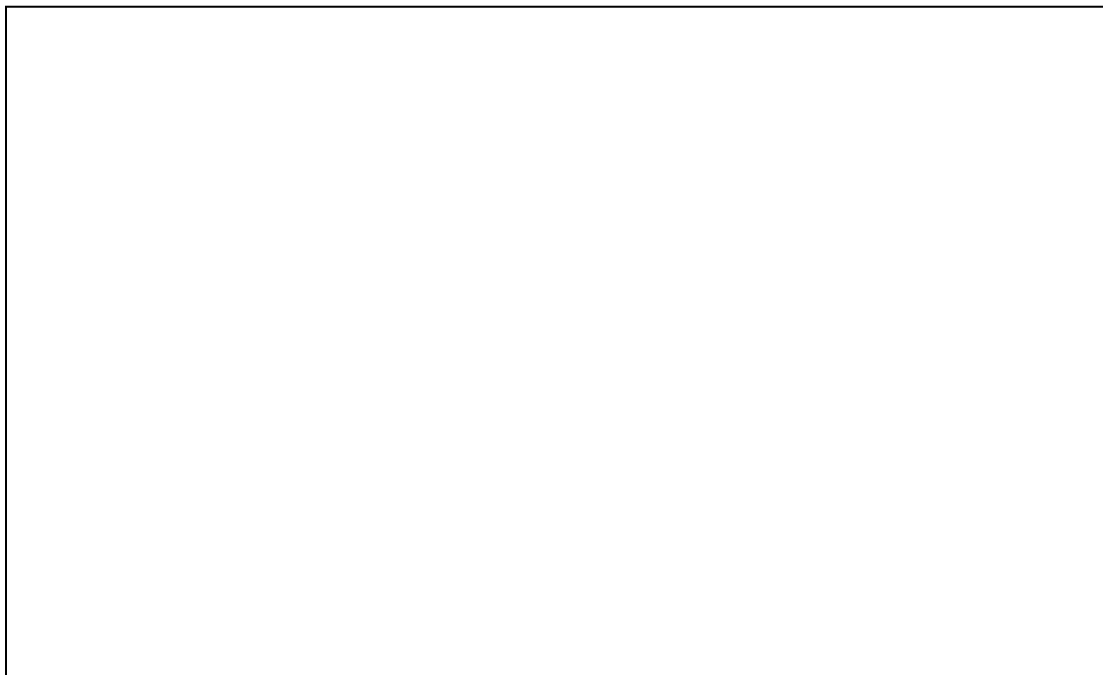
Se recentreaza, se atribuie lui 1 clusterul 2 si procedeul continua.

Alegerea lui k: In cele mai multe cazuri nu exista un motiv a priori pt selectarea lui k. De obicei se alege o valoare a lui k, se aplica alg, se evalueaza rezultatele, apoi se incerca o alta valoare si se analizeaza.



Se da următoarea mulțime de puncte { A(0,0), B(1,2), C(0, 2), D(3,2), E(3,1), F(3,0), G(4,0)}. Sa se determine o clusterizare cu k=3 clusterelor folosind algoritmul de clusterizare k-means. Se aleg centrele initiale punctele A, B, C.

- 1) Se dă următoarea mulțime de puncte $\{ A(0,0), B(1,2), C(0, 2), D(3,2), E(3,1), F(3,0), G(4,0) \}$. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare k-means. Se aleg centrele initiale punctele A, E, G.



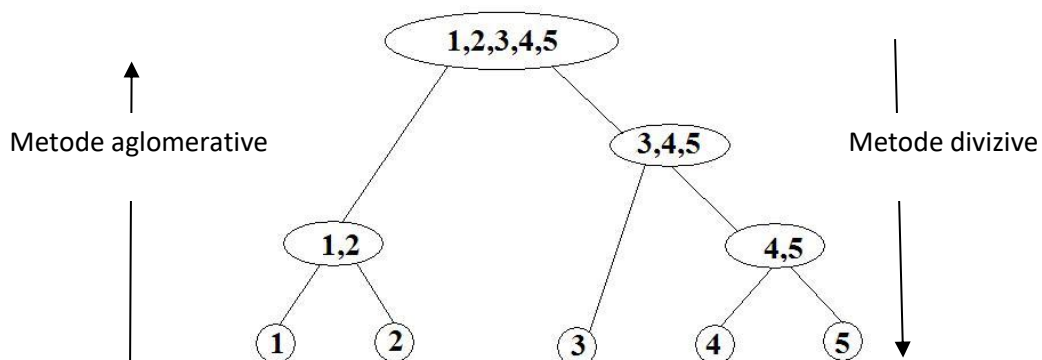
4.2 Lecția 2. Clusterizarea ierarhică. Metode aglomerative

Un alt tip de clusterizare este *clusterizarea ierarhică*. Aceasta se poate face folosind :

- Metode aglomerative de clusterizare – presupun o serie de fuziuni a înregistrărilor initiale din n cluster (cate înreg sunt) în grupuri din ce în ce mai putine de înreg, pana cand se obtine nr de cluster dorit.
- Metode divizive de clusterizare – pp ca initial toate înreg fac parte dintr-un singru cluster pe care apoi îl vor împarti în grupuri.

Metodele de clusterizare ierarhice se pot reprezenta printr-o diagrama 2D numita dendograma care prezinta fuziunile sau divizarile facute.

Exemplu de dendograma:



Metode aglomerative: Produc o serie de partitii ale inreg: P_n, P_{n-1}, \dots, P_1 unde $n = nr$ total de inreg

$$P_n = \{\{i_1\}, \{i_2\}, \dots, \{i_n\}\} \text{ n mt (n clustere)}$$

.

.

.

$$P_1 = \{i_1, i_2, \dots, i_n\} \text{ 1 cluster}$$

La fiecare pas metoda alege cate doua clustere – mai exact doua clustere care sunt cele mai apropiate:

$$\text{Alegem } R, S \text{ a.i. } D(R, S) = \min_{r, s \text{ clustere}} D(r, s). \text{ Unim } R \text{ si } S.$$

Diferenta dintre metode apare datorita modurilor diferite de a masura dist intre clustere.

4.2.1 Măsurarea distanței dintre clustere

1. Clusterizare cu leg simpla

Dist intre clustere= cel mai scurt drum intre clustere

$$D(r, s) = \min\{d(i, j) / i \text{ este in clusterul } r \text{ si } j \text{ in clusterul } s\}$$

2. Clusterizare cu leg completa

Dist intre clustere= distanta dintre cele mai departate inreg din clustere

$$D(r, s) = \max\{d(i, j) / i \text{ este in clusterul } r \text{ si } j \text{ in clusterul } s\}$$

3. Clusterizare cu leg medie

$$D(r, s) = \text{media}\{d(i, j) / i \text{ in clusterul } r, j \text{ sunt in clusterul } s\} = \frac{\sum d(i, j)}{\text{card}(r) \cdot \text{card}(s)}$$

4. Clusterizare cu leg medie de grup – r si s se unesc a.i. dupa unire dist medie din interiorul fiecarui cluster sa fie minima. Pp ca noul cluster format prin unirea lui r cu s este t. Atunci:

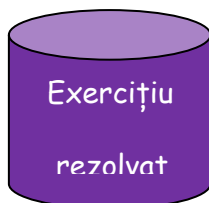
$$D(r, s) = \text{media}\{d(i, j) / i, j \text{ sunt in clusterul } t \text{ format prin fuzionarea lui } r \text{ cu } s\}$$

5. Clusterizare cu legatura Ward: Dist = cresterea in suma patratelor erorii ESS dupa fuzionarea celor doua clustere intr-unul singur. Se aleg pasi succesivi care sa minimizeze cresterea in ESS la fiecare pas.

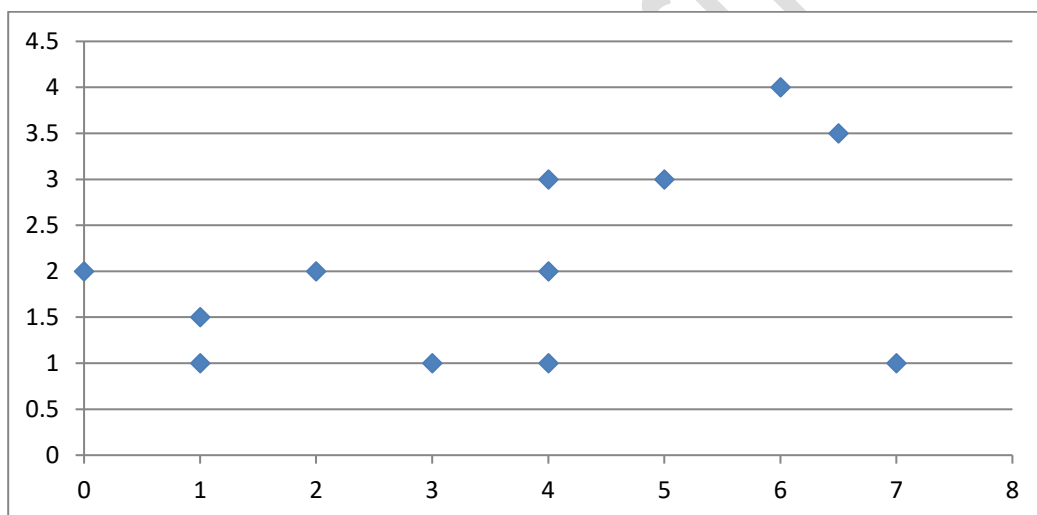
$X=mt$, $n = \text{nr elem ale lui } X$

$$ESS(X) = \sum_{i=1}^n \|x_i - media\|^2$$

$D(r,s)=ESS(t=\text{cluster obt prin fuzionarea lui } r \text{ si } s) - (ESS(r) + ESS(s))$



Sa se aplice metoda de clusterizare aglomerativă cu legătură simplă următoarelor punctelor din graficul de mai jos pentru $k=\text{nr clustere} = 3$.



Initial:

Cluster 1	punctul (0,2)	Cluster 7	punctul (4,2)
Cluster 2	punctul (1,1)	Cluster 8	punctul (4,3)
Cluster 3	punctul (1, 1.5)	Cluster 9	punctul (5,3)
Cluster 4	punctul (2,2)	Cluster 10	punctul (6,4)
Cluster 5	punctul (3,1)	Cluster 11	punctul (6.5, 3.5)

Cluster 6	punctul (4,1)	Cluster 12	punctul (7,1)
-----------	---------------	------------	---------------

Pas 1: Din reprezentarea grafica observam ca cele mai apropiate cluster sunt 2 si 3. Deci $R =$ cluster 2, $S =$ cluster 3, unim cele doua cluster dand noului cluster numarul cel mai mic intre 2 si 3 (doar pt a face o alegere) si avem la pasul urmator:

Cluster 1	punctul (0,2)	Cluster 7	punctul (4,2)
Cluster 2	(1,1), (1, 1.5)	Cluster 8	punctul (4,3)
Cluster 4	punctul (2,2)	Cluster 9	punctul (5,3)
Cluster 5	punctul (3,1)	Cluster 10	punctul (6,4)
Cluster 6	punctul (4,1)	Cluster 11	punctul (6.5, 3.5)
		Cluster 12	punctul (7,1)

Pas 2: Tot din reprezentarea grafica observam ca acum cele mai apropiate cluster sunt 10 si 11. Unim cele doua cluster si avem urmatoarele cluster:

Cluster 1	punctul (0,2)	Cluster 7	punctul (4,2)
Cluster 2	(1,1), (1, 1.5)	Cluster 8	punctul (4,3)
Cluster 4	punctul (2,2)	Cluster 9	punctul (5,3)
Cluster 5	punctul (3,1)	Cluster 10	[(6,4), (6.5, 3.5)]
Cluster 6	punctul (4,1)	Cluster 12	punctul (7,1)

Pas 3: Tot din reprezentarea grafica observam ca acum cele mai apropiate cluster sunt 6 si 7 sau 7 si 8. Pentru a face o alegere, consideram 6 si 7 cele doua cluster pe care le unim si avem urmatoarele cluster:

Cluster 1	punctul (0,2)	Cluster 8	punctul (4,3)
Cluster 2	(1,1), (1, 1.5)	Cluster 9	punctul (5,3)
Cluster 4	punctul (2,2)	Cluster 10	(6,4), (6.5, 3.5)
Cluster 5	punctul (3,1)	Cluster 12	punctul (7,1)

Cluster 6	(4,1), (4,2)		
-----------	--------------	--	--

Pas 4: Tot din reprezentarea grafica observam ca acum cele mai apropiate clustere sunt 6 si 8. Le unim si avem urmatoarele clustere:

Cluster 1	punctul (0,2)	Cluster 9	punctul (5,3)
Cluster 2	(1,1), (1, 1.5)	Cluster 10	(6,4), (6.5, 3.5)
Cluster 4	punctul (2,2)	Cluster 12	punctul (7,1)
Cluster 5	punctul (3,1)		
Cluster 6	(4,1), (4,2), (4,3)		

Pas 5: Din reprezentarea grafica observam ca acum cele mai apropiate posibile clustere ar fi:

Cluster 1, Cluster 2	Dist = $\min \{ d((0,2), (1,1)), d((0,2), (1,1.5)) \} = d((0,2), (1,1.5)) = \sqrt{1^2 + 0.5^2} = \sqrt{5/4} > 1$
Cluster 2, Cluster 4	Dist = $\min \{ d((2,2), (1,1)), d((2,2), (1,1.5)) \} = d((2,2), (1,1.5)) = \sqrt{1^2 + 0.5^2} = \sqrt{5/4} > 1$
Cluster 4, Cluster 5	Dist = $\sqrt{2} > 1$
Cluster 5, Cluster 6	Dist = $\min \{ d((3,1), (4,1)), d((3,1), (4,2)), d((3,1), (4,3)) \} = d((3,1), (4,1)) = 1$
Cluster 6, Cluster 9	Dist = $\min \{ d((5,3), (4,1)), d((5,3), (4,2)), d((5,3), (4,3)) \} = d((5,3), (4,3)) = 1$

Cluster 9, Cluster 10	Dist = $\min \{ d((5,3), (6,4)), d((5,3), (6.5,3.5)) \} = d((0,2), (1,1.5)) = d((5,3), (6,4)) = \sqrt{2} > 1$
Cluster 10, Cluster 12	Dist = $\min \{ d((7,1), (6,4)), d((7,1), (6.5,3.5)) \} = d((7,1), (6.5,3.5)) = \sqrt{0.5^2 + 2.5^2} = \sqrt{13/2} > 1$

Alegem Cluster 5 si 6. Le unim si avem urmatoarele cluster:

Cluster 1	punctul (0,2)	Cluster 9	punctul (5,3)
Cluster 2	(1,1), (1, 1.5)	Cluster 10	(6,4), (6.5, 3.5)
Cluster 4	punctul (2,2)	Cluster 12	punctul (7,1)
Cluster 5	(3,1), (4,1), (4,2), (4,3)		

Pas 6: Alegem apoi cluster 5 si 9 (distanța minimă fiind 1). Le unim si avem urmatoarele cluster:

Cluster 1	punctul (0,2)	Cluster 10	(6,4), (6.5, 3.5)
Cluster 2	(1,1), (1, 1.5)	Cluster 12	punctul (7,1)
Cluster 4	punctul (2,2)		
Cluster 5	(3,1), (4,1), (4,2), (4,3), (5,3)		

Pas 7: Din reprezentarea grafică observăm că acum cele mai apropiate posibile cluster ar fi:

Cluster 1, Cluster 2	Dist = $\min \{ d((0,2), (1,1)), d((0,2), (1,1.5)) \} = d((0,2), (1,1.5)) = \sqrt{1^2 + 0.5^2} = \sqrt{5/4} > 1$
Cluster 2, Cluster 4	Dist = $\min \{ d((2,2), (1,1)), d((2,2), (1,1.5)) \} = d((2,2), (1,1.5)) = \sqrt{1^2 + 0.5^2} = \sqrt{5/4}$
Cluster 4, Cluster 5	Dist = $\sqrt{2} > \sqrt{5/4}$
Cluster 5, Cluster 10	Dist = $\min \{ d((5,3), (6.5,3.5)) = d((5,3), (6,4)) = \min \{ \sqrt{1.5^2 + 0.5^2}, \sqrt{2} \} = \sqrt{2} > \sqrt{5/4}$

Cluster 10, Cluster 12	$\text{Dist} = \min \{ d((7,1), (6,4)), d((7,1), (6.5,3.5)) \} = d((7,1), (6.5,3.5)) = \sqrt{0.5^2 + 2.5^2} = \sqrt{13/2} > \sqrt{5/4}$
------------------------	---

Alegem Cluster 1 si 2. Le unim si avem urmatoarele cluster:

Cluster 1	(0,2),(1,1), (1, 1.5)	Cluster 10	(6,4), (6.5, 3.5)
Cluster 4	punctul (2,2)	Cluster 12	punctul (7,1)
Cluster 5	(3,1), (4,1), (4,2), (4,3), (5,3)		

Pas 8: Alegem Cluster 1 si 4 (distanța = $\sqrt{5/4}$). Le unim si avem urmatoarele cluster:

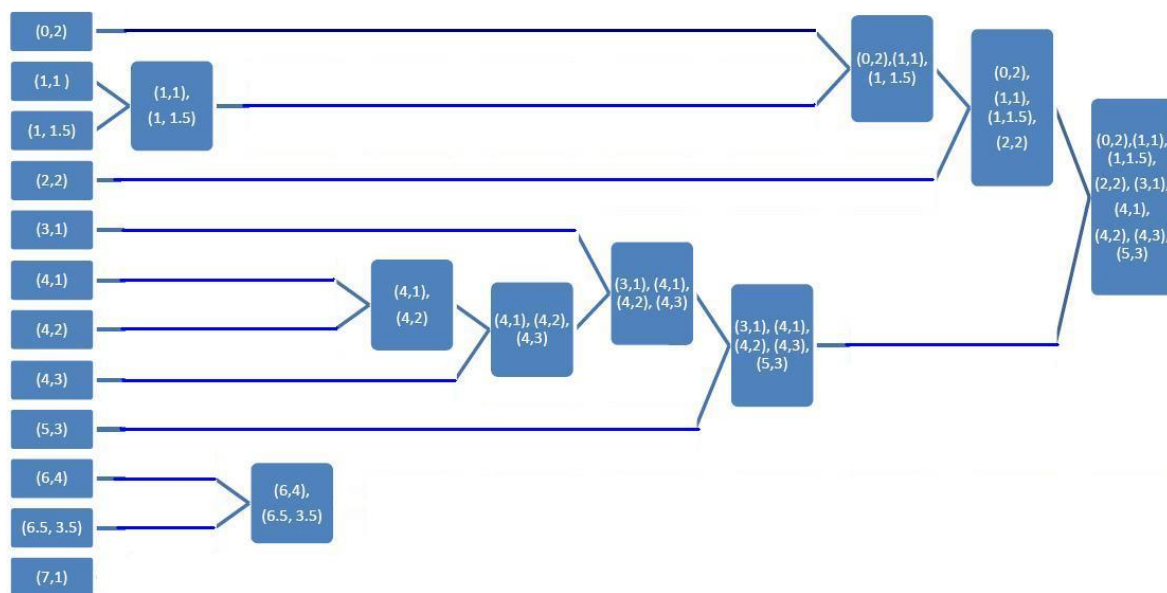
Cluster 1	(0,2),(1,1), (1, 1.5),(2,2)	Cluster 10	(6,4), (6.5, 3.5)
Cluster 5	(3,1), (4,1), (4,2), (4,3), (5,3)	Cluster 12	punctul (7,1)

Pas 9: Alegem Cluster 1 si 5 (distanța = $\sqrt{2}$). Le unim si avem urmatoarele cluster:

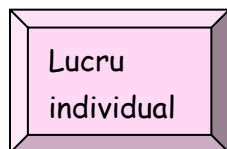
Cluster 1	(0,2),(1,1), (1, 1.5),(2,2), (3,1), (4,1), (4,2), (4,3), (5,3)
Cluster 10	(6,4), (6.5, 3.5)
Cluster 12	punctul (7,1)

Au fost gasite 3 cluster. Ne oprim.

Dendograma asociata este urmatoarea:



În același mod se pot obține clustere folosind și celelalte metode de calcul al distanțelor între clustere.



Se dă următoarea mulțime de puncte $\{A(0,0), B(1,2), C(0,2), D(3,2), E(3,1), F(3,0), G(4,0)\}$. Să se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativă cu legătură simplă.

Test de autoevaluare

6

- 2) Se dă următoarea mulțime de puncte $\{A(0,0), B(1,2), C(0,2), D(3,2), E(3,1), F(3,0), G(4,0)\}$. Să se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativă cu legătură completă.

4.3 Lecția 3. Clusterizarea ierarhică. Metode divizive

Algoritm:

Initial toate inreg sunt intr-un singur cluster.

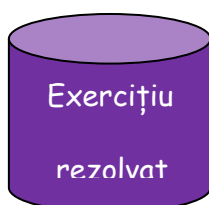
Se decide un prag pt dist.

Se calc dist dintre orice 2 inreg si se det perechea cu cea mai mare dist.

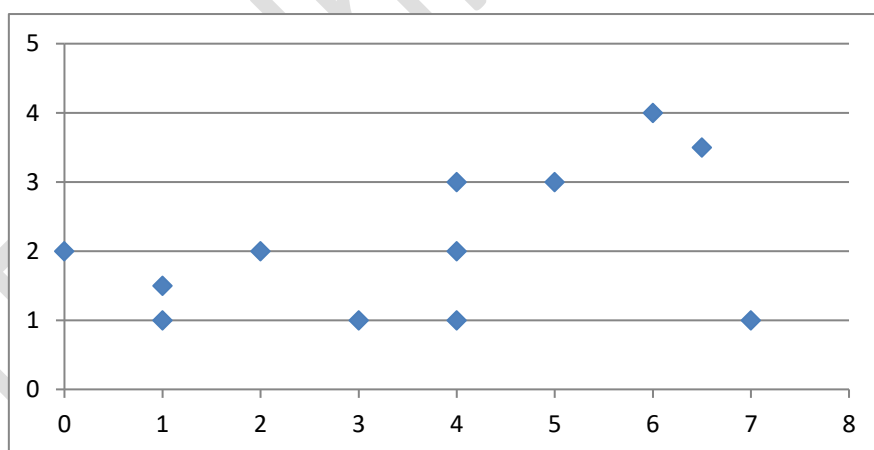
Dist max se compara cu pragul pt dist.

Daca dist max > prag pt dist atunci grupul se imparte in doua. Cele 2 inreg se pun in clusterse diferite iar celelalte pcte se pun in clusterul cel mai apropiat. Si se repeta procedeul.

Daca dist max < prag pt dist atunci stop.



Sa se aplice metoda de clusterizare divizivă următoarelor punctelor din graficul de mai jos pentru prag dist=3 .



Pas 1: Intai avem un singur cluster format din cele 12 puncte.

Cluster 1	(0,2),(1,1), (1, 1.5),(2,2), (3,1), (4,1), (4,2), (4,3), (5,3), (6,4), (6.5, 3.5), (7,1)
-----------	--

Fixam limita (pragul) maxima a distantei intre orice doua puncte: prag dist=3

Pas 2: Evident cele doua perechi aflate la cele mai departate sunt (0,2) si (7, 1).

Impartim celelalte puncte in fct de dist lor la cele doua puncte:

Cluster 1	(0,2), (1,1), (1, 1.5), (2,2), (3,1),
Cluster 2	(5,3), (6,4), (6.5, 3.5), (7,1), (4,3), (4,2), (4,1)

$$d((0,2),(4,3)) = \sqrt{4^2 + 1} = \sqrt{17}$$

$$d((7,1),(4,3)) = \sqrt{3^2 + 2^2} = \sqrt{13}$$

$$d((0,2),(4,2)) = \sqrt{4^2 + 1} = 4$$

$$d((7,1),(4,2)) = \sqrt{3^2 + 1^2} = \sqrt{10}$$

$$d((0,2),(4,1)) = \sqrt{4^2 + 1} = \sqrt{17}$$

$$d((7,1),(4,1)) = 3$$

Pas 3: În cadrul fiecarui cluster determinam cele mai departate puncte. Pentru clusterul 1 acestea

sunt: (0,2) si (3,1) cu o distanta = $\sqrt{3^2 + 1^2} = \sqrt{10}$

Pentru clusterul 2: acestea pot fi (7,1) si (4,3) cu o distanta $d((7,1),(4,3)) = \sqrt{3^2 + 2^2} = \sqrt{13}$ sau

$$d((6,4),(4,1)) = \sqrt{2^2 + 3^2} = \sqrt{13}$$

Alegem cele mai departate puncte deci (7,1) si (4,3) cu distanta > prag dist = 3 si deci clusterul 3 se va diviza. Toate punctele din acest cluster vor fi distribuite in fct de distanta lor fata de cele doua puncte (7,1) si (4,3).

Cluster 1	(0,2), (1,1), (1, 1.5), (2,2), (3,1),
Cluster 2	(5,3), (6,4), (4,3), (4,2), (4,1)
Cluster 3	(7,1), (6.5, 3.5),

$$d((6,4), (7,1)) = \sqrt{3^2 + 1^2} = \sqrt{10}$$

$$d((6,4), (4,3)) = \sqrt{2^2 + 1^2} = \sqrt{5}$$

Pas 4: În cadrul fiecarui cluster determinam cele mai departate puncte.

Pentru clusterul 1 acestea sunt: (0,2) si (3,1) cu o distanta = $\sqrt{3^2 + 1^2} = \sqrt{10}$

Pentru clusterul 2: acestea sunt (6,4) si (4,1) cu o distanta $d((6,4),(4,1)) = \sqrt{2^2 + 3^2} = \sqrt{13}$

Pentru clusterul 3: acestea sunt (7,1), (6.5, 3.5) cu o distanta = $d((7,1),(6.5,3.5)) = \sqrt{0.5^2 + 2.5^2} = \sqrt{13/2}$

Alegem cele mai departate puncte deci (6,4) si (4,1) cu distanta > prag dist = 3 si deci clusterul 2 se va diviza. Toate punctele din acest cluster vor fi distribuite in fct de distanta lor fata de cele doua puncte (6,4) si (4,1).

Cluster 1	(0,2),(1,1), (1, 1.5),(2,2), (3,1),
Cluster 2	(4,3), (4,2), (4,1)
Cluster 3	(6,4), (5,3)
Cluster 4	(7,1), (6.5, 3.5),

Pas 5: În cadrul fiecarui cluster determinam cele mai departate puncte.

Pentru clusterul 1 acestea sunt: (0,2) si (3,1) cu o distanta = $\sqrt{3^2 + 1^2} = \sqrt{10}$

Pentru clusterul 2: acestea sunt (4,3) si (4,1) cu o distanta = 2

Pentru clusterul 3: acestea sunt (6,4), (5,3) cu o distanta = $\sqrt{2}$

Pentru clusterul 4: acestea sunt (7,1), (6.5, 3.5) cu o distanta = $d((7,1),(6.5,3.5)) = \sqrt{0.5^2 + 2.5^2} = \sqrt{13/2}$

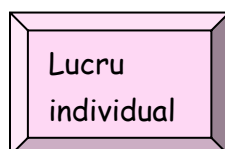
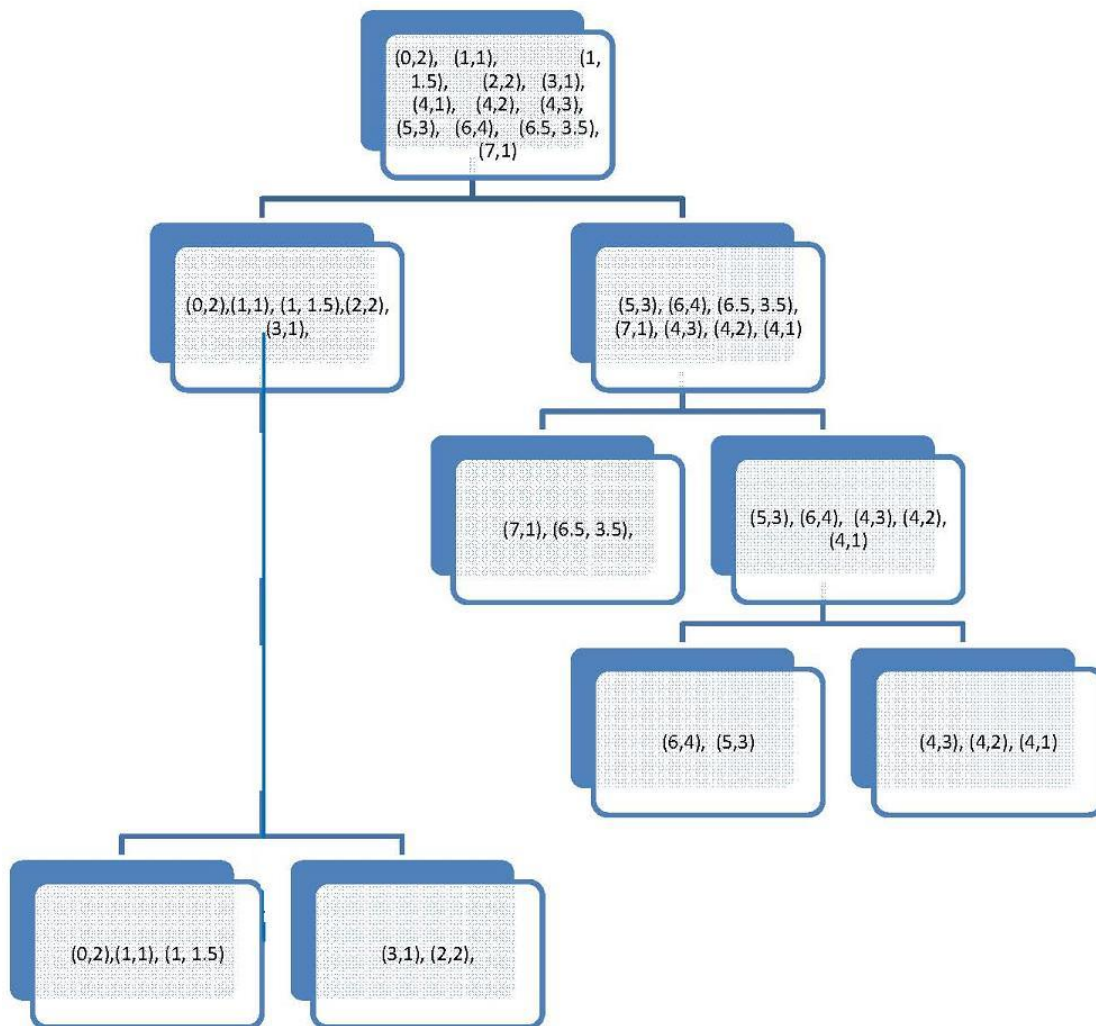
Alegem cele mai departate puncte deci (0,2) si (3,1) cu distanta > prag dist = 3 si deci clusterul 1 se va diviza. Toate punctele din acest cluster vor fi distribuite in fct de distanta lor fata de cele doua puncte (0,2) si (3,1).

Cluster 1	(0,2),(1,1), (1, 1.5)
Cluster 2	(3,1), (2,2),
Cluster 3	(4,3), (4,2), (4,1)
Cluster 4	(6,4), (5,3)
Cluster 5	(7,1), (6.5, 3.5),

Pas 6: În cadrul fiecarui cluster determinăm cele mai departate puncte.

Alegem cele mai departate puncte deci $(7,1)$, $(6.5, 3.5)$ cu o distanță $= d((7,1),(6.5,3.5)) = \sqrt{0.5^2 + 2.5^2} = \sqrt{13/2}$ cu distanță $< \text{prag dist}=3$ și deci procesul divizării s-a terminat.

Dendograma asociată este următoarea:



Se dă următoarea mulțime de puncte $\{ A(0,0), B(1,2), C(0, 2), D(3,2), E(3,1), F(3,0), G(4,0) \}$. Să se determine o clusterizare cu prag $\text{dist}=3$ clustere folosind algoritmul de clusterizare divizivă.

Test de autoevaluare



- 3) Se dă următoarea mulțime de puncte $\{ A(0,0), B(1,2), C(0, 2), D(3,2), E(3,1), F(3,0), G(4,0) \}$. Sa se determine o clusterizare cu prag $\text{dist}=2$ folosind algoritmul de clusterizare divizivă.

Probleme propușe

1. Se dau punctele $A(0,0)$, $B(1,2)$, $C(0,2)$, $D(3,2)$, $E(3,1)$, $F(3,0)$, $G(5,0)$,
 - a. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare k-means. Se aleg centrele initiale punctele A, B, C.
 - b. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativa cu legatura simpla.
 - c. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativa cu legatura completa.
 - d. Sa se determine o clusterizare folosind algoritmul de clusterizare diviziva si prag pentru distanta = 2 (pana cand distanta maxima in clustere ≤ 2)
2. Se dau punctele $A_1=(2,10)$, $A_2=(2,5)$, $A_3=(8,4)$, $A_4=(5,8)$, $A_5=(7,5)$, $A_6=(6,4)$, $A_7=(1,2)$, $A_8=(4,9)$.
 - a. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare k-means. Se aleg centrele initiale punctele A_1 , A_4 , A_7 .
 - b. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativa cu legatura simpla.
 - c. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativa cu legatura completa.
 - d. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativa cu legatura medie.
 - e. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativa cu legatura medie de grup.
 - f. Sa se determine o clusterizare cu $k=3$ clustere folosind algoritmul de clusterizare aglomerativa cu legatura Ward.
 - g. Sa se determine o clusterizare folosind algoritmul de clusterizare diviziva si prag pentru distanta = 2 (pana cand distanta maxima in clustere ≤ 2)

UNITATEA DE ÎNVĂȚARE 5

Tanagra

Obiective urmărite:

La sfârșitul parcurgerii acestei UI, studenții

- vor ști cum să importe baze de date din fișiere text sau din Excel în Tanagra
- vor ști să ruleze diferiți algoritmi de data mining folosind pachetul software Tanagra
- vor ști să alcătuiască diferite proiecte complexe care să aplice mai multe tehnici de data mining în Tanagra
- vor ști să rezolve probleme din viața reală folosind tehnici de data mining în Tanagra

Ghid de abordare a studiului:

Timpul mediu necesar pentru parcurgerea și asimilarea unității de învățare: 5h.

Lecțiile se vor parcurge în ordinea sugerată de diagramă.

Lecția

1

Rezumat:

În această UI este prezentat pachetul software pentru data mining Tanagra. Sunt prezentate modul de instalare, interfața, modul de importare a datelor în Tanagra, principalii operatori.

Cuvinte cheie:

Tanagra, define status, import date, fisier Excel, diagrama, componenta, operatori

5.1 Lectia 1. Tanagra

TANAGRA este un open source software pentru data mining ce poate fi folosit gratuit pentru invatare si cercetare. A fost creat de Ricco Rakotomalala, profesor la Universitatea din Lyon, Franta.

5.1.1 Instalare

Pentru instalarea acestui software se selecteaza din pagina de web a soft-ului :

<http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>

download si apoi Setup si run aplicatia setup_tanagra.exe.

5.1.2 Interfața

Interfata soft-ului cuprinde 3 parti:

1. Diagrama (cu o structura de arbore) care reprezinta pasii analizei asupra bazei de date.
2. Operatorii sau componentele (identificate in diagrama prin noduri), ce reprezinta o operatiile efectuate in baza de date. Multimea operatorilor posibili se gaseste in fereastra de jos a interfetei si este impartita in categorii numite tab-uri: Data visualization, Statistics, Clustering, etc.
3. Fisierul raport impartit in 2: descrierea parametrilor si rezultatele.

5.1.3 Tutoriale

Ca tutoriale exista cateva exemple detaliate despre cateva metode de analiza ce pot fi efectuate folosind TANAGRA. Ele pot fi descarcate tot din pagina de web a software-ului. Va recomand sa cititi urmatoarele tutoriale:

<http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enImportDataset.pdf>

<http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enBasics.pdf>

http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Deployment.pdf

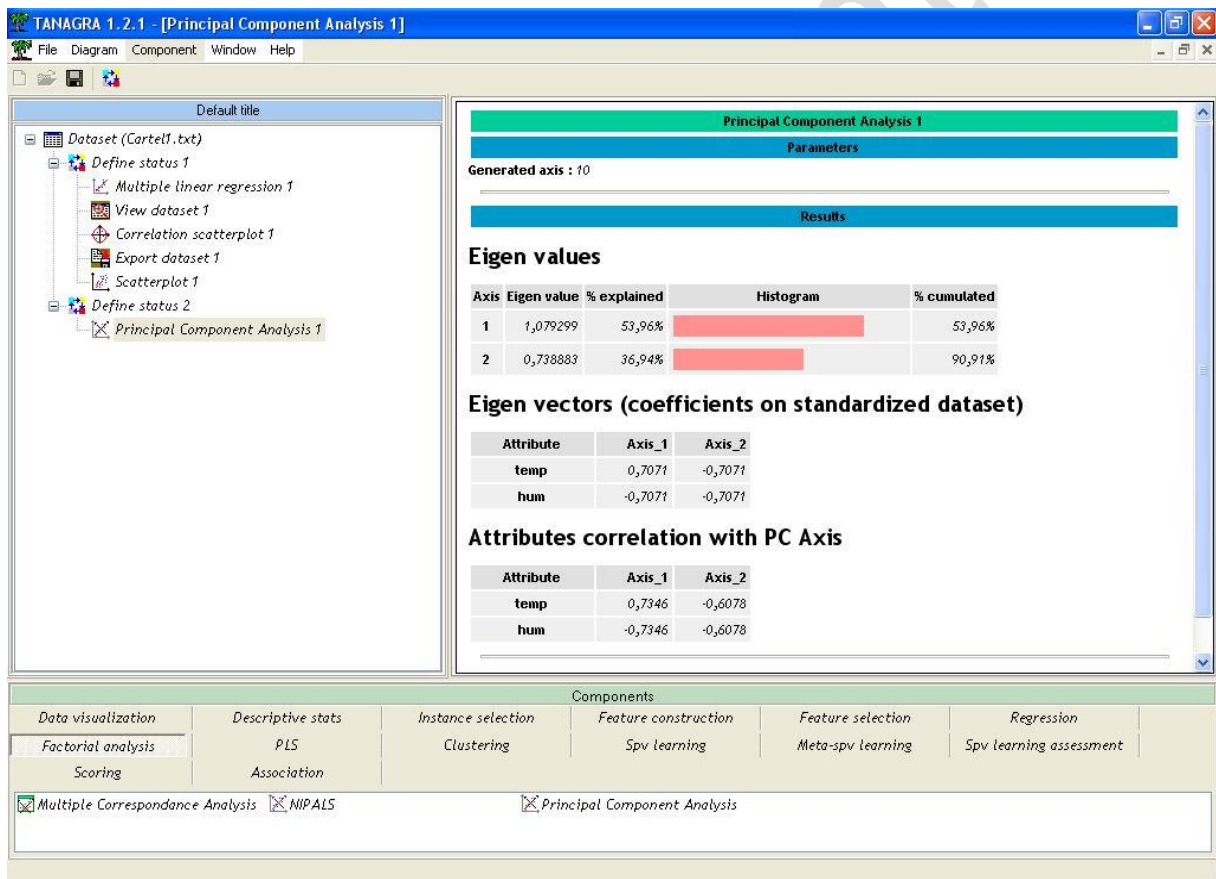
<http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enSupervisedDiscretization.pdf>

http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enHAC_IRIS.pdf

http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/en_Tanagra_Handle_Spreadsheet_File.pdf

<http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enDecisionTree.pdf>

http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/enFeature_Selection_For_Naive_Bayes.pdf

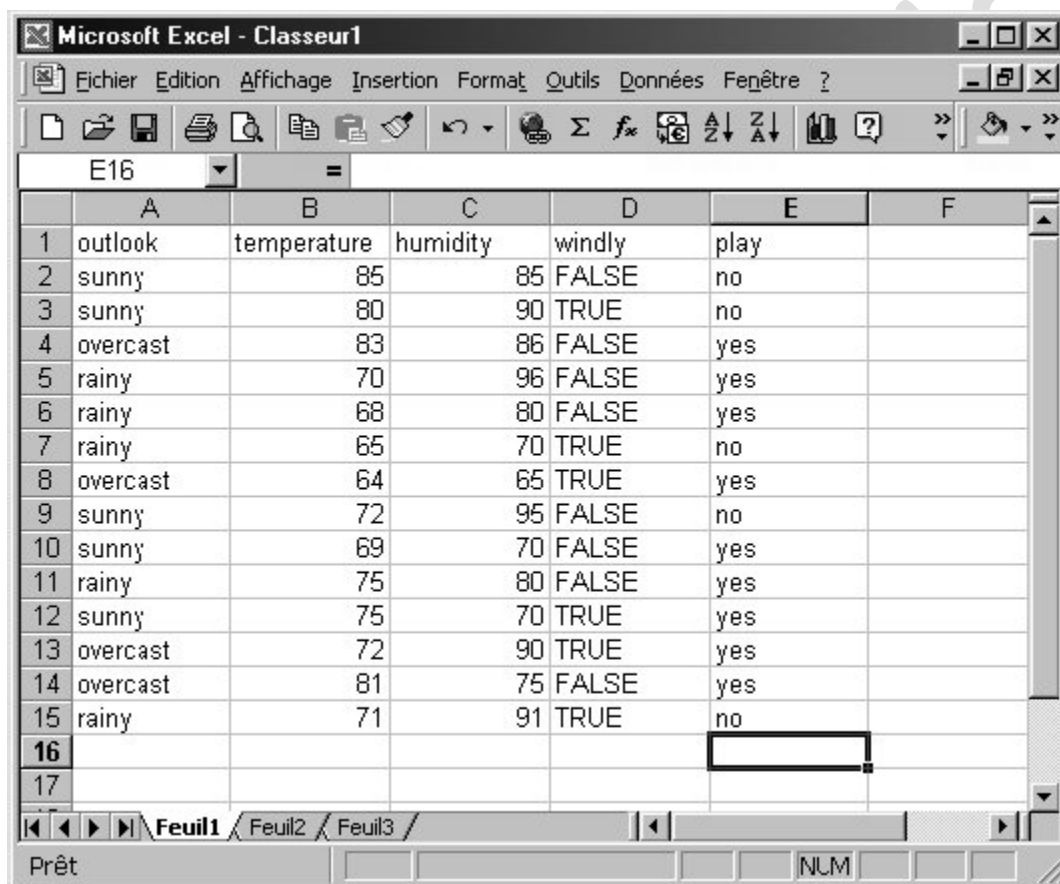


5.1.4 Baze de date

TANAGRA poate lucra cu o singura baza de date odata. Baza de date trebuie sa fie in formatul unui fisier text care contine pe prima linie numele atributelor bazei de date, separate de tab, iar pe urmatoarele linii valorile atributelor respective, câte o linie pentru fiecare înregistrare. Deoarece de multe ori bazele de date sunt memorate in formatul unui fisier Excel, vom arata cum poate fi importata o astfel de bază de date în TANAGRA.

Salvarea unui fisier Excel în format TANAGRA:

- Deschideti Excel
- Introduceti datele ca in exemplul de mai jos:



	A	B	C	D	E	F
1	outlook	temperature	humidity	windy	play	
2	sunny	85	85	FALSE	no	
3	sunny	80	90	TRUE	no	
4	overcast	83	86	FALSE	yes	
5	rainy	70	96	FALSE	yes	
6	rainy	68	80	FALSE	yes	
7	rainy	65	70	TRUE	no	
8	overcast	64	65	TRUE	yes	
9	sunny	72	95	FALSE	no	
10	sunny	69	70	FALSE	yes	
11	rainy	75	80	FALSE	yes	
12	sunny	75	70	TRUE	yes	
13	overcast	72	90	TRUE	yes	
14	overcast	81	75	FALSE	yes	
15	rainy	71	91	TRUE	no	
16						
17						

- Din meniul File selectati comanda "Save as...". Se va deschide o fereastră si la "Type of file" alegeti « Text (Tab delimited) ». Introduceti numele fisierului: weather.txt
- Click pe butonul Save. O fereastră dialog se va deschide și vă va avertiza că, în formatul text, doar un singur sheet poate fi salvat în același timp. Click OK.
- Datele sunt pregătite să fie importate în TANAGRA. Inchideti programul Excel.

Fisierul text poate fi creat cu orice editor de texte cu singura restrictie ca valorile atributelor aceleasi inregistrari sa fie separate de tab.

5.1.5 Importul datelor în TANAGRA

- Alegeți "File/New..." din meniul principal.
- Introduceți un titlu pentru diagrama.
- Introduceți un nume pentru fisierul în care veți lucra.
- Pentru Dataset alegeți fisierul pe care l-ați creat "weather.txt".
- Click OK.

O diagrama nouă s-a creat și puteți vedea conținutul ei în fereastra din dreapta: în primul rând, care este baza de date și ce atribute are și de ce tip (discrete sau continue).

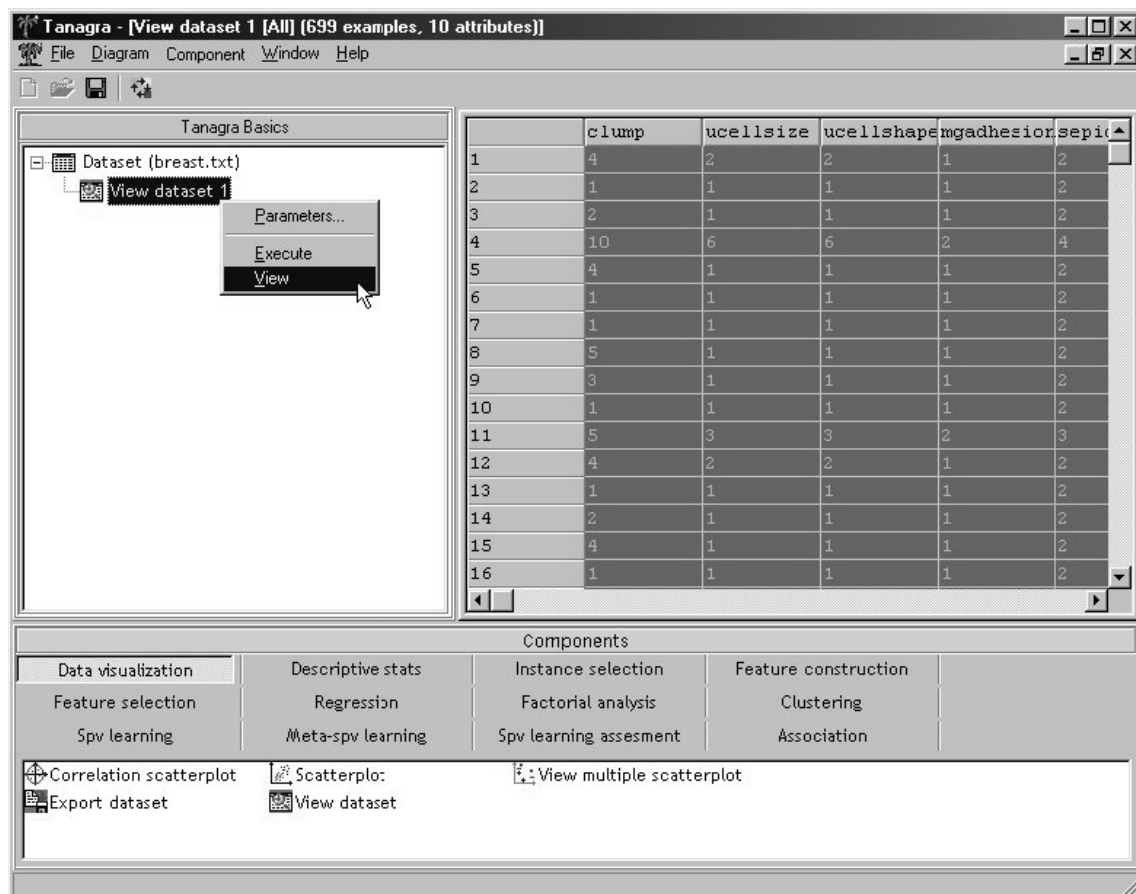
5.1.6 Precizări privind formatele diagramelor

Pentru crearea diagramei am folosit extensia bdm. Aceasta este extensia pentru un fisier TANAGRA ce conține, pe lângă descrierea diagramei, și baza de date importată (aici fisierul weather.txt). Astfel, se va pierde legătura cu fisierul text original, care poate fi modificat sau șters fără a avea vreo influență asupra fisierului TANAGRA. Pe de altă parte, tdm este extensia pentru un fisier TANAGRA ce conține, pe lângă descrierea diagramei și o referință la baza de date importată (aici fisierul weather.txt). Astfel, dacă se face o modificare în baza de date, data următoare când se va executa fisierul tdm, se vor obține rezultate diferite decât cele obținute inițial. De asemenea, dacă din greșeală, ați șters fisierul text, nu veți mai putea deschide diagrama creată și stocată în fisierul tdm.

5.1.7 Vizualizarea bazei de date

Pentru vizualizarea bazei de date se va adăuga un nou operator diagramei și anume:

- Click pe tab-ul Data visualization. Selectați componenta View Dataset și adăugați-o sub nodul "Dataset"
- Dublu-Click pe nodul "View dataset" (sau click dreapta și se alege View) și baza de date este afișată în fereastra din dreapta.



Fisierul text poate fi creat cu orice editor de texte cu singura restrictie ca valorile atributelor aceleasi inregistrari sa fie separate de tab.

5.1.8 Folosirea operatorului Define status

In TANAGRA aproape toate operatiile cer definirea atributelor care urmeaza sa fie folosite si cum vor fi folosite. Acest lucru se face folosind operatia Define status.

- Pentru aceasta, se alege tab-ul si de aici operatorul Define status. Se selecteaza si se pune sub nodul Dataset. Daca a fost pus gresit, puteti sa il stergeti alegand din meniul Diagram ->Delete component.
- Apoi click dreapta pe nodul "Define status" si din meniul care apare se alege comanda *Parameters*.
- In fereastra de dialog care apare puteti alege ca variabile ce attribute sa fie de intrare, care sa fie atributul tinta si care sa fie ilustrative. Observati ca attributele continue sunt marcate de un C albastru iar cele discrete de un D verde. Click OK.
- Dublu-click pe Define status pentru a executa si vizualiza operatia. In dreapta va apare descrierea operatiei efectuate.

5.1.9 Operatori

Prezentam mai jos o lista de operatori (componente) pe care ii vom folosi mai des:

Tab-ul	Operatorul (Componenta)	Descriere
Data visualization	View dataset	Vizualizarea continutului bazei de date
Data visualization	Export dataset	Exportul bazei de date curente in fisier text
Data visualization	ScatterPlot	Graficul unui atribut in fct de altul
Instance selection	Sampling	Dimensiunea esantionului: procent sau numar de exemple
Instance selection	Recover examples	Permutarea statutului exemplilor: cele active devin inactive si invers.
Feature construction	Formula	Crearea unui nou atribut plecand de la o formula algebrica
Feature construction	EqFreq Disc	Discretizare folosind intervale cu aceeasi frecventa. Parametrii: Numar de intervale. Numai pt attribute continue.
Feature construction	EqWidth Disc	Discretizare folosind intervale de lungimi egale. Parametrii: Numar de intervale. Numai pt attribute continue.
Regression	Multiple linear regression	Metoda celor mai mici patrate. Atribute de intrare continue.
Clustering	K-Means	Algoritmul lui Mc Queen. Se fac cateva incercari. Parametrii: - numar de clustere, numar de iteratii, numar de incercari
Supervised learning	Naive Bayes	Clasificare Bayesiană, Parametrii trebuie sa fie discreti
Supervised learning	ID3	Algoritmul ID3. Parametrii: dimensiunea minima a unui nod pentru a fi divizat, dimensiunea minima a unei frunze, adancimea maxima a arborelui de decizie, entropia maxima dobandita
Association	A priori	Algoritmul apriori.

5.1.10 Alte soft-uri pentru data mining:

Alternative comerciale

- [Clementine](#)
- [Insightful Miner](#)
- [SAS Enterprise Miner](#)
- [Tiberius](#)
- [Cart](#)

Alternative Open Source:

- [Tiberius](#)
- [Weka](#)
- [MiningMart](#)
- [Knime](#)
- [RapidMiner \(YALE\)](#)

FORMULAR DE FEEDBACK

În dorința de ridicare continuă a standardelor desfășurării activitatilor dumneavoastră, va rugăm să completați acest chestionar și să-l transmiteți îndrumatorului de an.

Disciplina: Tehnici de data mining

Unitatea de învățare/modulul: _____

Anul/grupa: _____

Tutore: _____

Partea I

1. Care dintre subiectele tratate in aceasta unitate/modul considerați că este cel mai util și eficient? Argumentati raspunsul.

2. Ce aplicatii/proiecte din activitatea dumneavoastra doriți să imbunatatiti/modificați/implementați în viitor în urma cunoștințelor acumulate în cadrul acestei unitati de învățare/modul?

3. Ce subiecte considerați că au lipsit din acesta unitate de învățare/modul?

4. La care aplicatii practice ati intampinat dificultati in realizare? Care credeti ca este motivul dificultatilor intalnite?

6. Timpul alocat acestui modul a fost suficient?

7. Daca ar fi sa va evaluati, care este nota pe care v-o alocati, pe o scala de la 1-10?. Argumentati.

Partea II. Impresii generale

1. Acest modul a întrunit așteptările dumneavoastră?

☐ În totalitate ☐ În mare măsură ☐ În mică măsură ☐ Nu

2) Aveți sugestii care să conducă la creșterea calității acestei unitati de invatare/modul?

3) Aveți propuneri pentru alte unitati de invatare?

Vă mulțumim pentru feedback-ul dumneavoastră!

UTM - Informatica