

Facultatea de Electronică, Telecomunicații și Tehnologie Informației
AIM AI Multimedia Lab
<https://www.aimmultimedialab.ro/>
Universitatea Politehnică din București

Tehnici de Analiză și Clasificare Automată a Informației

Prof. dr. ing. Bogdan IONESCU
<https://bionescu.aimmultimedialab.ro/>

București, 2022

1

Organizare disciplină

Prof. Bogdan Ionescu
<https://bionescu.aimmultimedialab.ro/>

Dr. Șerban Carată
<https://serbancarata.ro/>

Materiale curs:
<https://bionescu.aimmultimedialab.ro/courses.html#takai>

Punctaj:

- examen (scris) 50%;
- colocviu laborator (practic) 50%.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

2

Plan Curs

- M1. Introducere (concept, aplicații)
- M2. Prelucrarea și reprezentarea datelor de intrare
- M3. Tehnici de clasificare ne-supervizată ("clustering")
- M4. Tehnici de clasificare supervizată ("classification")
- M5. Evaluarea performanței clasificatorilor

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

3

Bibliografie

[1] Curs

[2] I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, 2011

[3] Internet
[găsiți punctual orice informație IEEE, ACM, Springer, etc.]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

4

> M1. Introducere

- 1.1. [Introducere]
- 1.2. [Aplicații concrete]
- 1.3. [Conceptul de învățare]
- 1.4. [Terminologie]
- 1.5. [Tehnici existente]
- 1.6. [Utilitare]

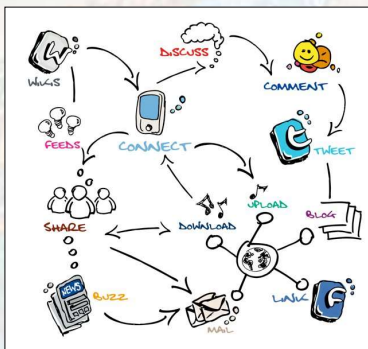
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

5

Informație

> volum imens de date (**Big Data**) care nu mai poate fi "gestionat" (vizualizat, analizat, înțeles, prelucrat) de către utilizator;

+ complexitatea datelor depășește de multe ori puterea de înțelegere și de procesare umană.



Exemplu date multimedia

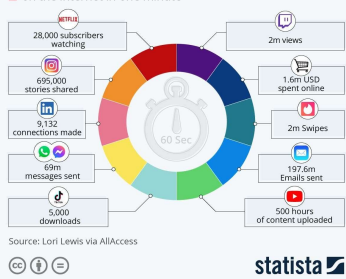
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

6

Informație (cont.)

A Minute on the Internet in 2021

Estimated amount of data created on the internet in one minute



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

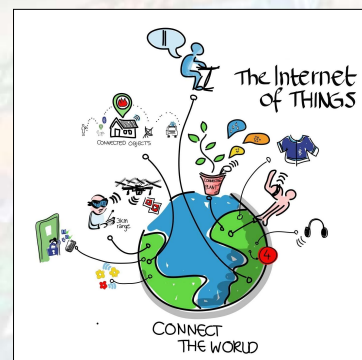
7

7

Informație (cont.)

[sursă imagine Wikipedia]

> în mod constant, apar noi "generatori" de informație, ex. dispozitive portabile, Internet-ul lucrurilor, roboți etc.



Exemplu date multimedia

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

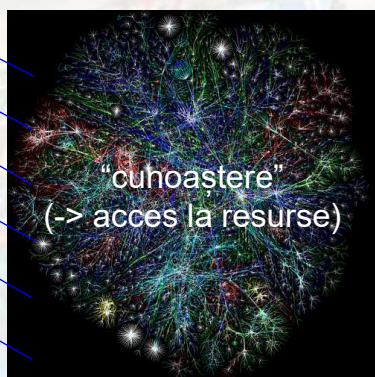
8

8

Internet (de facto)

[sursă imagine Patrick Barry, Flickr]

- știri
- rețele sociale
- divertisment
- securitate
- date financiare
- date medicale
- ...

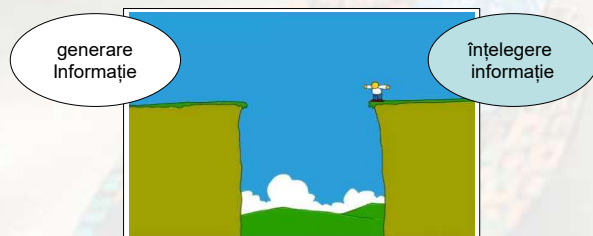


Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

9

9

Punerea problemei



> "bridge the gap", înțelegere și descoperire informație ascunsă (șabloane) ce poate fi utilă și care nu este exploatată;

[o problemă veche de când lumea: vânătorii încercau să înțeleagă comportamentul de migrare al animalelor, fermierii încercau să înțeleagă comportamentul culturilor, etc]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

10

10

Obiective clasificare informație

- > **reducerea volumului informațional:**
 - regruparea unui ansamblu de date în grupuri omogene și determinarea informație reprezentative;
 - eliminarea redundanței informaționale.
- > **punerea în evidență a "cunoașterii":**
 - localizarea într-un volum mare de date a unor grupuri de informații ce prezintă anumite caracteristici de interes;
 - o nouă înțelegere a relațiilor existente între date.
- > **punerea în evidență a datelor atipice:**
 - localizarea datelor ce nu corespund criteriilor considerate, în particular interesante prin natura acestora.
- > **rezolvarea unor probleme de calcul.**

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

11

11

Un exemplu, "weather problem"

[J.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Date vreme

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	foarte cald	ridică	nu	Nu
2	însorită	foarte cald	ridică	da	Nu
3	înnorat	cald	ridică	da	Da
4	ploioasă	cald	normală	nu	Da
5	ploioasă	rece	normală	nu	Da

> dispunem de patru atribute măsurabile: vreme (3), temperatură (3), umiditate (2) și vânt (2) și trebuie să determinăm automat dacă putem practica o activitate;

> conform datelor avem $3 \times 3 \times 2 \times 2 = 36$ de combinații posibile, din care dispunem doar de 5 seturi de date.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

12

12

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Date vreme

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	ridicată	da	Nu
3	înnorat	cald	ridicată	da	Da
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

Dacă (vreme==însorită) && (umiditate==ridicată) -> sport=Nu;
 Dacă (vreme==ploioasă) && (vânt==da) -> sport=Nu;
 Dacă (vreme==înnorat) -> sport=Da;
 Dacă (umiditate==normală) -> sport=Da;
 altfel -> sport=Da;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

13

13

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Dacă (vreme==însorită) && (umiditate==ridicată) -> sport=Nu;
 Dacă (vreme==ploioasă) && (vânt==da) -> sport=Nu;
 Dacă (vreme==înnorat) -> sport=Da;
 Dacă (umiditate==normală) -> sport=Da;
 altfel -> sport=Da;

> analizate în ordine clasifică corect toate exemplele din tabel?

> scoase din context nu mai sunt adevărate ceea ce înseamnă că un set de reguli depinde de modul în care este interpretat.

Dacă (umiditate==normală) -> sport=Da;

nr.	vreme	temperatură	umiditate	vânt	sport
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

14

14

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Date vreme

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	ridicată	da	Nu
3	înnorat	cald	ridicată	da	Da
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

> se poate merge mai departe, pe baza datelor să determinăm reguli de asociere care corelează atributele:

Dacă (temperatură==rece) -> umiditate=normală;
 Dacă (umiditate==normală) && (vânt==nu) -> sport=da;
 Dacă (vreme==însorită) && (sport==nu) -> umiditate=ridicată;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

15

15

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Date vreme

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	ridicată	da	Nu
3	înnorat	cald	ridicată	da	Da
4	ploioasă	rece	normală	da	Nu
5	ploioasă	rece	normală	nu	Da

(1) **date incomplete**: sistemul trebuie să fie capabil de generalizare pentru exemple noi, exemplu folosind cele 5 să putem prezice restul de 31 de situații?

nr.	vreme	temperatură	umiditate	vânt	sport
6	însorită	cald	normală	nu	?

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

16

16

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Date vreme

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	-	da	Nu
3	înnorat	-	ridicată	da	Da
4	ploioasă	-	-	da	Nu
5	ploioasă	rece	normală	nu	Da

(2) **date incomplete**: datele de intrare pot fi incomplete, sistemul trebuie să fie capabil de generalizare și în această situație:

nr.	vreme	temperatură	umiditate	vânt	sport
6	însorită	cald	normală	nu	?

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

17

17

Un exemplu, "weather problem" (cont.)

[I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Date vreme

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	foarte cald	ridicată	nu	Nu
2	însorită	foarte cald	0	da	Nu
3	înnorat	%&#	ridicată	da	Da
4	ploioasă	&@##	0	da	Nu
5	ploioasă	rece	normală	nu	Da

(3) **date eronate**: regulile clasifică corect exemplele dar datorită erorilor datelor (ex. zgomot) în realitate clasificatorul nu este capabil să clasifice corect chiar datele pe baza cărora a fost definit.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

18

18

Aplicații concrete

> accentul se pune pe abilitatea de a generaliza pe **date noi** despre care nu avem nici o informație a priori;

• Motoarele de căutare de pe Internet:

- învață folosind datele de la utilizator, ex. cuvintele folosite la căutare și gradul de satisfacție al utilizatorului;
- învață care pagini sunt mai relevante, ex. PageRank de la Google care definește "prestigiul" unei pagini în funcție de cât de corelată este cu alte pagini care la rândul lor sunt "prestigioase" etc;
- învață din istoricul de navigare pentru a recomanda produse și reclame, ex. platforme de comercializare cărți, filme, etc.

[I.H. Witten, E. Frank, M.A. Hall, "Data Mining: Practical Machine Learning Tools and Techniques"]

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

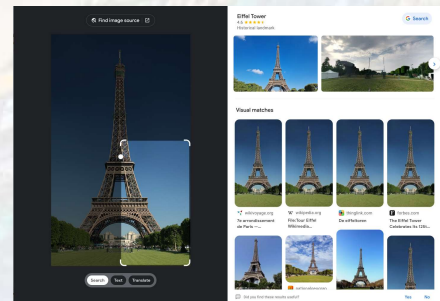
19

19

Aplicații concrete (cont.)

• Motoarele de căutare de pe Internet (cont.):

- > căutare după conținut date multimedia (audio, imagini, video);



Google Lens

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

20

20

Aplicații concrete (cont.)

• Sisteme decizionale: ex. acordarea unui credit

- *procedura tradițională*: solicitantul furnizează o serie de date personale și financiare iar creditorul folosește metode statistice și decizia umană pentru "accept" sau "refuz";
- în X% din cazuri, datele nu se încadrează clar în cele două clase iar decizia este luată de un operator uman ("**bordeline cases**");
- în ~1/2 din cele X% cazuri, decizia se dovedește greșită solicitantul creditat eșuând să returneze creditul -> problemă \$;
- *folosind clasificare*: predicție comportament al celor X% cazuri limită prin antrenare folosind date etichetate: cazuri limită din trecut pentru care se știe rezultatul -> *îmbunătățire decizie*.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

21

21

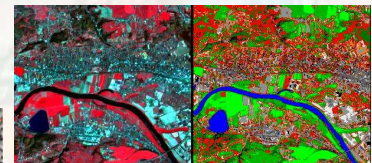
Aplicații concrete (cont.)

• Analiză imagistică:

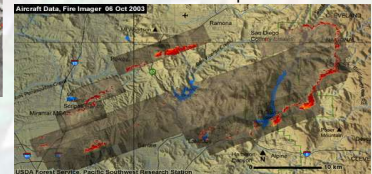
- *satelitară*;



detectare automată străzi



clasificare automată tip de teren



deteție automată incendii

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

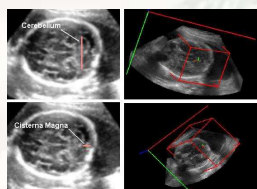
22

22

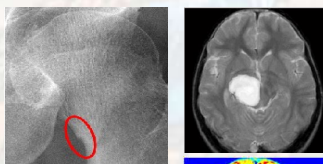
Aplicații concrete (cont.)

• Analiză imagistică (cont.):

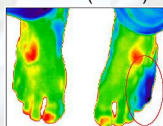
- *medicală*;



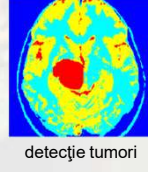
analiză automată (ultrasunete)



detectie automată fractură (raze X)



deteție automată probleme circulație (termografie)



detectie tumori

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

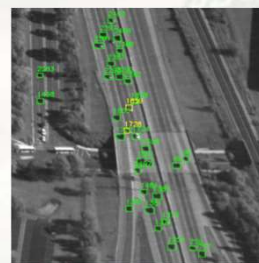
23

23

Aplicații concrete (cont.)

• Analiză imagistică (cont.):

- *securitate*;



monitorizare automată trafic



deteție automată persoane



deteție automată acțiuni

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

24

24

Experimentul 1.2.

> să se regrupeze datele similare:

date de intrare

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU 31

31

Experimentul 1.2. (cont)

Q1: care a fost criteriul de decizie al numărului de clase rezultate?

Q2: cum am decis că două date sunt similare sau nu?

Q3: ce s-a schimbat față de experimentul anterior (1.1.)?

Q4: a fost mai ușor să partiționăm "optimal" datele în această variantă?

date de intrare

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU 32

32

Concluzii experimentul 1.1. și 1.2.

- **clasificare ne-supervizată**, nu avem cunoștințe "a priori" despre semnificația și apartenența datelor;
- definitorii pentru partiționarea datelor au fost parametrii de conținut ai acestora (= descriptori) – aumite proprietăți se dovedesc a fi mai importante (= discriminatorii) decât altele (ex. culoare vs. formă);
- un număr mai mare de descriptori tinde să fie mai relevant pentru succesul (optimizarea) partiționării datelor;
- procesul este guvernat de definirea unei metrici (măsuri de distanță) pe baza căreia se evaluează similaritatea datelor;
- există date atipice, ce nu aparțin niciunei clase.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU 33

33

Experimentul 2.1.

clasa 1

clasa 2

clasa 3

clasa 4

date etichetate

NB: notați caracteristicile claselor pe hârtie.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU 34

34

Experimentul 2.2.

> să se regrupeze datele similare:

clasa 1

clasa 2

clasa 3

clasa 4

date de intrare

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU 35

35

Experimentul 2.2. (cont)

Q1: care a fost criteriul de decizie al numărului de clase rezultate?

Q2: cum am decis că două date sunt similare sau nu?

Q3: există o singură partiționare "optimală"?

Q4: ce se întâmplă cu datele care nu aparțin practic niciunei clase?

date de intrare

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU 36

36

Concluzii experimentul 2.1. și 2.2.

- **clasificare supervizată**, sistemul este "antrenat" în prealabil să răspundă la anumite clase de date;
- se cunoaște numărul de clase de ieșire (sau se determină "a priori" în funcție de datele de antrenare);
- definitorii pentru învățare (și astfel clasificare) au fost parametrii de conținut ai acestora (= descriptori);
- procesul este guvernat de definirea unei metrici (măsură de distanță) pe baza căreia se evaluează similaritatea datelor;
- învățarea nu este perfectă, clasificarea acelorași date de antrenare nu conduce la rezultate perfecte;
- există date atipice, acesta sunt asociate obligatoriu unei clase.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

37

37

Definiție învățare

Definiție "learning": a dobândi cunoștințe sau aptitudini prin studiu, practică, experimentare sau prin intermediul altor persoane.
[dicționar Merriam-Webster]

Definiție "machine learning": un sistem învață din experiența E cu privire la o anumită clasă de cerințe T și o anumită măsură de performanță P , dacă performanța acestuia la cerințele din T , așa cum este măsurată de P , se îmbunătățește pe baza experienței din E .
[Rossen Dimov, Seminar A.I. Tools]

Exemplu:

- T = joc de șah;
- P = procentul de partide câștigate;
- E = 1000 de înregistrări a unor jocuri de șah.

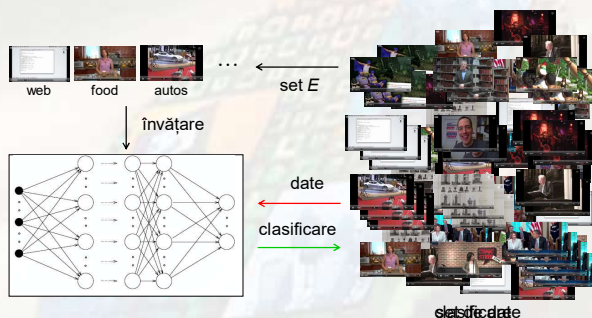
> posibilitatea unui sistem de a "învăța" pe baza unor date;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

38

38

Definiție învățare (cont.)



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

39

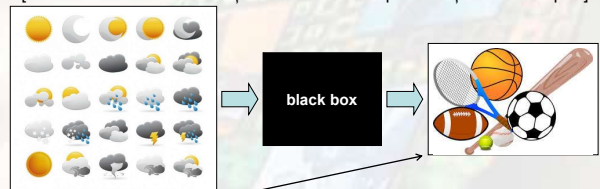
39

Terminologie

[sursă imagini Wikipedia]

concept = ceea ce se dorește a fi învățat de către sistem;

[clasificare automată în funcție de date meteo a posibilității de a face sport]



descriere concept = ceea ce produce sistemul de învățare (ieșire sistem – reprezentare concept prin sistem);

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

40

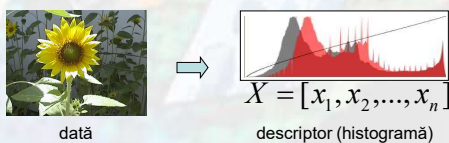
40

Terminologie (cont.)

[sursă imagini Wikipedia]

dată = o entitate de informație unitară, exemplu: o imagine, o secvență video, un document, o înregistrare a unor parametri etc.

descriptor (observație, vector de caracteristici) = reprezentarea unei date într-o formă mai compactă, de regulă vectorială. Valorile vectorului reprezintă măsurători ale unor proprietăți definitorii ale datei respective:



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

41

41

Terminologie (cont.)

[sursă imagine Wikipedia]

atribut (caracteristică, trăsătură) = o componentă a vectorului descriptor ce definește practic una dintre dimensiunile acestuia:

$$X = [x_1, x_2, \dots, x_n] \Rightarrow \begin{array}{l} \text{atribut 1: } x_1 \\ \text{atribut 2: } x_2 \\ \dots \\ \text{atribut n: } x_n \end{array}$$

descriptor

> descriptor = ansamblu de valori ale atributelor;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

42

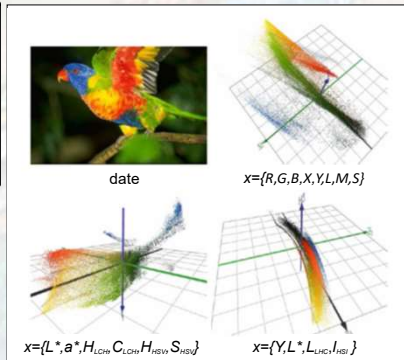
42

Terminologie (cont.)

spațiu de caracteristici

= spațiul definit de descriptorii datelor; axele acestuia sunt definite de atributele descriptorului:

În exemplu:
 > data = pixel imagine;
 > atribute = proiecțiile acestuia pe diferite spații de culoare;



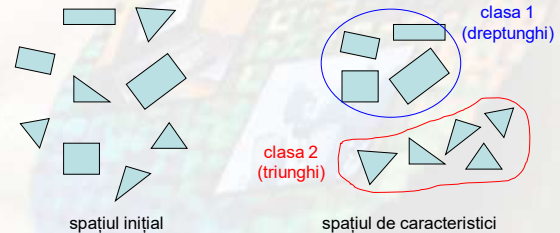
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

43

43

Terminologie (cont.)

clasă = o sursă de date a căror distribuție în spațiul de caracteristici este guvernată de o anumită densitate de probabilitate specifică; astfel, o clasă definește un anumit tip de date cu proprietăți similare din punct de vedere al unor anumite criterii:



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

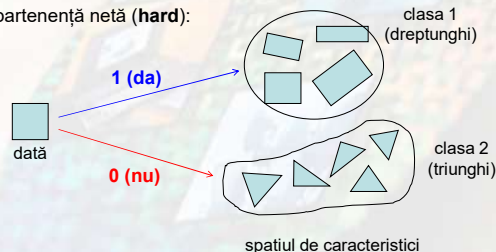
44

44

Terminologie (cont.)

apartenență la clasă = asocierea unei date la o anumită clasă și astfel determinarea faptului că proprietățile acesteia sunt reprezentative pentru specificul clasei respective;

> apartenență netă (**hard**):



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

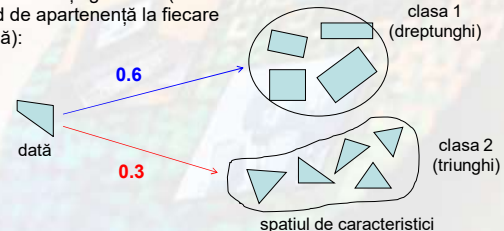
45

45

Terminologie (cont.)

apartenență la clasă = asocierea unei date la o anumită clasă și astfel determinarea faptului că proprietățile acesteia sunt reprezentative pentru specificul clasei respective;

> apartenență graduală (**soft** – grad de apartenență la fiecare clasă):



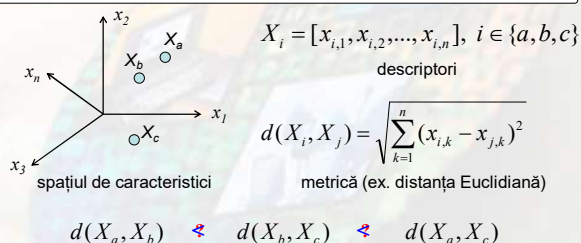
Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

46

46

Terminologie (cont.)

metrică (distanță, măsură de similaritate) = o măsură de evaluare a gradului de similaritate între date diferite. De regulă returnează o valoare mică când datele sunt similare (ex. aparțin aceleiași clase) și o valoare semnificativă când sunt diferite:



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

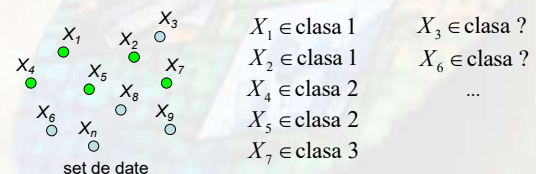
47

47

Terminologie (cont.)

set de date (bază de date) = ansamblul datelor ce urmează să fie analizate și clasificate;

date etichetate (ground truth) = o colecție de date pentru care se cunoaște "a priori" apartenența la clase; acestea sunt de regulă obținute pe baza expertizei umane:



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

48

48

Terminologie (cont.)

> rezumat: set de date

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicăta	nu	Nu
3	ploioasă	cald	ridicăta	nu	?
4	înnorat	rece	normală	da	?

set de date

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

49

49

Terminologie (cont.)

> rezumat (cont.): atribute

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicăta	nu	Nu
3	ploioasă	cald	ridicăta	nu	?
4	înnorat	rece	normală	da	?

atribute

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

50

50

Terminologie (cont.)

> rezumat (cont.): atribut clasă

nr.	vreme	temperatură	umiditate	vânt	sport
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicăta	nu	Nu
3	ploioasă	cald	ridicăta	nu	?
4	înnorat	rece	normală	da	?

atribut clasă

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

51

51

Terminologie (cont.)

> rezumat (cont.): instanță

nr.	vreme	temperatură	umiditate	vânt	play
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicăta	nu	Nu
3	ploioasă	cald	ridicăta	nu	?
4	înnorat	rece	normală	da	?

instanță (dată)

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

52

52

Terminologie (cont.)

> rezumat (cont.): date etichetate

nr.	vreme	temperatură	umiditate	vânt	play
1	însorită	cald	normală	da	Da
2	însorită	foarte cald	ridicăta	nu	Nu
3	ploioasă	cald	ridicăta	nu	?
4	înnorat	rece	normală	da	?

date etichetate

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

53

53

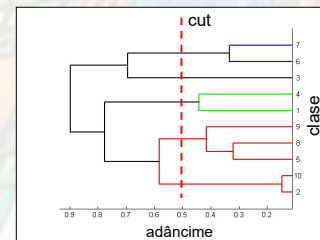
Tehnici existente

[sursă imagine Wikipedia]

1. Tehnici de clasificare nesupervizată (clustering)

- metode ierarhice:

[datele de intrare sunt grupate într-un număr variabil de clase sub forma unui arbore (dendrogram) pornind de la toate elementele într-o clasă și finalizând cu fiecare element într-o clasă separată]



Metode: Hierarchical Clustering cu diferite variante, agglomerative - "bottom up" sau divisive - "top down".

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

54

54

Tehnici existente (cont.)

1. Tehnici de clasificare nesupervizată (cont.)

- metode partiționale:

[produc o singură partiție și depind de alegerea numărului de clase de ieșire. Partiționarea se face folosind un criteriu de optimizare căutându-se prin încercări partiționarea optimă. Algoritmul este de regulă rulat repetitiv cu diferite puncte de plecare a partițiilor alegându-se în final varianta cea mai bună]

Metode:

- square error: k-means, ISODATA;
- graph-theoretic: Minimal Spanning Tree (MST);
- mixture resolving: Expectation Maximization (EM);
- nearest neighbor;
- fuzzy: fuzzy c-means (FCM).

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

55

55

Tehnici existente (cont.)

1. Tehnici de clasificare nesupervizată (cont.)

- o altă clasificare globală:

• **acumulative vs. partiționale:** clasificarea pornește de la o anumită partiție în clase, clase care pe parcursul algoritmului sunt fuzionate iterativ până când este satisfăcut un anumit criteriu de convergență vs. clasificarea pornește de la o singură clasă care este divizată iterativ până când criteriul de convergență este satisfăcut;

• **politetice vs. monotetice:** la stabilirea claselor sunt folosite toate atributele de intrare vs. atributele de intrare sunt folosite în mod secvențial pentru a constitui progresiv clasele, ex. atributul x_1 este folosit pentru a diviza datele în două clase, mai departe, atributul x_2 este folosit pentru divizarea claselor anterioare, și așa mai departe;

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

56

56

Tehnici existente (cont.)

1. Tehnici de clasificare nesupervizată (cont.)

- o altă clasificare globală (cont.):

• **nete vs. fuzzy:** datele sunt alocate unei singure clase, apartenența fiind binară (1 sau 0) vs. datele au un grad de apartenență la una sau mai multe clase - cu cât valoarea este mai mare cu atât este mai probabil să aparțină clasei respective;

• **deterministe vs. stohastice:** optimizarea claselor este deterministă pe baza unui algoritm determinist vs. se folosește o căutare aleatoare în spațiul format de toate clasificările posibile;

• **incrementale vs. non-incrementale:** volum foarte mare de date (ex. Big Data) - minimizare număr de citiri al datelor, reducere număr de repartiții în clase analizate, reducere date, partiționare progresivă crescând setul de date.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

57

57

Tehnici existente (cont.)

2. Tehnici de clasificare supervizată (classification)

- bazate pe criteriul Bayes:

[clasificator probabilistic, de regulă binar (două clase), ce se bazează pe o ipoteză de independență a atributelor de intrare (naivă); fiecărei clase i se asociază o probabilitate, $p(C_i | x_1, \dots, x_n)$ unde C_k - clasa k iar x sunt datele de clasificat; ieșirea clasificatorului este clasa cea mai probabilă (optimizare în funcție de datele de antrenare)]

Metode:

- Naive Bayes;
- Bayes Networks;
- AODE, etc.

$$p(C_k | x) = \frac{p(C_k) p(x | C_k)}{p(x)}$$

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

58

58

Tehnici existente (cont.)

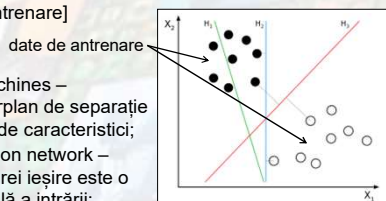
2. Tehnici de clasificare supervizată (cont.)

- bazate pe funcții:

[problema clasificării este modelată cu ajutorul unei reprezentări funcționale a datelor de intrare, reprezentare ce este optimizată folosind datele de antrenare]

Metode:

- Support Vector Machines - optimizează un hiperplan de separație a datelor din spațiul de caracteristici;
- Radial Basis Function network - rețea neuronală a cărei ieșire este o combinație funcțională a intrării;
- Linear Regression - asocierea optimă a unor funcții liniare perechilor de date de intrare-ieșire (date de antrenare), etc.



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

59

59

Tehnici existente (cont.)

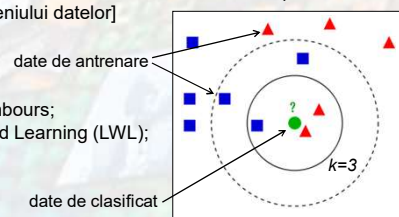
2. Tehnici de clasificare supervizată (cont.)

- metode "leneșe":

[antrenarea propriu-zisă este de fapt realizată cu întârziere de abia în momentul clasificării unei date noi; clasificarea este optimizată local, pentru fiecare dată nouă, ceea ce le face adaptabile la modificarea domeniului datelor]

Metode:

- k-Nearest Neighbours;
- Locally Weighted Learning (LWL);
- etc.



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

60

60

Tehnici existente (cont.)

2. Tehnici de clasificare supervizată (cont.)

- bazate pe reguli de decizie:

[se bazează pe generarea și optimizarea unui set de reguli de decizie de tip "dacă – atunci" folosind datele de antrenare; regulile nu sunt neapărat exclusive]

vârstă=16, prescripție="miopie", astigmatism=0, lacrimi="redușe" - lentile=0;
vârstă=14, prescripție="miopie", astigmatism=1, lacrimi="redușe" - lentile=0;
...

dacă (vârstă>14 && vârstă<16) && (prescripție=="miopie") &&
(lacrimi=="redușe") **atunci** lentile=0;

Metode:

- Decision Table;
- Ripple-Down Rule learner (RIDOR);
- etc.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

61

61

Tehnici existente (cont.)

[sursă imagine Wikipedia]

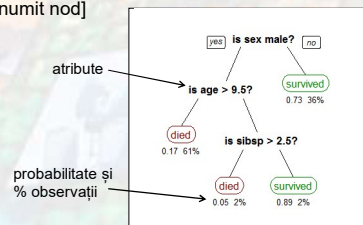
2. Tehnici de clasificare supervizată (cont.)

- bazate pe arbori:

[reprezintă informația sub formă arborescentă, unde clasele sunt reprezentate de frunzele arborelui, nodurile corespund atributelor datelor iar ramurile reprezintă relaționarea valorilor atributelor pentru a ajunge la un anumit nod]

Metode:

- Functional Trees;
- Random Trees;
- Random Forests;
- C4.5, etc.



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

62

62

Tehnici existente (cont.)

2. Tehnici de clasificare supervizată (cont.)

- meta-metode:

[reunesc mai mulți clasificatori existenți; de regulă învață în mod iterativ un set de clasificatori "slabi" și îi adaugă progresiv la clasificatorul global; pe măsură ce sunt adăugați, datele sunt ponderate, datelor clasificate eronat li se crește ponderea în timp ce datelor clasificate corect le scade ponderea; astfel încât următorul clasificator slab se va focaliza pe datele clasificate greșit]

Metode:

- AdaBoost;
- LogitBoost;
- AnyBoost, etc.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

63

63

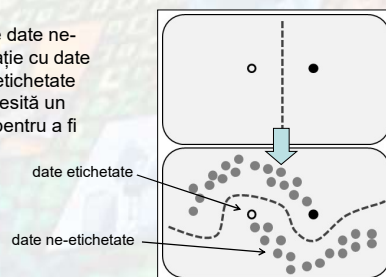
Tehnici existente (cont.)

[sursă imagine Wikipedia]

3. Tehnici de clasificare semi-supervizată

- se poziționează între tehnicile de clasificare nesupervizată și cele supervizate;

- **idee:** folosirea de date ne-etichetate în corelație cu date etichetate (datele etichetate sunt limitate și necesită un efort considerabil pentru a fi obținute);



Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

64

64

Utilitare



- Neural Network Toolbox;
- Bioinformatics Toolbox (Support Vector Machines);
- Statistics Toolbox (Hierarchical Clustering, K-Means, Gaussian Mixture Models, Naive Bayes, Discriminant Analysis, Nearest Neighbors, Classification Trees, Ensemble Classifiers, etc);
- poate fi completat cu alți clasificatori (Internet).



- platformă specializată, "open source", în Java;
- furnizează tot lanțul de prelucrare: pregătire date, selecție atribute, clasificare supervizată și nesupervizată, evaluare performanțe, analiză rezultate, etc;
- poate rula în linie de comandă (batch processing);
- poate rula multi-procesor.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

65

65

Utilitare (cont.)



- platformă generală de computer vision ce include și facilități de clasificare, "open source", în C++;
- Machine Learning Library (MLL): Statistical Models, Bayes Classifier, K-Nearest Neighbors, Support Vector Machines, Decision Trees, Boosting, Expectation Maximization, Neural Networks;
- poate rula multi-procesor.



- platformă specializată, comercială;
- mod de operare vizuală, fără programare;
- folosită la nivel global de companii care prelucrează și analizează date (ex. financiare);
- pachet de utilitare "puternice", validate.

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

66

66

> Sfârșit M1

Tehnici de analiză și clasificare automată a informației, Prof. Bogdan IONESCU

67