**Ce este analiza componentelor principale.Evidentiati 5 categorii de pb care pot fi solutionate cu ajutorul tehnicilor de analiza a componentelor principale.**

ACP este o tehnica de analiza multidimensionala care are ca scop descompunerea variabilitatii totale din spatiul cauzal initial sub forma unui nr. redus de componente si fara ca aceasta descompunere sa contina redundante informationale.ACP este o tehnica de analiza multidimensionala care are ca scop reducerea dimensionalitatii spatiului cauzal initial,in conditiile unei pierderi informationale minime.ACP repr.procedura algoritmica de analiza multidimensional ace consta in determinarea unor noi variabile,numite componente principale,definite sub forma combinatiilor liniare de variabile initiale cu varianta maxima.5 categorii de probleme sol.cu ACP:a)reducerea complexitatii datelor(data reduction)poate fi inlocuit un masiv de date de mari dimensiuni prin masive de dimensiuni mai mici.b)evidentierea si fixarea patternului asocierilor dintre variabile.c)det.variab.latente care se afla in spatele variab.masurate.

**Interpretati logica analizei componentelor principale(inclusiv d.p.d.v geometric).**

D.p.d.v geometric,variabilele numite componente principale definesc un nou spatiu al obiectelor.Cele mai interesante si mai utile aspect ale analizei componentelor principale sunt in primul rand legate,nu de aparatul mathematic pe care aceasta analiza se bazeaza,ci de multiplele si nuantele interpretari posibile pe care aceasta le ofera.Logica analizei componentelor principale se bazeaza pe ideea fundamentala ca se pot face anumite transformari asupra observatiilor initiale,care sa determine maximizarea variantei individuale pt.anumite variabile si minimizarea variantei pt.alte variabile.Et.1:se det.val.proprii corespunz.matricii de covarianta atasata var.initiale prin rezolvarea ecuatiilor.caracteristice:det|ԑ- λI|=0.Et.2:Cele n val.proprii anterior determ.sunt ordonate descrescator λ1≥ λ2≥ λ3≥ λ4≥…≥ λn.Et.3:Pt fiecare val.proprie luata in ordine descresc.se det.vectorul propriu atasat:ԑ\*αi=λi\* αi.Et.4:Se det.noile variab.drept combinatii liniare ale var.originale avand coef.dati de comp.vectorului propriu identificat la et.3.

**Definiti componentele principale si mentionati proprietatile acestora:**

Componentele principale sunt variabile vectoriale abstracte,definite sub forma unor combinatii liniare de variabile initiale si care au urmatoarele 2 proprietati fundamentale: a)sunt necorelate 2 cate 2 si suma patratelor coeficientilor care definesc combinatia liniara ce corespunde unei component principale este egala cu unitatea;b)prima componenta principala este o combinatie liniara normalizata a carei varianta este maxima,cea de-a doua componenta principala este o combinatie liniara necorelata cu prima component principal si care are o variant cat mai mare posibila,insa mai mica decat cea a primei componente. Proprietati:a)axele noului spatiu sunt ortogonale 2 cate 2 si definesc noile variabile numite componente principale;b)coordonatele obiectelor(generate prin proiectii pe axele noului spatiu)sunt denumite scoruri principale si au proprietatea ca pastreaza canfigurarea initiala a acestora;c)nr de componente principale este egal cu nr.de variabile originale;nu toate comp.princ. au insa o semnificatie informationala considerabila,astfel incat cele mai putin semnificative d.p.d.v informational sunt eliminate;d)componentele principale sunt combinatii liniare de varianta maximal ale variabilelor originale;e)componentele principale sunt scalate in functie de magnitudinea variantei acestora,prima fiind componenta principala cu varianta maxima,iar ultima component principal cu variant minima;f)componentele principale sunt necorelate 2 cate 2;g)suma variantelor comp. princ.coincide cu suma variantelor variabilelor originale,a.i. componentele principale preiau in totalitate variabilitatea continuta in variabilele originale.

**Formulati modelul matematic al ACP,definiti si interpretati marimile definitorii ale acestora.**

Opt ф (x,w)

AεMn\*k , unde criteriul de optim poate fi de maxim sau minim, in functie de natura functiei ф

SR: w=At \* t

Daca functia ф este de tip distanta,atunci criteriul de optim va fi reprezentat de minimizarea functiei ф. Daca functia ф este o masura a cantitatii de informative adusa de noua modelitate de repr. a obiectelor,criteriul de optim va fi repr. de max. functiei ф.Combinatiile liniare care definesc comp. princ. pot fi scrie de forma :

w1=α11 x1+α21x2+…+αn1xn =>modelul matematic al ACP:

w2= α12 x1+α22x2+…+αn2xn max Var(w)

wn=α1n x1+α2nx2+…+αnnxn AεMn\*n

w=At \*x

Cele n coloane ale matricii A repr. vectorii proprii normalizati ai matricii de covarianta ∑,iar varianta fiecarei comp.princ. wi,care este o varianta maximala in raport cu variantele comp.princ anterioare,este repr. de valoarea proprie λi a aceleeasi matrici de covarianta.Spatiul cauzal initial supus investigarii este det.de un nr.de n variabile explicative notate x1,x2,…xn.Pt det.comp. princ. wi este necesara det.coef.αji,care defines combinatia liniara coresp.acestei comp.princ.

**Ilustrati modul de deducere a componentelor principale.**

Determinarea coeficientilor combinatiei liniare ce defineste comp.principala w,in conditiile maximizarii variantei acestei comp.principale,este echivalenta cu a alege dintre cele n valori proprii ale mattricii de covarianta ∑ pe cea mai mare si a determina componentele vectorului de ponderi α ce defineste respective comp.principala prin calculul vectorului propriu al matricii ∑ asociat cu acea valoare proprie.

Pt.fiecare valoare proprie λi din cele n valori proprii ale matricii de covarianta ∑,avem cate o solutie a pb.de max: max αt \* ∑ \* α ,adica cate un vector αi si cate o comp.principala wi. α

Presupunand ca cele n valori proprii ale matricii de covarianta ∑ sunt SR: αt \*α=1 ordonate in asa fel incat: λ1≥λ2≥…≥ λn,prima component principal w1,care va avea variant maxima λ1,este data de combinatia liniara: w1=(α1)t \* x.

Vectorul α1este acel vector propriu al matricii de covarianta ∑ caruia ii corespunde valoarea proprie cea mai mare, λ1,adica este vectorul care verifica restrictiile urmatoare:

(∑-λ1\*I)\*α1=0

(α1)t\*α1=1 ,valoarea proprie λ1 este radacina a ecuatiei caracteristice: |∑- λi \*I|=0 iar I este notatia pt matricea unitate.Determinarea in acest fel a componentei principale w1,face ca aceasta sa aiba proprietatile illustrate prin relatiile urm.: E(w1)=(α1)t\*μ. Dupa det.primei comp.princ.w1,urmeaza det.celei de-a doua comp. princ.w,component care trb.caracterizata de urm.proprietati:sa aiba variant maximala si sa fie necorelata cu prima comp.princ.w1.

**Definiti si justificati 3 dintre componentele principale**

Componentele principale sunt vectori ortogonali care preiau cât mai mult din varianta variabilelor vector originale astfel :prima component principal preia maximul posibil din varianta variabilelor originale, a doua component principal preia maximul din varianta ramasa dupa ce este eliminata varianta preluata de prima componenta

a)sunt coordinate obiectilor(generate prin proiectii pe axele noului spatiu)sunt denumite scoruri principale si au proprietatea ca pastreaza configuratia initiala a acestora.

b) componentele principale conserva variant totala a variab initiale.

Var initiale x1,x2,..xn Var(t)=var(x1)+var(x2)+..+var(xn) Var(call) (x1,x2,..xn)=var(t)(w1,w2,…wn)

c) dependent de unitatile de masura.

Componentele principale sunt dependente de ordinal de masura asociat scalei variabilelor .aceasta inseamna ca orice schimbare a ordinului de masura pe scare asociata variabilelor, determina modificare a structurii componentelor principale.

d) sunt necorelate 2 cate 2 w1,w2…wp p<<n cov(wi,wj) ={0, i!=j si 1 i=j)

in situatia in care matricea initiala de date provine dintr-o repartitie normal multidimensional necorelarea componentelor principale este congruent cu independent statistica a acestor.

**Interpretati vectorii si valorile proprii ale matricii de covarianta**

Vectorii proprii ai matricei de covarianta (a unui vector aleator x) pot fi utilizati pentru construirea unei transformari unitare. In urma utilizarii acestei transformari, in noul spatiu de trasaturi rezultat, matricea de covarianta va fi una diagonala. In plus are loc o rotire a sistemului initial de coordonare in care vectorul x era, la inceput reprezentat.

Valorile proprii exprima varianta explicata prin fiecare noua componenta principala si corelatia dintre variabilele vechi.

Valorile proprii ale matricei de covarianta sunt intodeauna numere reale. Parte din aceste valori proprii pot sa fie zero sau/si sa existe grupuri de doua sau mai multe valori proprii care sa fie egale intre ele.

**Ce sunt scorurile principale si cum se determina?**

Coordonatele obiectelor în noul spatiu, adica proiectiile obiectelor pe axele acestuia, sunt evaluarile obiectelor

în raport cu noile variabile si se numesc *scoruri ale componentelor principale* sau *scoruri principale .*In analiza componentelor principale coordonatele obiectelor in spatial redus se mai numesc si scoruri principale.Daca vom presupune ca au fost retinute p componente principale si daca vom nota cu U matricea de dimensiunea n\*p , ale carei coloane sunt cei p vectori proprii care defines cele p component principale, atunci matricea scorurilor poate fi determinat astfel :W(p\*t)=A(p\*n)\* x(n\*t)

Liniile matricii W reprezinta scorurile corespunzatoare noilor variabile sau observatiile celor p componente principale .scorurile principale pot fi folosite in analiza ca substitut al observatiilor originale , simplificand baza informational initiala.Consideram ca este extreme de important sa facem precizarea ca scorurile principale sunt mai potrivite pt a fi folosite in analize deoarece sunt mai putin afectate de erori , ca au o anumita invariant in raport cu erorile, le face sa devina mai robuste in raport cu perturbatiile introduce de erori.

**Ce este matricea factor ?**

Component Matrix sau matricea factor arată corelația dintre variabilele inițiale și componentele principale. Pe baza acestei matrici și a reprezentărilor sale grafice se poate determina informația cuprinsă de fiecare componentă principala. Matricea factor este f.importanta in acp deoarece permite identificarea unor submultimi ale variabilelor originale care se grupeaza in jurul unei component principale permintand identificarea unui component real pt aceste constructii abstracte. Elementele matricii factor se numesc intensitatile factorilor. Ele sunt indicatori ai masurii in care variabilele originale participa la formarea componentelor principale .Cu cat este mai mare valoarea coeficientului de corelatie dintr-o variabila originala si o componenta principal , cu atat este mai adecvata si mai complete exprimarea informational a variabilei originale prin intermediul componentei principale respective.

**Detaliati modul in care pot fi interpretate componentele principale.**

Numarul de componente principale este egal cu numarul grupelor de variabile care sunt foarte puternic corelate între ele.

Analiza componentelor principale este o metoda de *reexprimare* a variabilelor originale sub forma unui numar mai mic de noi variabile, numite componente principale, care sunt combinatii liniare de varianta maxima ale variabilelor originale.

Cu ajutorul componentelor principale se poate defini o structura a dependentei dintre variabilele originale mai simpla si mai clara, deci mai usor de interpretat.

Fiecare component principala ofera informatii despre marimea cantatii de informatie preluata din variabilele originale.

**Criterii de alegere a nr de componente principale:**

1) Criteriul empiric:se determina matricea de covarianta, in matricea de covarianta se identifica grupuri de var. puternic corelate intre ele si ale caror componente sunt slab corelate cu componentele altor grupuri. Nr de comp. princ. necesare e dat de nr de astfel de grupuri identificate.

2) Criteriul procesului de acoperire: potrivit acestui criteriu, nr de comp princ este dat de un θk cat mai mare(e indicat sa fie mai mare de 70-75%):

θi=

VarT=

3)Criteriul povarnisurilor/grohotisurilor: nr de componente principale e obtinut prin reprezentarea grafica a valorilor proprii atasate matricii de covarianta initiale.

4)Criteriul lui Kaiser: nr de componente principale ce urmeaza a fi luat in calcul e dat de valorile porprii mai mari sau egale cu 1.

5)Criteriul statistic: este cel mai riguros.

**Analiza factoriala (definitii+tipuri de probleme ce pot fi rezolvate cu aceasta):**

Def: O analiza de tip multidimensional care are drept obiectiv evidentierea corelatiilor existente la nivelul unei multimi de indicatori cu ajutorul unui nr redus de factori necorelati si ordonati, in fct de variabilitatea impusa de ei. Tipuri de probleme ce pot fi tratate cu ajutorul analizei factoriale: detectarea structurii in relatiile dintre variabile; reducerea numarului variabilelor; rezolvarea multicoliniaritatii in regresia multipla; validarea constructiei unei scale compuse din mai multi itemi;

**Structura generala a modelului de analiza factoriala:**

Modelul analizei factoriale are la baza doua ipoteze fundamentale: prima ipoteza se refera la presupunerea ca nivelul sau valorile unui ansamblu de variabile aleatoare X1,X2,...,Xn se formeaza ca rezultat exclusiv al influentei a 3 categorii de factori:

O multime formata din p factori comuni, f1,f2,...,fp, a caror influenta se considera a se exercita asupra fiecareia dintre cele n variabile considerate

O multime formata din n factori unici, U1,U2,...,Un, a caror influenta se considera a se exercita in mod individual, fiecare factor unic influentand una si numai una din variabilele considerate

O multime de n factori reziduali, ε1,ε2,...εn a caror influenta se considera a fi exercitata tot in mod individual, fiecare factor rezidual influentand cate o singura variabila

Dpdv. Statistic, se considera ca influentele semnificative, care trebuie retinute in analiza, sunt cele exercitatede factorii comuni si unici, in timp ce influentele factorilor reziduali, se considera a avea caracter accidental, nesemnificativ. La nivelul fiecarei variabile, influenta factorului rezidual

corespunzator poate fi considerata a fi neglijabila și este asimilabila erorilor de masurare. Din acest motiv, factorii reziduali se mai numesc si erori.

In ceea ce priveste factorii comuni, exista posibilitatea ca in cazul anumitor variabile influenta lor asupra acestor variabile sa fie neglijabila sau chiar nula, ceea ce inseeamna ca factorii respectivi pot fi eliminati din lista factorilor pentru variabila respectiva. In aceste conditii, este posibil ca schema de influenta pt anumite variabile sa contina mai multi factori comuni, iar pt alte variabile mai putini. Nr de factori comuni cu influenta semnificativa asupra variabilei indicator determina complexitatea variabilei indicator respective.

**Defiiti si interpretati descompunerea variabilitatii in contextul analizei factoriale:**

In mod similar cu analiza componentelor principala,analiza factoriala isi propune sa reexprime variabilitatea continuta in spatiul initial intr-o maniera diferentiata in functie de rolul pe care il au in formarea acesteia factorii comuni, pe deoparte factorii unici pe de alta parte factorii . Vom trata modul in care varianta unei variabile aleatoare poate fi descompusa in componentele relevante dpdvd al interdependentelor cauzale. Prin utilizarea tehnicilor de analiza multidimensionala care au ca scop reducerea dimensionalitatii, variabilitatea spatiului cauzal n dimensional determinat de multimea variabilelor indicator x1,x2,…xn este conservata intr-o proportie mai mare sau mai mica prin intermediul variabilitatii induse de un numar mai redus de factori abstracti f1,f2…fp , p<n care sunt factori comuni. Impreuna cu factorul unic acesti factori determina un spatiu p+1 dimensional numit spatiu test ‘u’.

**Ce sunt scorurile factor, cum se calculeaza si cum se interpreteaza acestea:**

O anumita observatie, corespunzatoare unui factor dat, este determinata sub forma unui scor corespunzator respectivului factor, scor format pe baza contributiei variabilelor originale. Exprimarea generica a scorurilor pentru un anumit factor in functie de variabilele originale este data de urmatoarea relatie:

fi=bi1x1+bi2x2+...+binxn

unde bij reprezinta coeficientii scorurilor factor si sunt elemente ale transpusei matricii factor F. Sub forma matriciala, aceasta relatie poate fi scrisa astfel:

f=F’x

**Metode de estimare a modelului factorial**

AF=analiza de tip multidimen. care are drept obiectiv evidenta corelatiilor, existente la nivelul unei mult de indicat. cu ajutorul unui nr redus de fact.necorelati si ordonati in fct de variabilitatea indusa de ei.In fct de scopul urmarit, avem urmat modalitati: AF exploratorie-acea modalitate de utilize a AF care are drept obiectiv identif unei struct a dependentelor cauzale existente la niv unei mult. de var.indicatori si generarea fact.comuni. AF confirmatorie- acea modalit de utilize a AF in care se urmareste verif unor ip. formulate in legatura cu o struct cauzala deja identificata.

**Recun formelor +exemple.**

TRF reprez ansamblul met,teh, reg de analiza si decizie care au ca scop identif apartenentei obiectelor (entitati informationale) in cadrul unor clase cu individualitate bine definite. Scopul TRF este de a identif la nivelul unor mult. complexe si eterogene de date existent unor clase intr-o forma latenta a caror diferentiere se realizeaza pe baza asemanarilor si deosebirilor eistente intre elem multimilor. TRF este un domeniu multidisciplinar: th sistemelor, th deciz, th.probabilit, th.info., cercetari care pp. utilize unor massive mari de date, cotatii pe piata financiara. Aspectul cel mai important este in domeniul predictilor. Prob gen a TRF consta in anticiparea apartenentei unui obiect la o anumita clasa.

**Principalele concept ale recun formelor**

Forma reprez contextul informational distinct a carui reprezentare se realiz printr-un vector n-dimensional ale carui compon sunt associate caracteristicilor care fac obiectul procesului de recun. Mult ob ce urmeaza a fi clasificate se grup. intr-un spatiu vectorial real- spatiul formelor. Ipoteza fundamental care se face in cadrul TRF este aceea ca la niv spatiului formelor exista un anumit grad de eterogenitate ceea ce implica posibilitatea existentei claselor. Insa la nivelul aceleasi clase se considera ca obiectele sunt caracterizate printr-un anumit grad de omogenitate. Intre clasele identificate avem astfel urmatorele caract: similaritate si disimilaritate. In TRF, urmatorele notiuni sunt echivalente: clasa=grupa=categ=cluster. Clusterul=context informational bine individualizat format din mult formelor care au aceleasi val. pt caract. analizate sau variaza putin si care se deosebesc fundamental de ob. din celelalte clase.

**Prob gen a clasificarii.**

Consta in det. apartenentei unui obiect din Ω ={O1,..On} pop de obiecte, la una din clasele latente existente in pop initiala.1)caract. obiectelor incluse in proc. de clasif. pot fi exprimate sub forma unor var cantitative sau calitative, masurate pe una din cele 4 scale de masurare cunoscute:nominala,ordinala,de interval,de raport. 2)dintre var. incluse in clasif, unele dintre ele au putere de separare mai redusa iar altele mai ridicata, algoritmii de clasif fiind eficienti in sist. in care sunt incluse var care au putere de separare mai ridicata.

Pt cele n obiecte din mult analizata, valorile caracteristicilor in functie de care se face clasif se grupeaza sub forma unui subspatiu real(pt var cantitative) – spatiul formelor. Ob analizate se pot gasi intr-una din m.multe stari potentiale – stari ale naturii. Fie teta=1,k cele k stari ale naturii care pot fi associate grupelor din pop.initiala. Starile nat au 2 propr: exhaustivitatea(cele k stari desriu complet pop analizata) si exclusivitatea reciproca (la niv unui ob nu se pot manifesta 2 stari)c

Identificarea claselor reprez descomp mult Ω (pop de ob) in k clase w1…wk associate celor k stari care pe baza analogiei cu proprietatile de exhaustivitate si exclusivitate au urmat conditii: U(i=1,k)= Ω si wi ∩wj= 0, i#j, I,j=1,k.

**Sistemele de recun controlata si necontrolata**

1.SRC=acele sist in cadrul carora nr de clase este cun iar formele analizate sunt repartizate intr-una din aceste clase. Setul de forme a caror apartenenta la clasa este cun se mai numeste set de referinta/set de prototipuri/set de formare si reprez un esantion extras din pop analizata format din ob a caror apartenenta la clase este cun si care sunt fol in proc de clasif pt det reg de decizie ce det alocarea ob pe clase. Setul de formare contine inf refer la caract ob. dar si la nr de clase constituind baza inferentelor necesare det clasificatorilor. Alg specifici SRC:analiza discriminanta,alg de tip SVM, retele neuronale.

2.SRN=acele sist in care nu se cun inf referitoare la nr de clase si nici apartenenta ob la clase, constructia claselor realizandu-se pe masura parcurgerii mai multor forme,iar det nr de clase se realiz in faza finala a proc de clasif. Exceptie: in cazul sist de recun a formelor exista o categ speciala –alg de partitionare, in care se specifica nr de clase ce urm a fi det. fara a avea certitud. ca este impartirea optima a mult ob. Alg si proced. utilizate in cadrul SR a formelor mai sunt cun sub numele de tehnici de clasificare, clasificare necontrolata sau clusterizare. In functie de modul cum se det nr de clase, teh de clasif se imparti in alg de partitionare (kmeans) si alg de clasif ierarhici.

**Ce este analiza cluster, care sunt conceptele fundamentale ale acesteia si care sunt domeniile utilizarii ei**

***Analiza cluster*** poate fi definita ca reprezentand o multime de principii, metode si algoritmi de clasificare, avand ca scop organizarea datelor sub forma unor structuri informationale semnificative, relevante.

Din punct de vedere concret, efectuarea unei clasificari cu ajutorul metodelor si tehnicilor de analiza cluster consta în obtinerea unor *solutii cluster* sau a unor *partitii*, reprezentate de o multime de clase sau clustere, care verifica proprietatile mentionate anterior.

Analiza cluster este o *analiza explorativa*, de tip multidimensional, care are ca scop gruparea unor entitati informationale, cu natura fizica sau abstracta, în clase sau clustere alcatuite din entitati informtionale cu grad ridicat de similaritate.

Desi folosirea tehnicilor de analiza cluster nu este specifica doar pentru anumite domenii de activitate, totusi, utilizarea cea mai frecventa a acestora este întâlnita în domeniul marketingului, în investigatiile de natura psihosociala sau în evaluarile economico-sociale la nivel teritorial.

**Def scopurile analizei cluster si descrieti tipul informatiilor utilizate in analiza cluster**

Se poate spune ca scopul general al analizelor de tip cluster este acela de creare a asa-numitelor *taxonomii* sau *tipologii*.

Constructia tipologiilor este bazata pe analiza *asemanarilor s*i *deosebirilor* existente între obiectele unei multimi date.

Identificarea pe o mare cantitate de informatii brute a unor categorii, clase sau grupe informationale reprezinta deasemenea unul dintre scopurile generale si, în acelasi timp, principale ale oricarei analize cluster.

Problema cea mai importanta a oricarui tip de analiza cluster este aceea a modului în care poate fi masurat *proximitatea*, respectiv *gradul de apropiere* sau *gradul de departare*, dintre obiecte si dintre clustere.

În general, masurarea gradului de proximitate dintre obiecte se face cu ajutorul a doua grupe de indicatori, cunoscute sub numele de *indicatori de similaritate s*i *indicatori de disimilaritate.*

Cu cât valoarea unui indicator de similaritate este mai mare, cu atât obiectele sau variabilele pentru care acest indicator se evalueaza pot fi considerate a fi mai asemanatoare, respectiv mai apropiate.

*Indicatorii de disimilaritate* sunt marimi numerice care exprima cât de *deosebite* sau cât de *departate* sunt doua obiecte sau doua variabile.

**Definiti analiza cluster si aratati cum se clasifica metodele de analiza cluster**

***Analiza cluster*** poate fi definita ca reprezentand o multime de principii, metode si algoritmi de clasificare, avand ca scop organizarea datelor sub forma unor structuri informationale semnificative, relevante.

Din punct de vedere al naturii lor, al modului de operare si al tipului de solutii pe care le furnizeaza, metodele de analiza cluster pot fi împartite în dou| mari categorii: *metode de tip ierarhic s*i *metode de tip iterativ* sau *de partitionare*.

***Analiza cluster de tip ierarhic* sau *arborescent***este o metoda de clasificare bazata pe gruparea obiectelor pe baza de *agregare succesiva* în clase din ce în ce mai largi de obiecte sau de *dezagregare succesiva* în clase din ce în ce mai mici.

*Metode de clasificare ierarhica prin agregare-*se bazeaza în mod exclusiv pe evaluarea de distante intercluster (Metoda agregarii simple complete, medii, centroidului, Ward)

*Metode de clasificare ierarhica de divizare-* sunt analoage cu metodele aglomerative, cu

deosebirea ca derularea acestora se desfasoara într-o maniera *inversa*.

***Algoritmi de partitionare***

Algoritmii de partitionare includ o serie de metode de analiza cluster, cu mult mai performante decât metodele de clasificareierarhica. Dintre cei mai importanti algoritmi de partitionare, mentinam: algoritmul celor *K*-medii si algoritmul celor *K*-medoizi.

**Definiti conceptual de distanta si descrieti cateva modalitati de evaluare a distantelor dintre forme**

Distanta dintre doua clustere este, de fapt, o distanta dintre doua *multimi* de puncte, adica o distanta mai dificil de evaluat.

Ca distanta între doua *multimi de puncte*, distanaa dintre doua clustere poate fi masurata cu ajutorul uneia dintre mai multe metode posibile.

*Metoda celor mai apropiati vecini* evalueaza distanta dintre doua clustere ca distanta între doua obiecte, unul din primul cluster, iar celalalt din cel de-al doilea cluster, care sunt *cele mai apropiate* între ele în sensul distantei utilizate.

*Metoda celor mai departati vecini* evalueaza distanta dintre doua clustere ca distanta între doua obiecte, unul din primul cluster, iar celalalt din cel de-al doilea cluster, care sunt *cel mai departate* între ele în sensul distantei utilizate.

*Metoda distantei medii dintre perechi* evalueaza distanta dintre dou| clustere ca *medie* a distantelor dintre oricare doua obiecte care apartin celor doua clustere, unul primului cluster, iar celalalt din celui de-al doilea cluster.

*Metoda centroidului* evalueaza distanta dintre doua clustere ca *distanta între centroizii* celor doua clustere.

*Metoda lui Ward* evalueaza distanta dintre doua clustere *suma totaa| a patratelor abaterilor* la nivelul

configuratiei cluster rezultate din comasarea celor doua clustere pentru care se evalueaza distanta.

**Formulati criteriul general al clasificarii si aratati cum se evalueaza variab inter si intra clasa (Cazul uni-demns)**

**Criteriu general de clasificare**:Clasificarea obiectelor în clase se face în asa fel încât sa se asigure o *variabilitate minima în interiorul claselor s*i o *variabilitate maxima între clase.*

Metoda Ward este metoda de evaluare a distantelor dintre clustere care urmareste maximizarea gradului de omogenitate din interiorul clusterelor, ceea ce e echivalent cu minimizarea variabilitatii intraclusterelor. Minimizarea variabilitatii intracluster conduce automat la maximinizarea variabilitatii intercluster.

O problema dificila care apare în analiza cluster, este legata de *necesitatea evaluarii distantelor dintre clase* sau *clustere*. Dificultatea acestei probleme este data de faptul ca distantele dintre clase sau clustere sunt, de fapt, distante între *multimi de obiecte* sau distante între *multimi de variabile*.

Problema evaluarii distantelor dintre clustere apare în special în cazul *analizei cluster de tip ierarhic*, în care construirea arborelui de clustere poate fi facut pe baza *comasarii succesive* sau *divizarii succesive* a clusterelor. Comasarea clusterelor este numita *amalgamare* sau *agregare*, iar divizarea clusterelor este numita *dezagregare*.

Teoretic, procesul de agregare sau dezagregare succesiva a clusterelor se bazeaza pe definirea unei *distante limita* între clustere, distanta numita si *prag de agregare*, respectiv *prag de dezagregare*. În principiu, decizia de comasare a doua clustere sau de divizare a unui cluster este luata numai daca distanta dintre aceste clustere este mai mica, respectiv mai mare decât distanata limita fixata.

**Formulati criteriul general al clasificarii si aratati cum se evalueaza variab inter si intra clasa (Cazul n-dimens)**

**Criteriu general de clasificare**: Clasificarea obiectelor în clase se face în asa fel încât sa se asigure o variabilitateminima în interiorul claselor si o variabilitate maxima între clase.

**Analiza cluster de tip ierarhic**

Analiza cluster de tip ierarhic este o metoda de clasificare bazata pe gruparea obiectelor pe baza de

agregare succesiva în clase din ce în ce mai largi de obiecte sau de dezagregare succesiva în clase din ce în ce mai mici.

Ipoteza de baza a clasificarii ierarhice: În cadrul multimilor de obiecte analizate se diferentiaza o multitudine de

structuri de tip latent, care sunt caracterizate printr-o imbricare de natura arborescenta În cea mai mare parte a lor, algoritmii de clasificare ierahica sunt algoritmi de tip euristic. Exista însa si o categorie aparte de algoritmi de clasificare ierarhica, reprezentata de algoritmii de tip model formal, care genereaza structurile cluster pe baza maximizarii verosimilitatii.

Rezultatul utilizarii analizei cluster de tip ierarhic îl reprezinta o multime de structuri particulare de clustere, numita arbore ierarhic.

structurile cluster de tip ierarhic sunt caracterizate prin nivele diferite de agregare, cuprinse între un nivel minim si un nivel maxim.

Algoritmii de clasificare de tip ierarhic pot fi impartiti in doua catelogii :

• algoritmi de clasificare prin agregare, amalgamare sau combinare;

• algoritmi de clasificare prin dezagregare sau divizare.

**Metoda agregarii simple de analiza cluster**

Reprezinta metoda de clasificare ierarhica de tip ascendant care in fiecare etapa a procesului de clasificare comaseaza 2 dintre clusterele pentru care distanta intre cei mai apropiati vecini e minima in comparative cu alte perechi de clustere.

**Metoda agregarii complete de analiza cluster**

Este de tip ascendant, care in fiecare etapa a clasificarii comaseaza 2 clustere pentru care distanta dintre cei mai indepartati vecini din acele clustere. Clusterizarea de acest tip se mai numeste si analiza cluster de distanta maxima sau analiza cluster de tip MAX.

**Metoda centroidului de analiza cluster**

Este metoda dupa care distanta dintre doua clustere este masurata ca distanta între centroizii celor doua clustere. Centroidul sau centrul de greutate al unui cluster reprezinta obiectul, real sau abstract, ale carui caracteristici au ca valori chiar mediile caracteristicilor obiectelor care compun clusterul respectiv.

**Defini**t**ie**: Metoda centroidului evalueaza distanta dintre doua clustere ca distanta între centroizii celor doua clustere.

Evaluarea distantei dintre doua clustere cu ajutorul metodei centroidului se face calculând mai întâi centroizii celor doua clustere, dupa care se evalueaza distanta dintre clustere ca distanta între acesti centroizi.

**Metoda lui Ward**

Este o metoda de evaluare a distantei dintre doua clustere, care se bazeaza pe maximizarea gradului de omogenitate a clusterelor. De regula, gradul de omogenitate

a unui cluster se considera a fi cu atât mai mare, cu cât suma totala a patratelor abaterilor intracluster este mai mica.

Elementul caracteristic al metodei lui Ward este reprezentat de faptul ca prin comasarea a doua clustere se urmareste obtinerea unei omogenitati maxime la nivelul tuturor clusterelor care apartin unei configuratii date a obiectelor pe clustere

**Definitie:** Metoda lui Ward evalueaza distanta dintre doua clustere suma totala a patratelor abaterilor la nivelul configuratiei cluster rezultate din comasarea celor doua clustere pentru care se evalueaza distanta.

Spre deosebire de alte metode de calcul a distantelor între clustere, distanta Ward ofera o serie de avantaje. Aceste avantaje decurg din faptul ca ea este singura dintre metodele de evaluare a distantelor dintre clustere, care exprima distantele din punct

de vedere al minimizarii variabilitatii intracluster sau, ceea ce înseamna acelasi lucru, din punct de vedere al maximizarii variabilitatii intercluster.

**K-means**

Algoritmul K-means este o metodă de determinare a clusterelor pe care le formează mai multe pattern-uri. Procedura este una de instruire nesupervizată. Se presupune cunoscut numărul K al clusterelor. Fiecare cluster are un centroid. Algoritmul lucrează cu K clustere, deci K dintre punctele folosite la instruire vor fi centriozii celor K clustere. Întrucât initializarea centroizilor se face aleator, există posibilitatea ca mai multe rulări ale algoritmului să conducă la rezultate diferite.

Fiecare punct este asociat clusterului determinat de cel mai apropiat centriod.

Distanta dintre punct şi centriod poate fi calculată, de exemplu, ca distantă euclidiană.

**Dendrograma** Ca rezultat al algoritmului se obţine arborele de clasificare (dendrograma).

Prin secţionarea orizontală a dendrogramei se obţine o partiţie a mulţimii elementelor clasificate. Componentele partiţiei sunt clasele căutate.

Pe axa orizontală sunt elementele iniţiale (ordinea este cea care permite desenarea arborelui). Pe axa verticală sunt distanţele dintre obiecte, de exemplu, între obiectele 4 şi 6 este o distanţă egală cu 4.

**Cum se alege numarul de clustere in cazul clasificarilor de tip ierarhic?**

*Algoritmii* sau *metodele de tip ierarhic* au ca scop producerea *mai multor* soluţii cluster, soluţii numite *ierarhii cluster*.

În cazul metodelor de clasificare ierarhică, *numărul de clustere nu este cunoscut aprioric*.

Algoritmii de clasificare ierarhică furnizează *mai multe soluţii*, de tip *multinivel*, care se numesc *ierarhii cluster* şi care diferă între ele prin numărul de clustere pe care le includ şi prin gradul de agregare al clusterelor.

Cea mai sintetică soluţie a unei structuri cluster obţinute cu ajutorul metodelor de clasificare ierarhică este formată *dintr-un singur cluster*, care include toate obiectele analizate. Cea mai detaliată soluţie a unei structuri cluster de acest fel include unnumăr maxim de clustere, *egal cu numărul de obiecte analizate*, fiecare cluster conţinând un singur obiect. Aceasta înseamnăcă numărul posibil de soluţii dintr-o structură cluster obţinută cu ajutorul algoritmilor ierarhici este mai mic cu unu decâtnumărul de obiecte supuse clasificării. Acest număr este determinat de numărul de *nivele ierarhice* ale solutieisŞi este dat derelaŢia următoare:

NC=T-1

Alegerea dintre cele T-1 soluţii ale unei structuri cluster a celei mai potrivite soluţii cluster rămâne la latitudinea cercetătorului şi se face, în principal, în funcţie de obiectivele urmărite în analiză.

**37.Formulati problema generala a recunoasterii supervizate a formelor si mentionati cateva domenii de activitate.**

În mod frecvent,în analiza datelor apare necesitatea studierii unor populaţii care sunt eterogene d.p.d v. al caracteristicilor analizate,fapt care complică procesul de cunoaştere a acestor populaţii şi impune efectuarea unui demers ştiinţific specific. Expresia cea mai semnificativă a populaţilor de tip eterogen este întâlnită în special în domeniul statisticii, econometriei şi analizei datelor,fiind repr. chiar de cantităţile foarte mari de informaţie care trebuie prelucrată,sintetizată şi interpretată.În cazul cercetării unor populaţii de acest tip,pentru ca rezultatele investigării să capete consistenţă şi relevanţă,este necesară o împărţire, o divizare a acestor populaţii în subpopulaţii cu un anumit grad de omogenitate,urmând ca analizele şi procesul de modelare implicate în studierea respectivei populaţii să se facă în mod diferenţiat, pentru fiecare subpopulaţie în parte.Formularea unor concluzii corecte şi robuste cu privire la manifestarea populaţiilor caracterizate de un grad mai mare sau mai mic de eterogenitate nu este posibilă decât dacă analiza ia în considerare structurarea acestor populaţii pe categorii. în alte situaţii, cum sunt cele în care sunt analizate diverse entităţi economico-sociale, considerate a proveni din populaţii cu caracteristici foarte diferite, există interesul de a identifica,de a recunoaşte, originea acestor entităţi, şi de a obţine o încadrare corectă a acestora în anumite clase Reprezentative pentru populaţia de origine. situaţiile acest fel depăşesc sfera economico financiară, ele întâlnindu-se în mod frecvent într-o mare varietate de alte domenii importante ale ştiinţei, cum ar fi: informatica, biologia,antropologia,medicina, sociologia, geologia, meteorologia etc.

**Ce sunt clasificatorii de tip liniar?**

Un clasificator liniar este o combinaţie liniară de trăsături, care sunt componentele unui vector x. Clasificatorul poate fi scris ca:

g(x) = wT x + w0 = ∑wi xi + w0

Unde

• w este vectorul de ponderi

• w0 este deplasamentul (bias), sau ponderea pragului.

Definiţia unui clasificator

Fie S o mulţime de exemple {S1, S2,…, Sn} ce aparţin unor clase diferite {c1, c2, … , cm}. Fiecare exemplu are d trăsături associate x = {x1, x2,…, xd}. Un clasificator este o corespondenţă între spaţiul trăsăturilor şi etichetele claselor, {c1, c2,…,cm}.

**Definiti functiile discriminant liniare,variabilele discriminant si scorurile discriminant.**

***Functiile discriminant*** (***Fisher***) sunt combinaţii liniare de variabilele descriptor, de forma:

D(x)=B0+Bt\*x

unde x este vectorul variabilelor descriptor, iar B este vector propriu al matricii . Valorile funcţiilor discriminant se

numesc *scoruri discriminan*

***Variabilele discriminant*** sunt combinaţii liniare de variabilele descriptor, de forma:

d=B0+BT\*x

unde x si B au semnificaŢia din definiţia precedentă. Media şi varianţa variabilelor discriminant sunt:

E(d)= B0+Bt\*u

Var(d)=BT\*EPSILON\*B

Odată ce funcţiile discriminat au fost estimate, ele pot fi utilizate pentru efectuarea de predicţii cu privire la apartenenţa unor noi obiecte la clasele de predicţie.*t*.

Variabilele discriminant d1,d2…dp determină un nou spaţiu p-dimensional, numit *spaţiu discriminant*, ale cărui axe sunt reprezentate de vectorii B(i) Si în contextul căruia se va face clasificarea efectivă obiectelor. Valorile variabilelor discriminant di sunt rezultatul evaluării funcţiilor discriminate Di pentru un anumit obiect fixat şi se numesc *scoruri discriminat.* Scorurilediscriminant sunt utilizate ca *indicatori* în clasificarea propriu-zisă a obiectelor.

**Descrieti clasif. Bayesian si aratati cum poate fi utilizat in predictia apartenentei formelor.**

Principala probl care trebuie rezolvata in cadrul recunoasterii supervizate a formelor este aceea a determinarii criteriilor (regulilor) de clasificare astfel incat noi forme a caror apartenenta nu este cunoscuta sa poata fi incadrate in clase in care acestea se grupeaza in mod natural.

Regulile de clasificare sunt numite generic clasificatori in teoria recunoasterii formelor.

Determinarea clasificatorilor au o natura algoritmica, numindu-se proces de formare a clasificatorului.

In cadrul discriminarii mai multor obiecte, se poate utiliza una dintre abordarile:

**Clasif de cost minim** (Bayes)

**Clasif bazat pe distanta Mahalanobis** (clasif metric)

**Clasif bazat pe raport de verosimilitate** (Neyman Pearson)

**Clasif bazat pe entropie**

Ω-colectivitatea initiala a obiectelor, care se structureaza in k clase: ω1, ω2,.. ωk.

Colecticitatea initiala genereaza spatiul formelor R Ω.

In mod similar, cele k clase din Ω genereaza in spatiul formelor k regiuni de decizie (R1,R2,....Rk)

Problema recunoasterii formelor poate fi definita ca fiind echivalenta cu identificarea regiunilor de decizie in spatiul formelor in functie de care se grupeaza obiectele din Ω.

– regiune de decizie estimata

Geometric, determinare regiunilor de decizie estimate (este similara cu determinarea regiunilor de separare care sunt det pt acele forme pt care se verifica o relatie functionala de tipul Dij(x) = 0

Identificarea regiunilor de decizie se realizeaza in cele mai multe cazuri prin minimizarea costului total al clasificarii din Ω, adica minimizarea erorii totale a clasificarii.

**Descrieti clasificatorul liniar fisher si aratati cum poate fi utilizat acesta in predictia apartenetei formelor.**

Prima modalitate de abordare a problemelor de clasif. cu ajutorul tehnicilor de analiza discriminanta dateaza din anul 1933 si a fost propusa de Fisher.

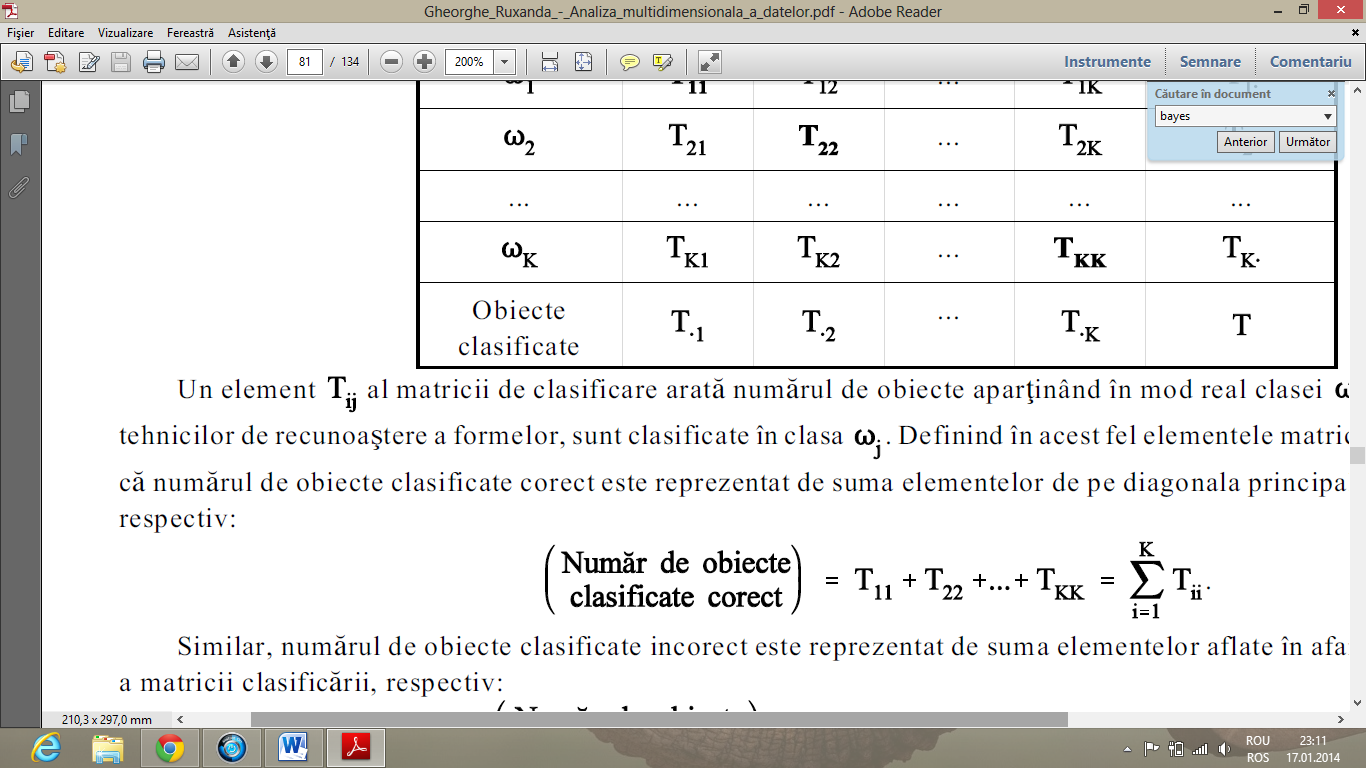
Metoda de analiza discriminanta propusa de Fisher este o metoda parametrica care nu necesita evaluarea distrib de probabilitate ale claselor ci numai estimatii ale unor parametrii la nivelul populatiei analizate (media, variatia, covarianta). Potrivit acestor metode de solutionare ale probl de clasificare supervizata,multimea initiala de obiecte Ω este impartita in ω1, ω2,.. ωk pe baza comparatiei dintre variabilitatea intraclasa si variabilitatea interclasa, inregistrate in cadrul claselor din populatia analizata.Analiza discriminanta pp. det. functiei discriminant care sa asigure separabilitatea cat mai buna a claselor si fiecare clasa sa fie caract. printr-un grad de omogenitate cat mai mare, echivalent cu variab intraclasa cat mai mica.Criteriul lui Fisher pt analiza discriminanta este caract. prin robustete si simplitate, informatiile furnizate fiind utilizate pt identificarea evolutiei fenomenelor analizate.In cadrul analizei discriminant de tip Fisher se urmareste det.a unei noi axe sau directii, avand la baza urm. criteriu mixt :maximizare variabilitatii interclasa si minimizarea variab intraclasa.

**Descrieti modul de stabilire a abilitatii predictive a unui clasificator si matricea corectitudinii clasificarii.**

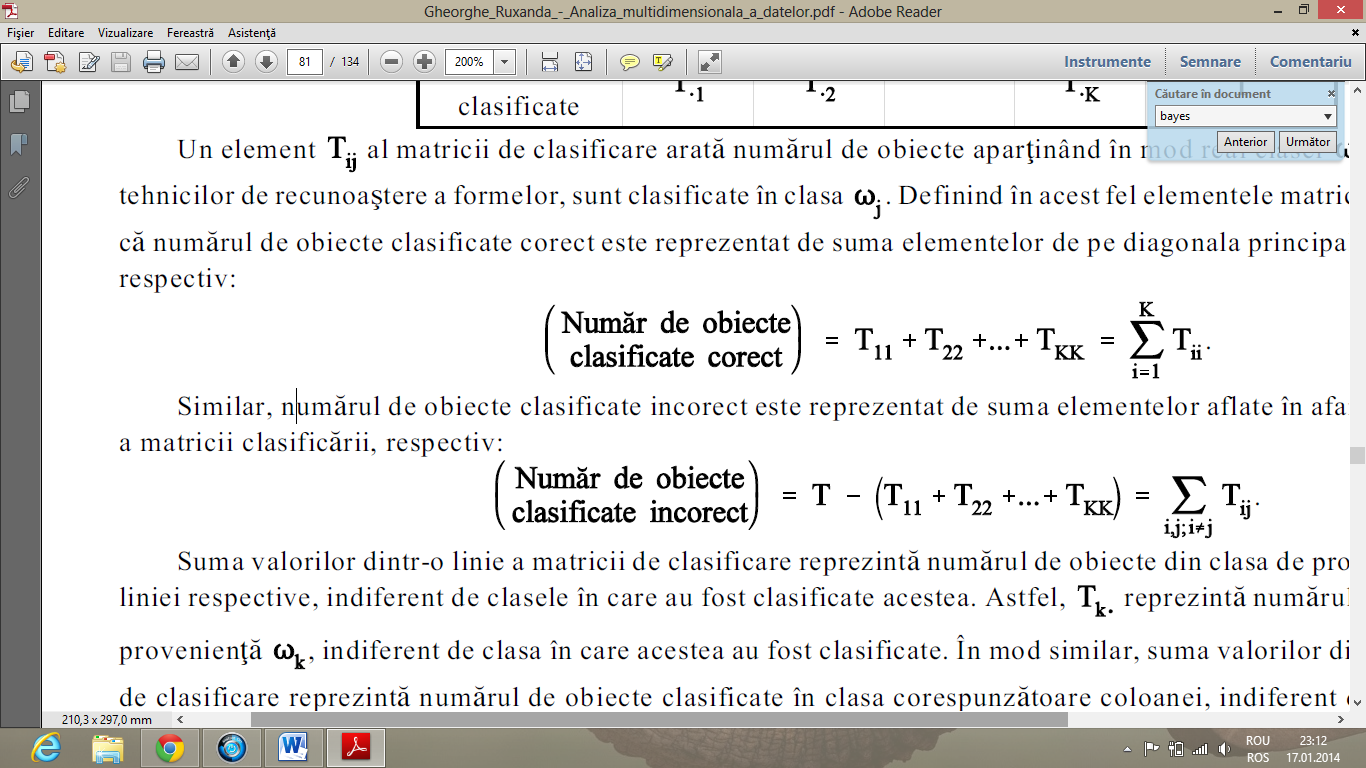
Modul în care un clasificator asigura clasif. obiectelor cu apartenenta cunoscuta poate fi descris prin intermediul unei matrici, numita matricea corectitudinii clasificarii sau, mai simplu, matricea clasificarii, care contine informatiile necesare pentru a aprecia corectitudinea clasificarii obiectelor.

Daca vom considera un esantion format din T obiecte,care apartin claselor ω1, ω2,... Ωk,atunci va rezulta o matricea de clasificare.Un element al matricii de clasif.arata nr. de obiecte apartinând în mod real clasei ωi care, prin utiliz.

tehnicilor de recunoastere a formelor, sunt clasif. în clasa ωj.Definind în acest fel elem. matricii de clasificare,rezulta ca nr. de obiecte clasificate corect este repr. de suma elem. de pe diagonala principala a matricii clasificarii, respectiv:



Similar,nr de obiecte clasif. incorect este repr. de suma elemen. aflate în afara diagonalei principale a matricii clasificarii :



Suma valorilor dintr-o linie a matricii de clasif. repr. nr. de obiecte din clasa de provenienta ce coresp. liniei respective, indiferent de clasele în care au fost clasif. acestea.Astfel, repr. nr. de ob. din clasa de provenienta,indiferent de clasa în care acestea au fost clasif.În mod similar,suma valorilor dintr-o coloana a matricii de clasif. repr.nr. de ob. clasif.în clasa coresp. coloanei,indiferent de clasa de provenienta a obiectelor. Rezulta ca repr. nr. de obiecte clasificate în clasa,indiferent de

clasa de provenienta a acestora.

Populaţia sau colectivitatea generală este reprezentată de mulţimea tuturor măsurătorilor efective sau conceptuale care prezintă interes pentru cercetător sau experimentator.

Eşantionul reprezintă o submulţime de măsurători selectate dintr-o populaţie, o submulţime a populaţiei statistice supusă investigaţiei ştiinţifice.

Variabila reprezintă o abstractizare a mulţimii de valori posibile pe care le poate înregistra o caracteristică a unui anumit fenomen. După natura pe care o au, variabilele pot fi de două tipuri: variabile calitative şi variabile cantitative.

**Scala nominală** este o scală non-metrică, pe baza căreia valorile variabilelor sunt definite prin intermediul simbolurilor nenumerice. Măsurarea variabilelor pe scala nominală este echivalentă cu procesul de codificare a variabilelor.

Scala ordinală este o scală non-metrică, prin intermediul căreia valorilor posibile ale caracteristicilor li se atribuie numere de ordine sau ranguri, în funcţie de poziţia acestor valori într-o ierarhie

Scala raport este o scală metrică, prin intermediul căreia valorilor posibile pe care le pot lua caracteristicile măsurate li se atribuie numere definite in raport cu o origine prestabilită.

**Tendinta centrala** poate fi evidentiata prin intermediul unor indicatori statistici, între care cei mai importanti sunt: *media, mediana* si *modulul*. Fiecare dintre acesti indicatori exprima, într-un fel sau altul, mai mult sau mai putin sugestiv, nivelulcaracteristicii analizate de-a lungul obiectelor.

***Media*** - se obţine împărţind suma valorilor individuale la efectivul populaţiei sau al eşantionului

***Mediana*** - este valoarea care, în cadrul seriei statistice, separă efectivul populaţiei în două părţi egale. Nu are o formulă atît de simplă cum este cea a mediei; mai mult, o valoare mediană propriu-zisă nu există decît dacă numărul n este fără soţ, cînd există, de fapt, un individ mijlociu (al [n+1]/2 lea) a cărui valoare este mediana. Dacă n este par, se iau indivizii de rang n/2 şi n/2 + 1, cu valori, să zicem, Xi şi xi+1, iar mediana poate fi orice valoare din intervalul (Xi,Xi+1); de regulă se ia media aritmetică a celor două valori.

***Modulu****l* - se utilizează numai cînd se lucrează cu frecvenţe, fiind valoarea luată cu cea mai mare frecvenţă. Se poate vorbi şi de valori modale relative atunci cînd frecvenţele mai multor clase, neînvecinate, le întrec pe cele din imediata lor vecinătate; avem de a face atunci cu serii bimodale (cu două moduri) sau plurimodale.

***Varianta simpla*** este o masura pt devierea de la medie, deviatie standard fata de medie. **Vs=Σi=1n  (xi- xmediu)2** . Varianta totala masoara variabilitateace caracterizeaza observatiile unei multimi de variabile si se defineste ca suma a variantelor individuale ale variabilelor: **Vt=ΣSi2**

Varianta totala ofera o imagine cuprinzatoare asupra variabilitatii globale ce carac observatiile var analizate, ca masoara aceasta variabilitate doar in sens individual, neluand in considerare variabilitatea comuna, simultana a observatiilor, adica variabilitatea generalizata.

Varianta generalizata masoara variabilitatea ce carac observatiile multimii de variabile, atat din pct de vedere individual, cat si din pct de vedere al simultanietatii, al interaxtivitatii informationale variabilele.

**Covarianta** este o masura a variatiei simultane a doua variabile, ea fiind, in valoare absoluta, cu atat mai mare cu cat valorile absolute ale variatiilor celor doua variabile in jurul mediei sunt mai apropiate ca magnitudine, evidentiind o anumita proportionalitate pe multimea subiectilor studiati.

**Definiti conceptul de distanta si descrieti cateva modalitati de evaluare a distantelor dintre forme**

*Indicatorii de disimilaritate* sunt marimi numerice care exprima cât de *deosebite* sau cât de *departate* sunt doua obiecte sau doua variabile. Indicatorii de disimilaritate se mai numesc si indicatori sau coeficienti de *deosebire* sau de *distantare* a obiectelor sau variabilelor. Cu cât valoarea unui indicator de disimilaritate este mai mare, cu atât cele doua obiecte sau cele doua variabile pentru care se calculeaza sunt mai diferite, adica mai distantate între ele. Cea mai importanta si cea mai utilizata categorie de indicatori de disimilaritate este reprezentata de indicatorii de tip *distanta*.

Prin natura lor numerica, variabilele de tip cantitativ, adica variabilele masurate pe scalele de tip raport, interval si, eventual, ordinal, permit o definire mai naturala a conceptului de distanta. Pentru variabilele de tip nominal, inclusiv variabilele de tip binar, distantele se calculeaza într-un mod specific, compatibil cu natura acestor variabile. Pentru evaluarea disimilaritatilor dintre obiectele ale caror caracteristici sunt de tip cantitativ sau dintre variabile de tip cantitativ, pot fi folosite mai multe tipuri de distante, cum ar fi: distanta *Euclidiana* (simpla, ponderata sau patrata), distanta *Manhattan*, distanta *Cebîsev*, distanta *Minkovski*, distanta *Camberra*, distanta *Mahalanobis,* distanta *Pearson*, distanta *Jambu* etc.

***Distan****t****a Euclidian****a*

Distanta *Euclidiana*, care mai este cunoscuta si sub numele de *norma de tip ,* este distanta cea mai frecvent utilizata în problemele de analiza cluster. Ea se calculeaza ca *radacina patrata a sumei patratelor diferentelor coordonatelor* celor doua obiecte sau variabile pentru care se evalueaza distansa.

**Distanta *Manhattan***, numita si distanta *rectangulara*, distanta “C*ity-Block*“ sau *norma de tip* , se calculeaza ca suma a valorilor absolute ale diferentelor coordonatelor celor doua obiecte sau celor doua variabile analizate Deoarece diferentele de coordonate utilizate în calculul sau nu sunt *amplificate* printr-o ridicare la o putere, distanta Manhattan este mai robusta în raport cu prezenta în date a valorilor aberante. Distanya Manhattan poate fi calculata si în varianta *ponderata*, calculul facându-se în mod similar cu cel al distantei Euclidiene ponderate. De asemenea, distanta Manhattan poate fi utilizata în cazul în care obiectele au caracteristici care sunt masurate pe scala de tip interval si pe scala de tip raport.

***Distanta Cebî****s****ev***

Distanta *Cebîsev*, cunoscuta si sub numele de “*maxim al dimensiunilor*“ sau *norma de tip* , este o distanta de tip *valoare absoluta* si se determina ca fiind valoarea maxima a valorilor absolute ale diferentelor dintre coordonatele obiectelor sauvariabilelor.Distanta Cebîsev poate fi utilizata atunci când se doreste ca doua obiecte sau variabile sa apara ca fiind diferite, daca eledifera chiar si doar din punct de vedere al unei caracteristici, respectiv al unui obiect.

***Distanta Mahalanobis***

Distanta *Mahalanobis* este una dintre cele mai cunoscute, mai importante si mai frecvent utilizate distante. Ea este o forma generalizata a conceptului de distanta. Distanta Mahalanobis reprezinta singurul tip de distanta care ia în considerare, într-o maniera completa, *gradul de* *dispersare* al multimii de obiecte sau al multimii de variabile analizate, precum si *gradul de corelare* al respectivelor entitati informationale. Utilizarea distantei Mahalanobis este recomandata, mai ales în situatiile în care variabilele care descriu obiectele sunt corelate între ele. Distanta Mahalanobis este utilizata si în cazul tehnicilor de clasificare controlata, pe baza acestei distante fiind dezvoltat chiar un criteriu operational de discriminare