Supervised and Unsupervised Classification for Pattern Recognition Purposes

Conf.dr. Cătălina-Lucia COCIANU
Catedra de Informatică Economică, ASE Bucureşti

A cluster analysis task has to identify the grouping trends of data, to decide on the sound clusters as well as to validate somehow the resulted structure. The identification of the grouping tendency existing in a data collection assumes the selection of a framework stated in terms of a mathematical model allowing to express the similarity degree between couples of particular objects, quasi-metrics expressing the similarity between an object an a cluster and between clusters, respectively. In supervised classification, we are provided with a collection of preclassified patterns, and the problem is to label a newly encountered pattern. Typically, the given training patterns are used to learn the descriptions of classes which in turn are used to label a new pattern. The final section of the paper presents a new methodology for supervised learning based on PCA. The classes are represented in the measurement/feature space by a continuous repartitions
Keywords*: clustering, supervised classification, pattern recognition, dissimilarity measures, PCA (principal component analysis).*

1 Generalităţi privind clasificarea formelor

Tehnicile de clusterizarea datelor fac parte din domeniul clasificării nesupervizate. Clasificarea supervizată (sau analiza discriminantă) presupune existenţa unei colecţii de forme preclasificate şi etichetarea unei forme neasignate încă nici unei clase. În general, formele deja clasificate (numite şi forme de antrenament) sunt utilizate la învăţarea descrierii claselor care, la rândul lor, sunt folosite la etichetarea unui nou exemplu. În cazul clasificării nesupervizate, problema de rezolvat este de a clasifica o mulţime de forme dată într-un set de clase "semnificative", pe baza unui criteriu de conformitate specificat. Într-un anumit sens, etichetele sunt asociate claselor, dar aceste categorii de etichete sunt esenţial determinate de datele observate (etichetarea este obţinută exclusiv din analizarea datelor disponibile).

În cazul majorităţii problemelor formulate în cazul clusterizării, informaţia disponibilă despre datele observate este în general neglijabilă. Metodologiile de clusterizare sunt în particular utile în determinarea interconexiunilor dintre datele observate, precum şi a structurii acestora.

În general, clusterizarea formelor presupune efectuarea următoarelor operaţii primitive asupra datelor (Jain, Murty, Flynn, 1999):

1. reprezentarea formelor (eventual extragerea caracteristicilor şi/sau selecţia caracteristicilor)
2. definirea unei măsuri de proximitate pe domeniul datelor observate
3. clusterizarea (gruparea) datelor
4. abstractizarea datelor (dacă este necesară)
5. estimarea rezultatelor (dacă este necesară).

În abordarea statică, validarea este realizată prin testarea unor ipoteze statistice. Există trei tipuri de studii asupra validării sistemului de clase obţinut. *Estimarea externă* a validităţii compară structura obţinută cu o structură a priori. *Examinarea internă* a validităţii sistemului de clase rezultat presupune verificarea faptului că structura este potrivită intrinsec datelor observate. Cea de-a treia abordare corespunde *testului relativ,* care compară două structuri şi măsoară meritul relativ al fiecăreia dintre ele (Dubes, 1993).

O formă (sau vector de caracteristici, observaţie sau dată) **x** este un vector de dimensiune *d*,

$\mathbf{x} = (x_1, x_2, ..., x_d)$.

Pentru fiecare $1 \le i \le d$, elementul scalar x_i din forma **x** este numit caracteristică (sau atribut) a lui **x**. O mulțime de forme este notată în continuare cu $\aleph$, unde
$\aleph = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$.
Cea de-a i-a formă din $\aleph$ este notată cu $\mathbf{x}_i = (x_{i,1}, x_{i,2}, ..., x_{i,d})$.
O clasă abstractă se referă la starea caracteristicii care guvernează procesul de generare a formelor. Concret, o clasă *C* poate fi privită ca o sursă de date a cărei distribuție în spațiul caracteristicilor este guvernată de o densitate de probabilitate specifică clasei *C* . Tehnicile de clusterizare revin la gruparea formelor astfel încât clasele astfel obținute să reflecte diferitele procese de generare a formelor reprezentate în mulțimea formelor $\aleph$.
Tehnicile de clusterizare "hard" asociază fiecărei forme $\mathbf{x}_i$ o unică etichetă l_i care identifică o anumită clasă. Mulțimea tuturor etichetelor asignate formelor din $\aleph$ este notată cu $L = \{l_1, l_2, ..., l_n\}$. Pentru orice $1 \le i \le n$, $l_i \in \{1,2,...,k\}$, unde *k* este numărul de clase.
Procedurile de clusterizare fuzzy asociază fiecărei forme de intrare $\mathbf{x}_i$ o probabilitate de apartenență la clasa $j, f_{i,j}$,
$1 \le i \le n, 1 \le j \le k, \sum_j f_{i,j} = 1$.
Convențional, formele sunt reprezentate prin vectori multidimensionali, unde fiecare dimensiune corespunde unei caracteristici (atribut). Caracteristicile pot fi
1. cantitative: valori continue (de exemplu ponderi); valori discrete; valori de tip interval (de exemplu durata unui eveniment).
2. calitative: nominale sau neordonate (de exemplu culori); ordinale.
O altă modalitate de structurare a caracteristicilor este reprezentarea arborescentă, în care un nod parental reprezintă o generalizare a nodurilor descendenți direcți. (Jain, Murty, Flynn, 1999). Reprezentarea generalizată a formelor, numite în acest caz obiecte simbolice. a fost propusă de Diday (1988). Obiectele simbolice sunt definite prin conjuncții logice de evenimente.
În multe situații practice este utilă izolarea caracteristicilor celor mai descriptive ale mulțimii de intrare şi utilizarea exclusiv a acestora în analizele ulterioare. Tehnicile de selecție a caracteristicilor identifică o submulțime a caracteristicilor existente, care va fi utilizată în analiza ulterioară, în timp ce tehnicile de extragere a caracteristicilor presupun calculul unor noi caracteristici din mulțimea inițială de forme observate. În ambele situații, scopul este acela de a îmbunătăți fie performanța clasificatorului, fie eficiența calculului.
Disimilaritatea dintre două forme din mulțimea $\aleph$ este exprimată prin intermediul unei distanțe definite pe spațiul formelor. În continuare vom presupune ca atributele formelor sunt definite pe spații continue. Cea mai utilizată metrică pentru caracteristici definite pe spații continue este distanța euclidiană,
(1.1) $d_2(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$,
unde $1 \le i, j \le n$.
Distanța definită prin (1.1) este un caz particular al distanței Minkowski,
(1.2) $d_p(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_p$,
unde $1 \le i, j \le n$.
Distanța euclidiană este de obicei utilizată în evaluarea distanței dintre două obiecte în spații bidimensionale sau tridimensionale; de asemenea, algoritmii de clusterizare ce folosesc distanța euclidiană sunt aplicați mulțimilor de date care sunt partiționate în clase compacte, "izolate" (Mao, Jain, 96). Inconvenientul folosirii directe a metricii Minkowski este tendința de „dominare" a caracteristicilor cu valori absolute mari.
Măsurile de tip distanță pot fi de asemenea influențate de corelațiile lineare dintre caracteristicile formelor. Acest tip de distorsiune poate fi eliminată fie prin aplicarea unei transformări asupra setului de date observate pentru obținerea unei selecții de medie vectorul nul şi matrice de covarianță matricea unitate (procesul de "albire" a datelor), fie prin utilizarea distanței Mahalanobis. Pentru $1 \le i, j \le n$, distanța Mahalanobis dintre formele $\mathbf{x}_i$ şi $\mathbf{x}_j$ este definită prin,
(1.3) $d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)^T$,
unde $\boldsymbol{\Sigma}$ este fie matricea de covarianță de selecție (calculată exclusiv pe baza formelor disponibile din mulțimea $\aleph$), fie matricea de

covarianță teoretică a procesului care a generat formele observate. Implicit se presupune că densitățile condiționale ale claselor sunt unimodale şi gaussiene. O serie de algoritmi de clusterizare utilizează matrice de vecinătăți în locul mulțimii inițiale de forme $\aleph$. În situația în care o parte a caracteristicilor formelor nu sunt definite pe spații continue este posibil ca atributele respective să nu fie comparabile. Noțiunea de vecinătate este definită în cazul atributelor cu valori nominale binar. În practică au fost însă definite diferite măsuri de vecinătate pentru tipuri eterogene de caracteristici. De exemplu, Wilson şi Martinez (1997) au propus o combinație dintre o metrică de tip Minkowski pentru caracteristici definite pe spații continue (caracteristicile cantitative) şi o distanță bazată pe numărare pentru atributele nominale (caracteristicile calitative).

De asemenea, în literatura de specialitate sunt definite o serie de distanțe care înglobează informații referitoare la influența vecinătăților asupra formelor comparate. Punctele din vecinătatea unei forme sunt numite context. Pe baza contextului, similaritatea dintre două puncte $\mathbf{x}_i$ şi $\mathbf{x}_j$ este definită prin, (Jain, Murty, Flynn, 1999)

(1.4) $s(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i, \mathbf{x}_j, \mathcal{E})$,

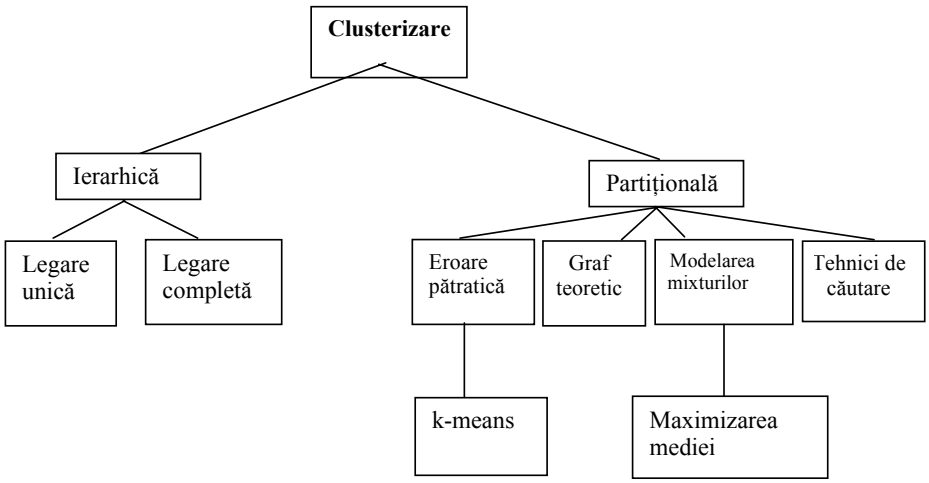
unde $\mathcal{E}$ este contextul în care este calculată distanța (mulțimea punctelor vecine).

O altă abordare a măsurării similarității formelor este bazată pe context şi pe un set de concepte predefinite. De exemplu, în cazul clasificării conceptuale, similaritatea dintre formele $\mathbf{x}_i$ şi $\mathbf{x}_j$ este definită prin,

(1.5) $s(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i, \mathbf{x}_j, \mathcal{E}, \mathcal{C})$,

unde $\mathcal{E}$ este contextul în care este calculată distanța şi $\mathcal{C}$ este mulțimea conceptelor predefinite.

Metodele de clusterizarea a formelor dezvoltate până în prezent pot fi prezentate ierarhic prin figura 1. (Jain, Murty, Flynn, 1999).



2. Clusterizarea ierarhică a formelor

Algoritmii de clusterizare de tip ierarhic au ca rezultat o *dendrogramă*, adică un arbore care reprezintă grupările de forme la diferite niveluri de similaritate. Diferența dintre două niveluri de similaritate este reflectată de modificarea claselor.

Reprezentarea ierarhică rezultată este caracterizată astfel. Clasele de pe fiecare nivel al ierarhiei sunt create prin reunirea claselor situate pe nivelul imediat inferior. Nivelul inferior corespunde claselor ce conțin câte o singură observație, în timp ce nivelul superior este asociat unei singure clase ce conține toate datele. Startegiile de clusterizare ierarhică sunt de tip *aglomerare* (bottom-up), respectiv de tip *divizare* (top-down). (Hastie, Tibshirani şi Friedman, 2001)

Tehnicile de tip aglomerare au ca punct de plecare nivelul inferior şi, la fiecare etapă, sunt reunite recursiv în câte un cluster o pereche selectată de clase. Este obținută astfel gruparea de la nivelul imediat superior. Perechea de clustere selectate pentru reunire este

astfel încât disimilaritatea inter-clase să fie minimă. Metodele de tip divizare consideră inițial nivelul superior şi, la fiecare etapă, împart recursiv un cluster de la nivelul curent în două noi clase. Procesul de divizare a unei clase este realizat astfel încât grupările rezultate să maximizeze disimilaritatea inter-clase. În ambele variante de implementare a clasificării ierarhice, fiecare nivel al reprezentării constituie o grupare particulară a datelor în clase care nu au forme comune. Întreaga ierarhie reprezintă o secvență ordonată de astfel de grupări. În continuare alegerea nivelului corespunzător clasificării "naturale" a datelor observate (în sensul că observațiile fiecărui grup sunt suficient de similare între ele şi diferă suficient de mult de formele asignate celorlalte clase) cade în sarcina utilizatorului reprezentării arborescente (Hastie, Tibshirani şi Friedman, 2001).

Clasificările ierarhice de tip aglomerare sunt implementate în variantele legare unică, legare completă şi algoritmul de minimizare a varianței. Cele mai utilizate metode sunt legarea unică şi legarea completă; în aplicații, algoritmul de tip legare completă s-a dovedit a produce ierarhii mai utile decât cele obținute prin aplicare algoritmului legare unică. (Jain şi Dubes, 1988).

În cazul legării unice, distanța dintre două clase este calculată prin minimul distanțelor dintre fiecare două forme, fiecare situată în câte una din cele două clase. În cadrul algoritmului legare completă, distanța dintre două clase este maximul distanțelor dintre câte o formă dintr-o clasă şi o formă din cealaltă clasă. În ambele variante, două clase sunt reunite pe baza criteriului distanței minime. (Jain, Murty, Flynn, 1999)

Algoritmul de clusterizare de tip aglomerare prin legare unică

Etapele algoritmului sunt următoarele.

(1) Contruieşte nivelul inferior, format din clase cu câte o singură formă, precum şi o listă, L, a distanțelor dintre oricare două forme distincte, neordonate. Sortează crescător lista obținută.

(2) Parcurge L şi, pentru fiecare disimilaritate distinctă d_k, construieşte un graf în care fiecare două forme situate la o distanță mai mică decât d_k sunt conectate printr-o muchie. Dacă toate formele sunt membre în graful conectat, stop. Altfel, repetă (2).

(3) Rezultatul algoritmului este reprezentarea ierarhică formată din grafuri şi care poate fi "secționată" la nivelul de disimilaritate dorit, formând o partiție identificată prin componentele conectate unic în graful corespunzător.

Algoritmul de clusterizare de tip aglomerare prin legare completă

Etapele algoritmului sunt următoarele.

(1) Contruieşte nivelul inferior, format din clase cu câte o singură formă, precum şi o listă, L, a distanțelor dintre oricare două forme distincte, neordonate. Sortează crescător lista obținută.

(2) Parcurge L şi, pentru fiecare disimilaritate distinctă d_k, construieşte un graf în care fiecare două forme situate la o distanță mai mică decât d_k sunt conectate printr-o muchie. Dacă toate formele sunt membre în graful complet conectat, stop.

(3) Rezultatul algoritmului este reprezentarea ierarhică formată din grafuri şi care poate fi "secționată" la nivelul de disimilaritate dorit, formând o partiție identificată prin componentele complet conectate în graful corespunzător.

3. Clusterizarea partițională

Algoritmii de clusterizare de tip partiție sunt utilizați în general în cazul selecțiilor de volum mare, pentru care construirea unei dendrograme necesită un timp de calcul foarte mare. Metodele de tip partiție realizează clusterizarea formelor din $\aleph$ prin optimizarea unei funcții criteriu, definită local, pe o submulțime a datelor observate, respectiv global, pe mulțimea $\aleph$.

Dintre metodele de clusterizare de tip partiție, tratăm în continuare cele mai des utilizate, şi anume,

- clusterizare în contextul algorimtului EM (maximizarea mediilor) pentru estimarea parametrilor unei mixturi de distribuții normale;
- algoritmi de tip eroare pătratică;
- clusterizare fuzzy;

- algoritmi de clusterizare în abordarea evoluționistă.

3.1. Clusterizarea în contextul algoritmului EM

Teorema EM. Algoritmul de maximizare a mediei

Teorema EM este bazată pe inegalitatea Jensen: pentru orice distribuții discrete de probabilitate, $p(x)$ și $q(x)$,

(3.1) $\sum_x p(x)\log_b p(x) \geq \sum_x p(x)\log_b q(x)$.

Fie y data observabilă, $P_{\theta'}(y)$ distribuția de probabilitate a lui y în cadrul unui model caracterizat de parametrii θ' (θ' reprezintă totalitatea parametrilor ale căror valori intervin în specificarea distribuției $P_{\theta'}$) și $P_\theta(y)$ distribuția de probabilitate corespunzătoare setului de parametri θ. Problema este de a determina condițiile în care apariția datei y este mai probabilă în modelul caracterizat de θ comparativ cu modelul specificat prin θ' (în acest caz, θ este o îmbunătățire a lui θ').

Teorema EM (maximizarea mediei)

Dacă

(3.2) $\sum_t P_{\theta'}(t \mid y)\log_b P_\theta(t,y) > \sum_t P_{\theta'}(t \mid y)\log_b P_{\theta'}(t,y)$

rezultă (3.3) $P_\theta(y) > P_{\theta'}(y)$ (Jelinek, 1997)

Teorema de maximizare a mediei are următoarea interpretare. Dacă inițial parametrii modelului sunt setați pe θ' și sunt selectați parametrii θ astfel încât relația (3.2) să fie îndeplinită, atunci data y este observată cu o probabilitate mai mare în modelul caracterizat de parametrii θ comparativ cu modelul specificat prin θ'.

Pe baza teoremei EM este obținut algoritmul de maximizare a mediei, astfel. (Hastie, Tibshirani și Friedman, 2001)

Pas1. Selectează θ', valoarile inițiale ale parametrilor modelului.

Repetă

Pas2. Calculează θ care maximizează membrul stâng al inegalității (3.2)

Pas3. $\theta' \leftarrow \theta$

cât timp *C*.

Condiția *C* este prestabilită și asigură încheierea calculului; specificarea condiției *C* poate fi realizată utilizând următoarea observație. Pe măsură ce algoritmul calculează noi valori ale parametrilor θ', probabilitatea $P_{\theta'}(y)$ crește până la o anumită valoare, datorită faptului că $P_{\theta'}(y) \leq 1$.

Tehnicile de clusterizare în contextul algoritmului EM reprezintă abordarea de tip modelarea mixturilor. În cadrul acestei abordări, ipoteza de lucru este aceea că formele de clasificat provin din diverse distribuții, scopul fiind acela de a identifica parametrii fiecărei distribuții și, eventual, numărul distribuțiilor din care provin formele. În general, au fost dezvoltați algoritmi de clusterizare EM în cazul parametrilor densităților de tip Gauss.

În contextul EM, a fost propusă următoarea abordare a clasificării formelor unei mulțimi date $\aleph$ (Mitchell, 1997). Parametrii densităților de repartiție componente ale mixturii, cât și parametrii de mixare sunt necunoscuți și vor fi estimați din datele observate. La momentul inițial, procedura EM dispune de o anumită estimație a vectorului parametru. Iterativ, algoritmul recalculează scorul formelor în densitatea mixtură rezultată din vectorul parametru. Cu noile scoruri astfel calculate este modificat vectorul parametru de estimat. În contextul clasificării, scorul asociat fiecărei forme (care măsoară, în esență, probabilitatea ca fiecare formă să provină dintr-o componentă particulară a mixturii) este similar unei indiciu asupra clasei din care provine forma. Toate acele forme care, prin scorul lor, sunt plasate într-o componentă particulară a mixturii sunt grupate în aceeași clasă.

3.2. Algoritmi de tip eroare pătratică

Una dintre cele mai utilizate funcții criteriu pentru clusterizare este eroare pătratică. Situațiile în care este recomandată aplicarea acestui criteriu sunt cele în care clasele sunt izolate și compacte.

Fie $\aleph$ mulțimea formelor de clasificat, *K* numărul de clase și $\Im$ un cluster. Eroarea pătratică asociată clusterizării $\Im$ (sau gradul de împrăștiere a punctelor în interiorul clasei $\Im$) este,

(3.4) $e^2(\aleph,\Im)=\sum_{j=1}^{K}\sum_{i=1}^{n_j}\left\|\mathbf{x}_i^{(j)}-\mathbf{c}_j\right\|^2$,

unde $\mathbf{x}_i^{(j)}$ este cea de-a *i*-a formă din clasa *j*, iar $\mathbf{c}_j$ este centroidul clasei *j*.

Forma generală a unui algoritm de clusterizare de tip eroare pătratică poate fi descrisă astfel.

Inițializare. Selectează o partiție inițială a formelor din $\aleph$, cu un număr fixat de clustere, fiecare clasă având un centru dat

Faza 1.

Repetă

Asociază fiecare formă **x** clusterului cu centroidul cel mai apropiat de **x** şi recalculează centrele claselor astfel modificate.

până când structura clusterelor se stabilizează (nu au loc alte reasignări ale formelor)

Faza 2.

Unește şi respectiv împarte clusterele astfel obținute pe baza unor informații de tip euristic şi, opțional, reia Faza 1.

Aloritmul k-means

Cel mai utilizat algoritm din această clasă este algoritmul *k*-means. Fie *K* numărul de clase, fiecare formă fiind asignată unic unei clase prin etichetarea ei cu un număr *k*, $1\le k\le K$. Procesul de asignare a formelor poate fi caracterizat de o mapare de tipul mai-mulți la unul, definit printr-un codor $k=C(i)$, care asociază cea de-a *i*-a observație din mulțimea $\aleph$ clasei *k*. Disimilaritatea dintre două forme este măsurată în general în termenii distanței euclidiene.

Utilizând aceste notații, gradul de împrăştiere a punctelor corespunzător unei clusterizări $\Im$ este,

(3.5.) $e^2(\aleph,\Im)=\sum_{k=1}^{K}\sum_{C(i)=k}\left\|\mathbf{x}_i-\mathbf{c}_k\right\|^2$,

unde $\mathbf{c}_k$ este vectorul medie asociat celei de-a *k*-a clase.

Criteriul de clasificare utilizat este minimizarea erorii pătratice. Sistemul de clase optim din acest punct de vedere, notat cu $\Im^*$, este definit prin,

(3.6) $\Im^*=\min_{\Im}\sum_{k=1}^{K}\sum_{C(i)=k}\left\|\mathbf{x}_i-\mathbf{c}_k\right\|^2$.

Sistemul de clustere $\Im^*$ poate fi obținut pe baza următoarei observații.

Observație Pentru orice mulțime de observații *S*,

(3.7) $\mathbf{c}_S=\arg\min_{\mathbf{m}}\sum_{i\in S}\left\|\mathbf{x}_i-\mathbf{m}\right\|^2$.

Rezultă,

(3.8) $\Im^*=\min_{\Im,\{\mathbf{m}_k\mid 1\le k\le K\}}\sum_{k=1}^{K}\sum_{C(i)=k}\left\|\mathbf{x}_i-\mathbf{m}_k\right\|^2$.

Pe baza relațiilor (3.7) şi (3.8) rezultă algoritmul k-means.

Clusterizarea k-means

Repetă

Pas1. Pentru un sistem de clase dat, $\Im$, varianța totală definită în (3.8) prin $\sum_{k=1}^{K}\sum_{C(i)=k}\left\|\mathbf{x}_i-\mathbf{m}_k\right\|^2$ este minimizată în raport cu $\{\mathbf{m}_k\mid 1\le k\le K\}$şi sunt obținuți centroizii $\{\mathbf{c}_k\mid 1\le k\le K\}$ corespunzător relației (3.8)

Pas 2. Pe baza sistemului curent de centroizi $\{\mathbf{c}_k\mid 1\le k\le K\}$, cantitatea $\sum_{k=1}^{K}\sum_{C(i)=k}\left\|\mathbf{x}_i-\mathbf{c}_k\right\|^2$ este minimizată prin asignarea fiecărei observații clasei cu centroidul cel mai apropiat, deci $C(i)=\arg\min_{1\le k\le K}\left\|\mathbf{x}_i-\mathbf{c}_k\right\|^2$

până când structura clusterelor se stabilizează (nu au loc alte reasignări ale formelor)

3.3. Clusterizarea în context fuzzy

Tehnicile hard de analiza claselor generează partiții ale mulțimii formelor $\aleph$ şi astfel încât fiecare observație este asignată unic unei clase. Clusterizarea fuzzy extinde noțiunea de analiză hard prin asocierea fiecărei forme la o funcție de apartenență. Rezultatul tehnicilor de clusterizare fuzzy este tot un sistem de clase, dar nu o partiție.

Tehnica generală de clusterizare fuzzy poate fi descrisă astfel. Fie *N* numărul de forme ale mulțimii $\aleph$ şi *K* numărul de clustere.

În cadrul clusterizării fuzzy, fiecare cluster este o mulțime fuzzy a tuturor formelor din $\aleph$. Rezultatul reprezentării prin intermediul sistemului de clase este, pentru fiecare clasă *C*, setul perechilor ordonate (i,μ_i), unde *i* este ce-a de-a *i*-a formă, iar μ_i este gradul de apartenență a formei la clasa *C*. Valori mari

ale gradului de apartenență corespund unei valori mari de încredere în asignarea observației respective clasei considerate.

Metodele de clusterizare fuzzy determină obținerea de partiții crisp (sau hard) prin stabilirea unui prag pentru gradele de apartenență rezultate.

Unul dintre cele mai folosite modele de analiză a claselor în context fuzzy probabilist este modelul *c-means* (Bezdek, 1981). În acest caz, funcția criteriu (care trebuie minimizată) este definită pe baza unui grad de fuzzyficare, $m > 1$, prin

$$(3.9)\quad E^2(\aleph, U, \mathbf{c}) = \sum_{i=1}^{N}\sum_{k=1}^{K} u_{ij}^m \left\|\mathbf{x}_i - \mathbf{c}_k\right\|^2 ,$$

unde $\mathbf{c}_k$ este centroidul celui de-al *k*-lea cluster.

În acest caz, argumentele $U^*, \mathbf{c}^*$ care minimizează (3.9) sunt calculate prin (Bezdek, 1981)

$$(3.10)\quad u_{ik}^* = \left[\sum_{j=1}^{K}\left(\frac{\left\|\mathbf{x}_i - \mathbf{c}^*{}_k\right\|^2}{\left\|\mathbf{x}_i - \mathbf{c}^*{}_j\right\|^2}\right)^{\frac{2}{m-1}}\right]^{-1}$$

$$(3.11)\quad \mathbf{c}^*{}_k = \frac{\sum_{i=1}^{N}(u)^{*m}{}_{ik}\,\mathbf{x}_i}{\sum_{i=1}^{N}(u)^{*m}{}_{ik}}$$

pentru $1 \le i \le N, 1 \le k \le K$.

Pentru descrierea algoritmului, vom considera, la fiecare moment de timp *t*, $0 \le t \le T$, perechea de parametri $(U^t, \mathbf{c}^t)$, T prestabilit. Fie $\varepsilon \ge 0$ eroarea admisă pentru calculul centroizilor sistemului de clase optimal.

Algoritmul c-means în contextul fuzzy probabilist

Pas1. Selectează sistemul inițial de parametri, $(U^0, \mathbf{c}^0)$, $t = 0$

Repetă

Pas 2. Actualizează sistemul parametrilor pe baza relațiilor (3.10) şi (3.11), astfel, pentru $1 \le i \le N, 1 \le k \le K$

$$u_{ik}^{t+1} = \left[\sum_{j=1}^{K}\left(\frac{\left\|\mathbf{x}_i - \mathbf{c}^t{}_k\right\|^2}{\left\|\mathbf{x}_i - \mathbf{c}^t{}_j\right\|^2}\right)^{\frac{2}{m-1}}\right]^{-1},$$

$$\mathbf{c}_k^{t+1} = \frac{\sum_{i=1}^{N}\left(u^t\right)_{ik}^m \mathbf{x}_i}{\sum_{i=1}^{N}\left(u^t\right)_{ik}^m}$$

$t = t + 1$

până când $t = T$ *sau* $\left\|\mathbf{c}^{t+1} - \mathbf{c}^t\right\| \le \varepsilon$

3.4. Clusterizarea în context evoluționist

Algoritmii genetici reprezintă o abordare a problemei de căutare a unei soluții optimale. La fiecare iterație *t*, algoritmul menține o populație de soluții potențiale, numite cromozomi, $P(t)=\{x_1^t, x_2^t, \ldots, x_n^t\}$. Cromozomii sunt reprezentări binare ale soluțiilor considerate la momentul respectiv. Fiecare cromozom $x_i^t, t = \overline{1,n}$ este evaluat în scopul măsurării performanțelor sale din punct de vedere al funcției obiectiv considerate, aplicate pentru soluția a cărei reprezentare o constituie. La iterația *t+1* este selectată o nouă populație, pe baza calității indivizilor populației de la momentul anterior. O parte a membrilor acestei populații suferă modificări de tip mutație şi crossover (încrucişare) şi determină noi indivizi.

Un algoritmul genetic este descris astfel. (Banzhaf şi Reeves, 1999)

procedure genetic
begin t=0
inițializează P(t)
evaluează P(t)
while not(condiție de terminare)
begin t=t+1
selectează P(t) din P(t-1)
modifică P(t)
evaluează P(t) end
end.

Condiția de terminare a algoritmului se referă, în general, la posibilitatea îmbunătățirii funcției obiectiv la iterația curentă.

Abordarea evoluționistă a clusterizării este justificată intuitiv de o evoluție naturală a populației soluțiilor pe baza operatorilor evoluționişti către o partiție optimă a datelor. So-

luţiile candidate la obţinerea sistemului de clase sunt reprezentate prin cromozomi. Funcţia de evaluare determină probabilitatea fiecărui cromozom de a "supravieţui" în generaţia următoare.

Algoritm generic pentru clusterizare evoluţionistă

Pas 1. Selectează aleator populaţia iniţială corespunzătoare soluţiei. În acest context, fiecare soluţie corespunde unei *k*-partiţii a datelor. Fiecărei soluţii îi este asociată o valoare de tip fitness. În general, performanţa unui cromozom este invers proporţională cu valoarea eroare pătratică: o soluţie cu o eroare pătratică mică este creditată cu o valoare fitness mare.

Repetă

Pas 2. Utilizează operatorii de selecţie, recombinare şi mutaţie pentru generarea următoarei populaţii de soluţii. Evaluează performanţa soluţiilor obţinute

până când CT este îndeplinită,

unde *CT* reprezintă condiţia terminală.

Într-o abordare hibridă a analizei sistemelor de clustere, algoritmul genetic este utilizat pentru determinarea unor centroizi iniţiali adecvaţi, partiţia finală fiind ulterior determinată prin aplicarea algoritmului *k-means*. (Babu şi Murty, 1993).

4. Tehnici de clasificare supervizată utilizând PCA

Fie $X_1, X_2, ..., X_N$ forme de dimensiune *n*, clasificate până la momentul curent în clasele $C_1, C_2, ..., C_M$. Prima ipoteză de lucru este aceea că fiecare clasă corespunde unui proces stochastic staţionar.

Pentru orice $1 \le i \le M$, clasa C_i este definită prin,

$C_i = \{X_1^i, X_2^i, ..., X_{N_i}^i\}$, unde

$\sum_{i=1}^{M} N_i = N$.

Conform ipotezei de lucru, $\{X_1^i, X_2^i, ..., X_{N_i}^i\}$ este o mulţime de realizări ale unui proces stochastic staţionar, pentru $1 \le i \le M$.

Fie Σ_i matricea de covarianţă şi $\hat{\Sigma}_i$, $\hat{\mu}_i$ matricea de covarianţă de selecţie, respectiv vectorul medie de selecţie ai formelor clasificate în C_i, pentru orice $1 \le i \le M$,

$$\hat{\mu}_i = \frac{1}{N_i}\sum_{k=1}^{N_i} X_k^i$$

$$\hat{\Sigma}_i = \frac{1}{N_i - 1}\sum_{k=1}^{N_i}\left(X_k^i - \hat{\mu}_i\right)\left(X_k^i - \hat{\mu}_i\right)^T .$$

În continuare vom presupune că fiecare clasă este suficient de bogată astfel încât matricele de covarianţă de selecţie să aproximeze suficient de bine matricele de covarianţă teoretice.

Fie X_{N+1} noua formă de clasificat. *Ideea care stă la baza tehnicilor de clasificare propuse este de a clasifica forma X_{N+1} în clasa C_i dacă modificările aduse componentelor principale ale matricei de covrianţă de selecţie $\hat{\Sigma}_i$ sunt "suficient" de mici*.

Pentru fiecare $1 \le i \le M$, vom nota cu $\hat{\Sigma}_{i,N+1}$ matricea de covarianţă de selecţie a clasei C_i "îmbogăţită" cu forma X_{N+1}. Rezultă,

$$\hat{\Sigma}_{i,N+1} = \frac{N_i - 1}{N_i}\hat{\Sigma}_i + \frac{1}{N_i + 1}\left(X_{N+1} - \hat{\mu}_i\right)\left(X_{N+1} - \hat{\mu}_i\right)^T$$

Utilizăm în continuare următoarele notaţii

- $\psi_i^1, \psi_i^2, ..., \psi_i^n$ vectorii proprii ai matricei $\hat{\Sigma}_i$, corespunzători valorilor proprii ordonate, $\lambda_i^1 > \lambda_i^2 > ... > \lambda_i^{m_i} \ge \lambda_i^{m_{i+1}} \ge ... \ge \lambda_i^n$, $1 \le i \le M$;
- $\psi_i^1, \psi_i^2, ..., \psi_i^{m_i}$ componentele principale ale formelor clasificate în C_i, $1 \le i \le M$;
- $\psi_{i,N+1}^1, \psi_{i,N+1}^2, ..., \psi_{i,N+1}^n$ vectorii proprii ai matricei $\hat{\Sigma}_{i,N+1}$, corespunzători valorilor proprii ordonate, $\lambda_{i,N+1}^1 > \lambda_{i,N+1}^2 > ... > \lambda_{i,N+1}^{m'_i} \ge \lambda_{i,N+1}^{m'_{i+1}} \ge ... \ge \lambda_{i,N+1}^n$, $1 \le i \le M$;
- $\psi_{i,N+1}^1, \psi_{i,N+1}^2, ..., \psi_{i,N+1}^{m'_i}$ componentele principale ale formelor din setul $\{X_1^i, X_2^i, ..., X_{N_i}^i, X_{N+1}\}$, $1 \le i \le M$
- $A_i = \hat{\Sigma}_{i,N+1} - \hat{\Sigma}_i$, $1 \le i \le M$.

Observaţie Pentru $1 \le i \le M$, valorile proprii $\lambda_{i,N+1}^1 > \lambda_{i,N+1}^2 > ... > \lambda_{i,N+1}^{m'_i} \ge \lambda_{i,N+1}^{m'_{i+1}} \ge ... \ge \lambda_{i,N+1}^n$

sunt calculate prin aproximări de ordinul I, (State, Cocianu, Vlamos, Ştefănescu, 2006)

$\lambda_{i,N+1}^k \cong \left(\psi_i^k\right)^T \hat{\Sigma}_{i,N+1} \psi_i^k$, $1 \le k \le n$.

Metodologia de clasificare a formei X_{N+1} dirijată de componentele principale ale formelor fiecărei clase C_i poate fi descrisă astfel.

Fie $C_{j_1}, C_{j_2}, ..., C_{j_t}$ cu proprietatea că

(4.1) $m_{j_i} = m'_{j_i}$, $1 \le i \le t$

Fie $\psi_{j_i,N+1}^1, \psi_{j_i,N+1}^2, ..., \psi_{j_i,N+1}^{m_{j_i}}$ componentele principale, corespunzătoare matricei de covarianţă de selecţie $\hat{\Sigma}_{j_i,N+1}$. Utilizând teoria perturbaţiei rezultă,

$\psi_{j_i,N+1}^k = \psi_{j_i}^k + \sum_{l=1}^{m_{j_i}} \frac{\left(\psi_{j_i}^l\right)^T A_i \psi_{j_i}^k}{\lambda_{j_i}^k - \lambda_{j_i}^l} \psi_{j_i}^l$ (State, Cocianu, Vlamos, Ştefănescu, 2006)

Criteriile posibile de selecţie: X_{N+1} este clasificată în clasa C_{j_i} dacă

1.

$$D = \frac{1}{m_{j_i}} \sum_{k=1}^{m_{j_i}} d\left(\psi_{j_i}^k, \psi_{j_i,N+1}^k\right) = \min_{1 \le l \le t} \frac{1}{m_{j_l}} \sum_{k=1}^{m_{j_l}} d\left(\psi_{j_l}^k, \psi_{j_l,N+1}^k\right)$$

şi

(4.2) $D < \varepsilon$, ε parametru dat

Sau

2. Selectează clasa cu număr maxim de valori proprii semnificative şi aplică 1. doar acelor clase

Sau

3. Selectează clasa cu număr minim de valori proprii semnificative şi aplică 1. doar acelor clase

Bibliografie

BABU, G. P., MURTY, M. N., AND KEERTHI, S. S. 2000. Stochastic connectionist approach for pattern clustering. *IEEE Trans. Syst. Man Cybern.*.

BANZHAF, W., REEVES C. (Eds.), 1999. *Foundations of Genetic Algorithms*, Morgan Kaufmann Publ. Inc.

BEZDEK, J. C. 1981. *Pattern Recognition WithFuzzy Objective Function Algorithms*. Plenum Press, New York, NY.

BREIMAN, L. 1998. Arcing classifiers (with discussion). *Ann. Stat., 26,* 801–849.

DIDAY, E. 1988. The symbolic approach in clustering. In *Classification and Related Methods*, H. H. Bock, Ed. North-Holland Publishing Co., Amsterdam, The Netherlands.

DUBES, R. C. 1993. Cluster analysis and related issues. In *Handbook of Pattern Recognition& Computer Vision*, C. H. Chen, L. F. Pau,and P. S. P. Wang, Eds. World Scientific Publishing Co., Inc., River Edge, NJ, 3–32.

FABER, V. 1994. Clustering and the continuous k-means algorithm. *Los Alamos Science 22*, 138–144.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. 2001. *The Elements of Statistical Learning Data Mining, Inference, and Prediction.* Springer-Verlag

JAIN, A.K., MURTY, M.N. AND FLYNN, P.J. 1999. Data clustering: a review, ACM Computing Surveys, Vol. 31, No. 3, September 1999

JELINEK F., 1997. *Statistical Methods for Speech Recognition*, MIT Press

PARIDA, L., GEIGER, D., AND HUMMEL, R. 1998. Junctions: detection, classification, and reconstruction. *IEEE Trans. Pattern Anal. Machine Intell., 20*

STATE, L. AND COCIANU C., 1997. *Determinarea caracteristicilor lineare optimale din punct de vedere informational in compresia/decompresia datelor*, Informatica Economica, Vol. 1, Nr. 4, 1997

STATE, L., COCIANU C., VLAMOS P. AND ŞTEFĂNESCU V. *PCA-Based Learning Algorithm for Solving Recognition Tasks*, Proceedings of KCC 2006, Orlando, July 16-19 2006, în curs de apariţie

STATE L., 1997. *Analiza în componente principale pentru compresia/ restaurarea datelor*, Informatica Economică, Nr. 2/1997

WAH, B. W., Ed. 1996. Special section on mining of databases. *IEEE Trans. Knowl. Data Eng.* (Dec.).

WANG, Y. AND STAIB, L. H. 2000. Boundary finding with prior shapes and smoothness models. *IEEE Trans. Pattern Anal. Machine Intell., 22*

WILSON, D. R. AND MARTINEZ, T. R. 1997.Improved heterogeneous distance functions. *J. Artif. Intell. Res. 6*, 1–34.