O tehnica fuzzy de partitionare si inductie automata bazata pe extensia fuzzy a distantei c^2

Conf.dr. Vasile GEORGESCU
Universitatea din Craiova, vgeo@central.ucv.ro

Lucrarea propune un sistem de achizitie automata a cunostintelor ce include mecanisme fuzzy de partitionare a domeniului variabilelor continue si de generare inductiva a arborelui de decizie. Ambele procese au la baza adaptarea distantei $?^2$ pentru a opera în tabele de contingenta fuzzy, în care frecventele se definesc prin grade de apartenenta la elementele unor partitii fuzzy. În scopul punerii de acord a unui test probabilist cu o descriere fuzzy a datelor s-a introdus un criteriu ce restrictioneaza alegerea schemelor de acoperire a domeniului unei variabile continue cu multimi fuzzy. Aplicarea criteriului ne permite sa interpretam vectorul gradelor de apartenenta atasat fiecarei valori $x \in Dom(X)$ în termenii unei distributii de probabilitati. Cu ajutorul codificarii fuzzy astfel definite, un algoritm de discretizare induce partitiile fuzzy optimale ale tuturor predictorilor. Arborele de decizie fuzzy este generat apoi recursiv prin selectia optimala a predictorului în raport cu care are loc ramificarea în fiecare nod, selectie ce apeleaza tot la distanta $?^2$ adaptata tabelelor de contingenta fuzzy.

Cuvinte cheie: *fuzzy, inductie automata, clasificare, discretizare, algoritmi, codificare.*

Algoritmi de inductie si clasificare automata

Tehnicile actuale de achizitie a datelor produc colectii masive de informatii. Generarea automata a cunostintelor din date este un proces complex ce apeleaza la mecanisme de inductie automata si face posibila transformarea informatiei amorfe, latente, într-o forma structurata, inteligibila si deci "digerabila" pentru comunitatea stiintifica din domeniul de expertiza respectiv.

Sistemele de învatare automata sunt reunite de obicei sub numele de *algoritmi de inductie* (*inducers*) si au drept scop sa genereze *mecanisme de clasificare automata* (*classifiers*) cu mare putere predictiva, prin extragerea, abstractizarea si codificarea informatiei continute într-o baza de date. În acest context, algoritmii ce induc arbori de decizie reprezinta instrumente particulare deosebit de utile. Ei genereaza proceduri de clasificare ce utilizeaza structura de arbore pentru a selecta regula de decizie cea mai potrivita într-un context precizat. Cunoasterea astfel structurata poate fi imediat transpusa într-un formalism de reprezentare compatibil cu sistemele inferentiale bazate pe reguli. Avantajul lor major este gradul înalt de inteligibilitate si usurinta cu care pot fi procesate cunostintele pentru a infera noi concluzii, decizii, sau predictii.

În continuare vom introduce câteva notatii si definitii. Fie T un set de date destinat învatarii regulilor de decizie (training), ce constau din n sentinte (numite si exemple), fiecare având atasata o anumita eticheta.

Orice sentinta etichetata are doua parti:

- o parte neetichetata, notata prin x_k $(k = \overline{1,n})$, care este un vector ale carui elemente x_k^S sunt valori ale atributelor corespunzatoare $(X^s)_{s=\overline{1,m}}$;
- un atribut nominal special, eticheta, notata prin Y si reprezentând variabila tinta (sau decizia), adica variabila ce focalizeaza procesul de învatare si despre care dorim sa facem predictii (o valoare actuala y_k a etichetei Y, cuplata cu vectorul atributelor x_k, specifica o clasificare cunoscuta, adica o combinatie ce trebuie folosita pentru a învata regulile de decizie).

Atributele (X^s) sunt numite uneori *predictori* si au un rol explicativ în clasificare. Fie $Dom(X^s)$ domeniul predictorului X^s. Fiecare sentinta neetichetata este un element al spatiului sentintelor neetichetate $X = Dom(X^1) \times ... \times Dom(X^m)$, unde m este numarul atributelor.

Daca desemnam prin Y multimea valorilor posibile ale etichetelor, atunci o sentinta etichetata $(x_k \in X, y_k \in Y)$ este un element al spatiului $X \times Y$.

Sarcina unui algoritm de inductie I este sa produca un mecanism eficient de clasificare dintr-o multime de sentinte etichetate, adica sa aplice un set de date T destinat învatarii, în multimea clasificarilor posibile C. Mecanismul de clasificare astfel generat (bazat pe asocierile pertinente dintre partea neetichetata a sentintelor si o anumita eticheta) poate fi folosit apoi pentru a clasifica orice alta sentinta neetichetata $x \in X$, prin punerea sa în corespondenta cu o eticheta $y \in Y$, care este tinta predictiei.

Presupunem ca setul de date este alcatuit din sentinte independente si identic distribuite, motiv pentru care vom asocia atributelor o interpretare conjunctiva (partea conditionala a regulii reprezinta o conjunctie de premise simple).

În particular, un mecanism de clasificare cu structura de arbore, contine arborele de decizie indus, care pune în corespondenta o sentinta neetichetata cu o clasa prin parcurgerea drumului de la radacina la o frunza si returneaza eticheta ce desemneaza aceasta clasa în frunza.

În continuare, scopul nostru este sa investigam câteva tehnici pentru generarea automata a mecanismelor de clasificare (clasificatoarelor) fuzzy, folosind algoritmi de inductie special modificati, capabili sa produca reguli de decizie cu premize fuzzy.

Arbori de decizie fuzzy

Arborii de decizie fuzzy pot fi priviti ca o generalizare naturala a arborilor de decizie traditionali. Implementând algoritmi de învatare discriminativi si operând prin partitionare recursiva, ideile lor de baza sunt aceleasi: sa partitioneze spatiul observatiilor în subesantioane dupa criterii de selectie induse de date si sa reprezinte partitiile sub forma unui arbore. Mecanismul inferential este destul de asemanator: cunoasterea indusa este folosita pentru a produce alte clasificari, prin analiza corespondentei dintre atributele unui nou exemplu si conditiile întâlnite pe traseele de la radacina la frunzele arborelui.

Cu toate acestea, sunt multe aspecte în care cele doua abordari difera.

În mod esential, arborii de decizie fuzzy folosesc reprezentari fuzzy, care furnizeaza un cadru simbolic pentru achizitia cunostintelor în contexte imprecise si incerte si permite îmbunatatirea metodologiilor existente în cele doua componente majore ale acestora: procedurile de generare a arborelui si procedurile inferentiale.

· **Acoperirea fuzzy, comparativ cu partitionarea stricta a unui domeniu continuu**

În sistemele de învatare automata traditionale, doua tipuri de domenii se întâlnesc în mod curent: cele corespunzând variabilelor simbolice (nominale, definite lingvistic printr-un numar mic de modalitati) si cele corespunzând variabilelor numerice (discrete sau continue).

Cu toate ca variabilele continue se întâlnesc frecvent în practica, numerosi algoritmi de inductie dispun doar de mecanisme de învatare în spatiul atributelor nominale. Pentru a opera cu variabile continue, astfel de algoritmi utilizeaza în mod uzual tehnici de partitionare a domeniului variabilelor continue, pentru a produce atribute nominale. Asemenea tehnici se mai numesc *proceduri de discretizare* (*discretizers*).

Data fiind o partitie a unui domeniu real în intervale, un "discretizor" este o functie de codificare ce pune în corespondenta valorile unei variabile reale continue ce cad

într-un anumit interval, cu modalitatile (categoriile) asociate ale unui atribut nominal.
Atât partitiile stricte cât si acoperirile fuzzy trebuie sa fie *complete*, adica fiecare valoare a domeniului trebuie sa apartina cel putin unui interval (sau multime fuzzy).
Contrar unei partitii, care trebuie sa fie *consistenta*, acoperirea fuzzy poate fi *inconsistenta*, datorita unui anumit grad de acoperire reciproca a multimilor fuzzy (altfel spus, o valoare a domeniului poate cadea în mai multe multimi fuzzy, cu anumite grade de apartenenta).
Termenii lingvistici definiti în mod gradual de elementele unei scheme de acoperire fuzzy confera abilitatea de a modela detalii de cunoastere foarte fine, precum si posibilitatea de a trata date afectate de perturbatii sau având atribute lipsa.

· **Utilizarea restrictiilor fuzzy în partitionarea recursiva a arborelui**

Arborii de decizie fuzzy difera totodata de cei traditionali prin capacitatea de a folosi criterii de ramificare ce suporta nu numai restrictii stricte, ci si restrictii fuzzy.
Generarea arborelui se bazeaza pe o partitionare recursiva. Radacina arborelui de decizie reprezinta întregul spatiu de observatii atât timp cât nu sunt impuse restrictii. În consecinta, el contine toate sentintele destinate învatarii.
Restrictiile (fie stricte, fie fuzzy) sunt definite în raport cu termenii lingvistici A_i^s ai fiecarui atribut X^s si au forma: "X^s este A_i^s", unde A_i^s denota fie o categorie simbolica, fie unul din elementele ce acopera domeniul (intervale, sau multimi fuzzy).
În scopul ramificarii unui nod N în timpul procesului de expandare recursiva a arborelui, trebuie sa selectam un atribut X^s (ce nu apare pe drumul de la radacina catre N), de obicei cel care maximizeaza o anumita masura de similaritate, sau câstig de informatie, pentru sentintele destinate învatarii prezente în nod. Acum, nodul poate fi ramificat prin partitionarea sentintelor sale potrivit restrictiilor "X^s este A_i^s" care sunt folosite apoi pentru a genera conditii care sa conduca spre nodurilor fii (N_i).
Aceeasi procedura este repetata recursiv pentru fiecare nod fiu, pâna se atinge un criteriu de stop.
Dupa cum stim, arborii de decizie clasici opereaza cu variabile continue prin discretizarea domeniului lor în partitii definite în mod strict, ale caror componente sunt multimi clasice (intervale, în cele mai multe cazuri). Din acest motiv, valoarea uneia dintre variabilele ce descriu o sentinta satisface exact una din posibilele restrictii ce se ramifica din nodul respectiv (aceea asociata multimii clasice a carei functie caracteristica este 1). Prin urmare, fiecare sentinta cade exact într-un subnod si astfel exact într-o frunza (cu exceptia cazurilor când lipsesc valorile unor atribute).
Dimpotriva, arborii de decizie fuzzy sunt capabili sa proceseze variabile continue prin generarea de acoperiri fuzzy (secvente ordonate de multimi fuzzy ce se suprapun partial). Pentru ca fiecare restrictie fuzzy este definita de functia de apartenenta a multimii fuzzy, devine posibil ca un exemplu dat sa satisfaca mai multe conditii si astfel sa cada în mai multe noduri fiu.
Arborii de decizie fuzzy pot procesa sentinte (exemple) care contin date exprimate atât prin valori numerice cât si prin termeni fuzzy. Pentru a face fata acestei situatii este suficient sa dispunem de un operator apt sa masoare gradul de concordanta (similaritate) dintre entitati nu numai de acelasi tip, ci si de tipuri diferite (fie ele valori numerice, sau termeni fuzzy).
De exemplu, fie X^s un atribut, $U^s = Dom(X^s)$ universul sau de discurs, iar A_i^s o restrictie fuzzy. În cazurile în care atributul ia o valoare numerica, notata x^s, o masura a gradului în care x^s satisface restrictia fuzzy A_i^s este data prin chiar functia de apartenenta a lui A_i^s, evaluata în x^s: $\mu_{A_i^s}(x^s)$.

Când atributul ia o valoare fuzzy notata F^s, un indice de concordanta fuzzy poate fi exprimat, de exemplu, prin masura de posibilitate:

$$\Pi(F^s, A_i^s) = \sup_{u \in U^s} \mu_{F^s \cap A_i^s}(u) = \sup_{u \in U^s} \min(\mu_{F^s}(u), \mu_{A_i^s}(u))$$

Pentru ca de obicei multimile fuzzy se suprapun, iar indicele de concordanta este un numar din intervalul [0, 1], cea mai mare parte a sentintelor satisfac în acelasi timp mai multe restrictii fuzzy (dar numai partial) si astfel cad în mai multe frunze, fiecare dintre ele cu o pondere cuprinsa în [0,1].

În nodul radacina, toate cele n sentinte au aceeasi pondere: $w_k^{root} = 1 \ (k = \overline{1, n})$. Pentru a atinge un nod N, un anumit numar de restrictii fuzzy trebuie sa fie satisfacute. În acest nod, o sentinta data e_k are ponderea w_k^N calculata într-un mod conjunctiv, pe baza indicilor de concordanta ce desemneaza gradul în care sunt satisfacute restrictiile fuzzy de-a lungul drumului de la radacina la N. Fie N_i fiul de rang i al lui N si sa presupunem ca se cunoaste w_k^N. Atunci, ponderea în N_i, notata prin $w_k^{N_i}$, poate fi evaluata recursiv:

$$w_k^{N_i} = T\left(w_k^N, \Pi(e_k^s, A_i^s)\right)$$

unde T este o T-norma (de obicei *min*, sau *produs*), iar A_i^s este termenul fuzzy asociat cu restrictia fuzzy ce conduce la N_i. Cumulând ponderile pentru toate sentintele ce cad în N_i obtinem frecventa de aparitie a termenului lingvistic A_i^s:

$$\sum_{k=1}^{n} T\left(w_k^N, \Pi(e_k^s, A_i^s)\right)$$

În forma relativa, aceasta frecventa estimeaza probabilitatea ca o sentinta prezenta în N_i sa satisfaca restrictia asociata termenului lingvistic A_i^s:

$$p_{i\cdot}^N = p_{A_i^s}^N = \frac{\sum_{k=1}^{n} T\left(w_k^N, \Pi(e_k^s, A_i^s)\right)}{\sum_{i=1}^{p} \sum_{k=1}^{n} T\left(w_k^N, \Pi(e_k^s, A_i^s)\right)}$$

Analog, probabilitatea ca o sentinta aflata în N sa reprezinte clasa lui Y asociata termenului lingvistic B_j este estimata de:

$$p_{\cdot j}^N = p_{B_j}^N = \frac{\sum_{k=1}^{n} T\left(w_k^N, \Pi(e_k^y, B_j)\right)}{\sum_{j=1}^{q} \sum_{k=1}^{n} T\left(w_k^N, \Pi(e_k^y, B_j)\right)}$$

Discretizare prin codificare fuzzy standard. O interpretare probabilista

Codificarea fuzzy este un instrument general care poate fi adaptat cu usurinta pentru diferite scopuri, dar mai specific el apare în discretizarea variabilelor aleatoare continue pe intervalele unei partitii finite a lui $\Re$. Concret, el consta în generarea codificarii Dirac (care atribuie fiecare observatie unui interval) prin asocierea unei observatii (eventual) la mai multe intervale, cu ponderi (grade de apartenenta) proprii fiecarui interval.

Fie Z o variabila aleatoare continua, definita pe spatiul de probabilitate (Ω, A, μ). De obicei, statisticianul nu se ocupa cu fenomenul subiacent concretizat prin Z, ci numai cu un esantion de volum n $(X_1(\omega), ..., X_n(\omega))$ extras dintre valorile unei variabile observate X, care ar putea fi o "aproximare" a lui Z.

Pentru a furniza un model al fuzzicitatii dintre X si Z, putem folosi o codificare fuzzy a lui X, care sa transforme esantionul original într-un esantion codificat, asociat unei variabile aleatoare nominale ordonata.

Conceptul de *codificare fuzzy standard* ne permite sa obtinem o interpretare probabilista a procedurii de discretizare: aceasta înseamna sa asociem fiecarei observatii $x \in X(\omega)$ a unei variabile reale aleatoare X, o distributie de probabilitati P_x pe $\Re$, înzestrata cu o σ-algebra finita, generata de o partitie a lui $\Re$ în intervale. În scopul obtinerii acestei interpretari probabiliste, ponderile observatiilor referitoare la fiecare interval trebuie sa fie numere pozitive, a caror suma este 1.

În termeni generali, o *codificare fuzzy standard* este o aplicatie $C=(c_1,...,c_p)$ de la $\Re$ la $[0,1]^p$, care satisface conditiile:

(i) C este masurabila în raport cu σ-algebrele Boreliene ale lui $\Re$ si $[0,1]^p$;

(ii) $\forall x\in\Re,\ \sum_{i=1}^{p} c_i(x)=1$.

Functiile $c_1,...,c_p$ sunt numite functii de codificare fuzzy standard.

Când o partitie $(A_1,...,A_p)$ a lui $\Re$ în intervale este data pentru o variabila aleatoare reala, obtinem o interpretare a functiilor de codificare în termenii unor probabilitati de apartenenta ale unei observatii x la diverse clase. În acest caz, o distributie de probabilitati P_X pe $(\Re, B_p)$, definita de $P_X(A_i)=c_i(x)$, este asociata fiecarei observatii $x=X(\omega)$, unde B_p este σ -algebra generata pe $\Re$ de partitia $(A_1,...,A_p)$.

Mai formal, o codificare fuzzy standard asociata unei partitii $(A_1,...,A_p)$ este bine definita când este data o structura statistica $(\Re, B_p, \{P_X; x\in(\Re, B_\Re)\})$. Functiile masurabile $c_1,...,c_p$ din $(\Re, B_\Re)$ în $([0,1], B_{[0,1]})$ definite de $c_i(x)=P_X(A_i)$ sunt numite functii de codificare fuzzy standard. Potrivit acestei definitii, $P=(P_X)_{X\in\Re}$ este o distributie de probabilitati de tranzitie pe $(\Re, B_\Re)\times B_p$, adica $\forall B\in B_p$, $P_X(B)$ este o functie masurabila si $\forall x\in\Re$, P_X este o distributie de probabilitati pe B_p. Astfel, esantionului initial $(X_1(\omega), ..., X_n(\omega))$ i se va asocia un esantion codificat $(P_{X_1(\omega)},...,P_{X_n(\omega)})$ reprezentând o distributiei de probabilitati. Mai mult, pentru orice variabila aleatoare reala X, variabila p-dimensionala $\hat{X}_p=(c_1\circ X, ..., c_p\circ X)$ este numita variabila codificata fuzzy standard.

În particular, avem de a face cu o codificare de tip interval atunci când utilizam codificarea Dirac $P_X=\delta_X$, unde δ_X este masura Dirac $\delta_X(A_i)=c_i(x)=\chi_{A_i}(X)$, iar χ_{A_i} este functia caracteristica a intervalului A_i.

Tabele de contingenta pentru atribute fuzzy

Fie (X,Y) un cuplu de doua variabile aleatoare. Când cel putin una dintre aceste variabile este continua, o tabela de contingenta nu ar putea fi construita fara ca variabila sa fie mai întâi discretizata. Consideram doua cazuri:

Cazul I: O codificare fuzzy standard este aplicata lui X si o codificare Dirac lui Y

Presupunând ca X este continua, decidem sa o transformam într-o variabila fuzzy-evaluata, folosind o codificare fuzzy standard care genereaza o acoperire fuzzy. Presupunem, de asemenea, ca Y este fie nominala, fie continua (în ultimul caz discretizarea se va face prin aplicarea unei partitionari stricte în intervale).

Fie $(A_1,...,A_p)$ o acoperire fuzzy a domeniului lui X, $(B_1,...,B_q)$ o partitie stricta pe domeniul lui Y si $(x_k,y_k)_{k=\overline{1,n}}$ un esantion de volum n. Pentru a clasifica cele n valori ale lui Y pe partitia $(B_1,...,B_q)$ si a obtine distributia sa de frecvente avem nevoie doar de aplicarea codificarii Dirac:

$$n_{\cdot j}=\sum_{k=1}^{n} c_j(y_k)=\sum_{k=1}^{n}\delta_{y_k}(B_j)=\sum_{k=1}^{n}\chi_{B_j}(y_k);\quad j=\overline{1,q}$$

unde χ_{B_j} este functia caracteristica a lui B_j: $\chi_{B_j}(y)=\begin{cases}1 & y\in B_j\\ 0 & y\notin B_j\end{cases}$, **cu** $\sum_{j=1}^{q}\chi_{B_j}(y)=1$.

Pentru a clasifica cele n valori ale lui X pe acoperirea fuzzy $(A_1,...,A_p)$ si a genera distributia sa de frecvente, putem folosi o codificare fuzzy standard:

$$n_{i\cdot}=\sum_{k=1}^{n} c_i(x_k)=\sum_{k=1}^{n}\mu_{A_i}(x_k);\quad i=\overline{1,p}$$

unde functia de codificare fuzzy $c_i(x)$ este definita în mod natural de functia de apartenenta A_i: $\mu_{A_i}(x)$, supusa conditiei de a avea o interpretare probabilista:

$$\sum_{i=1}^{p} c_i(x) = \sum_{i=1}^{p} \mu_{A_i}(x) = \sum_{i=1}^{p} P_{x_k}(A_i) = 1, \ \forall x \in Dom(X)$$

Din aceasta conditie (satisfacuta exclusiv de o codificare fuzzy standard) putem deduce urmatoarea proprietate:

$$\sum_{i=1}^{p} n_{i\cdot} = \sum_{i=1}^{p}\sum_{k=1}^{n} \mu_{A_i}(x_k) = \sum_{k=1}^{n}\left(\sum_{i=1}^{p} \mu_{A_i}(x_k)\right) = \sum_{k=1}^{n} 1 = n$$

Fie acum $e_k = (x_k, y_k)$ un exemplu extras dintr-un esantion dat de volum n. De fapt, e_k desemneaza realizarea unui eveniment compus. Din acest motiv, clasificarea acestui exemplu în raport cu acoperirea fuzzy $(A_1, ..., A_p)$ si partitia $(B_1, ..., B_q)$ poate fi interpretata ca o conjunctie a doua functii de codificare. Deoarece modul obisnuit de a reprezenta o conjunctie este folosirea unei T-norme (*min* si *prod* sunt cele mai populare), avem:

$$n_{ij} = \sum_{k=1}^{n} T\left(\mu_{A_i}(x_k), \ \chi_{B_j}(y_k)\right); \quad i = \overline{1, p}; \ j = \overline{1, q}$$

Mai mult:

$$\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} T\left(\mu_{A_i}(x_k), \ \chi_{B_j}(y_k)\right) = \sum_{k=1}^{n}\sum_{i=1}^{p}\left(\sum_{j=1}^{q} T\left(\mu_{A_i}(x_k), \ \chi_{B_j}(y_k)\right)\right) =$$

$$= \sum_{k=1}^{n}\sum_{i=1}^{p} \mu_{A_i}(x_k) = \sum_{k=1}^{n} 1 = n$$

Într-adevar, ca rezultat al faptului ca o valoare data y_k cade exact într-un interval B_j, exista exact un termen în suma $\sum_{j=1}^{q} T\left(\mu_{A_i}(x_k), \ \chi_{B_j}(y_k)\right)$ pentru care $\chi_{B_j}(y_k) = 1$. Atunci, aplicând câteva proprietati ale T-normei (adica $T(u,1) = u$ si $T(u, 0) = 0$) putem deduce ca:

$$\sum_{j=1}^{q} T\left(\mu_{A_i}(x_k), \ \chi_{B_j}(y_k)\right) = T\left(\mu_{A_i}(x_k), \ 1\right) = \mu_{A_i}(x_k)$$

Cazul II: O codificare fuzzy standard este aplicata atât lui X cât si lui Y

Contrar cazului precedent, nu numai valorile atributului X satisfac mai multe restrictii fuzzy, ci si unele valori $y_k \in Dom(Y)$ au clasificari multiple. Putem vorbi acum numai despre gradele de apartenenta ale acestor exemple la diverse clase lingvistice. Din acest motiv, numarul exemplelor $e_k = (x_k, y_k)$ ce satisfac un cuplu (A_i, B_j) de restrictii fuzzy creste semnificativ.

Totusi, frecventele asociate tabelelului de contingenta pot fi definite analog:

$$n_{ij} = \sum_{k=1}^{n} T\left(\mu_{A_i}(x_k), \ \mu_{B_j}(y_k)\right); \quad i = \overline{1, p}; \ j = \overline{1, q}$$

Principala diferenta este ca $\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$ nu poate fi egala cu n, datorita algebrei T-normelor, chiar daca utilizam o codificare fuzzy standard.

Extensia fuzzy a masurilor de similaritate, bazate pe statistica testului χ^2

Testul χ^2 de independenta a doua variabile aleatoare sintetizate într-un tabel de contingenta $p \times q$ are ca obiect investigarea diferentelor în frecventa când un esantion de volum n este clasificat în p clase dupa primul atribut si în q clase dupa al doilea. Frecventele exemplelor în fiecare clasificare pot fi reprezentate simbolic ca în tabelul urmator:

X \ Y		Clase pentru Y					$n_{i\cdot}$
		B_1	...	B_j	...	B_q	
Clase pentru X	A_1	n_{11}	...	n_{1j}	...	n_{1q}	$n_{1\cdot}$
	⋮	⋮	...	⋮	...	⋮	⋮
	A_i	n_{i1}	...	n_{ij}	...	n_{iq}	$n_{i\cdot}$
	⋮	⋮	...	⋮	...	⋮	⋮
	A_p	n_{p1}	...	n_{pj}	...	n_{pq}	$n_{p\cdot}$
$n_{\cdot j}$		$n_{\cdot 1}$	...	$n_{\cdot j}$	...	$n_{\cdot q}$	n

Statistica testului este:

$$\chi^2 = n \cdot \sum_{i=1}^{p} \sum_{j=1}^{q} \left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n} \right) \frac{1}{n_{i\cdot}\, n_{\cdot j}}$$

Aceasta urmeaza o distributie χ^2 cu (p-1)·(q-1) grade de libertate. Daca χ^2 depaseste valoarea critica, atunci ipoteza nula conform careia cele 2 atribute sunt independente unul fata de celalalt este respinsa.

Inconvenientul acestui test statistic este ca el depinde de numarul gradelor de libertate. Limita superioara ce poate fi atinsa a fost stabilita astfel:

$$\frac{\chi^2}{n} \le \min\,(p-1, q-1)$$

Pentru a facilita comparatiile dintre astfel de valori ale testului pentru tabele de contingenta de diferite dimensiuni, au fost propuse alte masuri de similaritate, cu valori în [0, 1]. Cele mai cunoscute sunt:

- coeficientul lui Tschuprow:

$$T = \sqrt{\frac{c^2}{n \cdot \sqrt{(p-1)\cdot(q-1)}}} = \sqrt{\frac{\Phi^2}{\sqrt{(p-1)\cdot(q-1)}}},$$ unde

$$\Phi^2 = \frac{c^2}{n}$$

- coeficientul de contingenta al lui Cramer:

$$C = \sqrt{\frac{\chi^2}{n \cdot \min(p-1, q-1)}} = \sqrt{\frac{\Phi^2}{\min(p-1, q-1)}}$$

- coeficientul lui K. Pearson:

$$P = \sqrt{\frac{\chi^2}{n+\chi^2}}$$

Cea mai convenabila solutie pentru a asigura comparabilitatea dintre valorile testului calculate pentru tabele de contingenta de diferite dimensiuni este sa evaluam (prin intregrare numerica) masura de probabilitate:

$$P\left(\chi_\nu^2 < \chi^2\right) = \frac{1}{2^{\nu/2} \cdot \Gamma\left(\frac{\nu}{2}\right)} \int_0^{\chi^2} e^{-\frac{x}{2}}\, x^{\frac{\nu}{2}-1}\, dx$$

Nici o alta masura nu discrimineaza cu mai multa acuratete între atributele aflate în competitie si, mai important, aceasta masura ne da posibilitatea de a controla probabilistic procesul de ramificare a arborelui.

Pentru a pune de acord astfel de teste probabiliste cu o descriere fuzzy a datelor, putem folosi codificarea fuzzy standard pentru a genera acoperiri fuzzy corespunzatoare variabilelor continue si pentru a construi tabele de contingenta, asa cum am sugerat în sectiunea precedenta. Se obtine astfel un test mai putin sensibil la fluctuatiile modelului (alegerea intervalelor, fluctuatiile de esantionare, etc.).

Detalii privind implementarea algoritmului si testele numerice efectuate vor face obiectul unei lucrari viitoare.

Bibliografie

[1] Breiman, L., Friedman J.H., Olsen, R.A. & Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth.

[2] Fayyad, U.M & Irani, K.B. (1993). "Multi-interval discretization of continuous-valued attributes for classification learning", *in Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, pp. 1022-1027.

[3] Georgescu, V. (1998). **"**Flexible estimation of cost functions based on fuzzy logic modeling approach". *Fuzzy Economic Review*, Vol.III, No.2, pp. 49-68.
[4] Georgescu, V. (1998). **"**New estimation methods in fuzzy regression analysis, based on projection theorem and decoupling principle". *Fuzzy Economic Review*, Vol.III, No.1, pp. 21-38
[5] Georgescu, V. (1996). "A fuzzy generalization of principal components analysis and automatic classification, based on a new metric concept: the dissimilarity between fuzzy sets". *Proceedings of the Third Congress of SIGEF*, Buenos Aires, Paper 2.25
[6] Georgescu, V. (1995). *Expert system designing in fuzzy logic and possibilistic logic*. Craiova, Ed. Intarf
[7] Hawkins, D.M. & Kass, G.V. (1982), "Automatic Interaction Detection", in Hawkins, D.M. (ed.*), Topics in Applied Multivariate Analysis*, pp. 267-302, Cambridge Univ. Press: Cambridge.
[8] Quinlan, J.R. (1986). "The Effect of Noise on Concept Learning" in *Machine Learning* II, R. Michalski, J. Carbonell & T. Mitchell (eds), Morgan Kaufmann.
[9] Quinlan, J.R. (1986). "Induction on Decision Trees", in *Machine Learning*, Vol. 1, pp. 81-106.
[10] Quinlan, J.R. (1987) "Decision Trees as Probabilistic Classifiers", in *Proceedings of the Fourth International Workshop on Machine Learning*, pp. 31-37.
[11] Quinlan, J.R. (1993). "Combining Instance-Based and Model-Based Learning", in *Proceedings of International Conference on Machine Learning*, Morgan Kaufmann, pp. 236-243.
[12] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.