Metode de selecţie a trăsăturilor

1. Consideraţii teoretice

Selecția trăsăturilor reprezintă etapa din procesul de data mining în urma căreia păstrăm în reprezentarea setului de date folosit doar acele trăsături pe care le considerăm mai bune (mai relevante) pentru reprezentarea datelor. De obicei noua dimensiune a vectorilor de reprezentare este mult mai mică decât reprezentarea curentă.

2. Entropia şi Câştigul Informaţional (IG)

O măsură numită *Entropy Based Discretization* este o măsură utilizată în mod frecvent în teoria informaţiei, care caracterizează (im)puritatea unei colecţii arbitrare de eşantioane (vectori de antrenament, documente), fiind o măsură a omogenităţii setului de eşantioane. Entropia poate fi utilizată pentru a împărţi recursiv valorile unui atribut numeric.

Câştigul informaţional şi entropia sunt funcţii ale distribuţiei probabilistice care susţin procesul de comunicare. Entropia este o măsură a incertitudini unei variabile aleatoare. Dându-se o colecţie S de *n* eşantioane grupate în *c* concepte ţintă (clase), entropia lui S relativă la clasificare eşantioanelor este:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i) \qquad (1)$$

unde p_i este procentajul din *S* care aparţine clasei *i*.

În toate calculele care implică entropia considerăm ca $0*\log_2 0$ este 0. Observăm că entropia este 0 dacă toţi membrii din S aparţin aceleaşi clase şi entropia este maximă ($\log_2 c$) când eşantioanele sunt egal distribuite în clase. Dacă eşantioanele, din colecţie, sunt inegal distribuite în clase, entropia este între 0 şi $\log_2 c$.

O interpretare din teoria informaţiei a entropiei este că ea specifică numărul minim de biţi de informaţie care sunt necesari pentru a codifica clasificarea după un membru arbitrar din S. Dacă p_i este 1, receptorul ştie că eşantionul este întotdeauna 1, astfel nici un mesaj nu trebuie trimis de la emiţător şi atunci entropia este 0. Pe de altă parte, dacă p_i este 1/c, sunt necesari c/2 biţi pentru a indica în ce clasă se găseşte eşantionul curent.

Logaritmul va fi în baza 2 (chiar dacă avem mai mult de 2 clase) deoarece entropia este o măsură a lungimii de codificare aşteptată, măsurată în biţi.

Pentru a înţelege aceasta presupunem următorul exemplu. Considerăm o colecţie S de 12 eşantioane (prezentată în tabelul următor). Această colecţie conţine 3 clase „a”, „b”, şi „c”. 5 eşantioane sunt clasificate în clasa „a”, 3 în clasa „b” şi 4 în clasa „c”. Pentru aceasta folosim notaţia S=[5a, 3b, 4c]. Entropia lui S relativă la această clasificare este:

$$Entropy[5a,3b,4c] = -\frac{5}{12}\log_2\frac{5}{12} - \frac{3}{12}\log_2\frac{3}{12} - \frac{4}{12}\log_2\frac{4}{12} = 1.5545$$

Atribute(cuvinte) / Eşantioane (documente)	$atrib_1$	$atrib_2$	$atrib_3$	$atrib_4$	$atrib_5$	$atrib_6$	clasa
v_1	1	5	7	0	1	1	a
v_2	1	1	0	0	7	5	b
v_3	0	0	1	7	1	0	c
v_4	1	3	7	1	0	1	a
v_5	2	1	8	0	0	0	a
v_6	0	0	0	0	10	7	b
v_7	0	0	7	10	7	0	c
v_8	0	0	5	3	8	1	c
v_9	2	2	5	2	0	1	a
v_{10}	0	1	0	1	9	5	b
v_{11}	0	1	4	1	9	2	c
v_{12}	1	4	6	1	0	0	a

3. Câștigul informațional

Pe baza entropiei este definită o măsură a gradului de eficienţă în selecţia trăsăturilor. Această măsură este numită *Câştigul informaţional* şi reprezintă de fapt reducerea în entropie, cauzată de gruparea eşantioanelor în acord cu un atribut. Mai precis, câştigul informaţional pentru un atribut relativ la o mulţime de eşantioane S este definit ca:

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

unde *Values(A)* este mulţimea de valori posibile pentru atributul A şi S_v este submulţimea din S pentru care atributul A are valoarea v.

Utilizând ecuaţia de mai sus, pentru fiecare trăsătură se calculează câştigul informaţional obţinut dacă mulţimea este împărţită utilizând acea trăsătură. Se obţin valori între 0 şi 1 fiind apropiate de 1 dacă trăsătura împarte mulţimea originală în două submulţimi cu dimensiuni apropiate.

Continuând cu exemplu de mai sus, vom calcula câştigul informaţional. Pentru simplificarea exemplului considerăm că atributele pot lua doar 2 valori: 0 dacă atributul nu este în eşantion (document) şi 1 altfel. Vom nota setul de eşantioane care conţine un atribut $S_{attribi>0}$ şi cel care nu conţine acel atribut cu $S_{attribi=0}$. De exemplu pentru $atrib_1$ cele 12 eşantioane sunt împărţite astfel:

- 6 în $S_{attrib1>0}$ din care 5 în clasa *“a”* şi 1 în clasa *“b”*;

- 6 în $S_{attrib1=0}$ din care 2 în clasa *„b”* şi 4 în clasa *“c*.

Câştigul informaţional în raport cu fiecare atribut este calculat astfel:

$$Gain(S, atrib_1) = Entropy(S) - \frac{6}{12} Entropy(S_{attrib_1=0}) - \frac{6}{12} Entropy(S_{attrib_1>0})$$
$$= 1{,}5545 - 0{,}5 * 0{,}6498 - 0{,}5 * 0{,}9182 = 0{,}7705$$

$$Gain(S, atrib_2) = Entropy(S) - \frac{8}{12} Entropy(S_{attrib_2>0}) - \frac{4}{12} Entropy(S_{attrib_2=0})$$
$$= 1.5545 - 0.66 * 1.2987 - 0.333 * 0.811278 = 0.418296$$

$$Gain(S, atrib_3) = Entropy(S) - \frac{9}{12} Entropy(S_{attrib_3>0}) - \frac{3}{12} Entropy(S_{attrib_3=0})$$
$$= 1.5545 - 0.75 * 0.991076 - 0.25 * 0.0 = 0.811278$$

$$Gain(S, atrib_4) = Entropy(S) - \frac{8}{12} Entropy(S_{attrib_4>0}) - \frac{4}{12} Entropy(S_{attrib_4=0})$$
$$= 1.5545 - 0.666 * 1.4056 - 0.333 * 1.0 = 0.284159$$

$$Gain(S, atrib_5) = Entropy(S) - \frac{8}{12} Entropy(S_{attrib_5>0}) - \frac{4}{12} Entropy(S_{attrib_5=0})$$
$$= 1.5545 - 0.666 * 1.4056 - 0.333 * 0.0 = 0.617492$$

$$Gain(S, atrib_6) = Entropy(S) - \frac{8}{12} Entropy(S_{attrib_6>0}) - \frac{4}{12} Entropy(S_{attrib_6=0})$$
$$= 1.5545 - 0.666 * 1.561278 - 0.333 * 1.0 = 0.180315$$

La sfârşit, după ce calculăm câştigul informaţional şi aplicăm un prag de 0.5 vom lua în considerare doar atributele $atrib_1$, $atrib_3$, $atrib_5$. Astfel discretizarea pe baza entropiei poate reduce dimensiunea setului de atribute utilizând informaţiile despre clase şi permite selectarea celor mai bune atribute în raport cu acele clase.

4. Temă laborator 3

Problemă: Pornind de la fișierul rezultat la tema Extragerea Trăsăturilor să se calculeze câștigul informațional pentru fiecare atribut din fișier. Programul care trebuie implementat are 2 etape:

1. Etapa de calcul a relevanței unui atribut. În această etapă se parcurge fișierul de la tema „Extragerea Trăsăturilor” și pentru fiecare atribut în parte se va calcula câștigul informațional care este obținut în cazul în care se împarte setul de date folosind acel atribut. Pentru calculul câștigului informațional se va folosi ecuația (2) și se va considera că fiecare atribut are reprezentare binară („0” dacă atributul nu există și „1” dacă atributul există indiferent de

numărul de apariții ale atributului în documentul curent). În această etapă rezultatele se vor salva într-un fișier în care, în cadrul secțiunii @attribute se va trece și valoarea obținută pentru acel atribut. În această etapă secțiunea @data nu se modifică (se va copia din vechiul fișier în noul fișier). Pentru exemplul de mai sus fișierul poate arăta astfel (considerăm primele 8 exemple ca fiind de antrenare și ultimele 4 de testare):

@attribute $atrib_1$ 0.7705
@attribute $atrib_2$ 0.418296
@attribute $atrib_3$ 0.811278
@attribute $atrib_4$ 0.284159
@attribute $atrib_5$ 0.617492
@attribute $atrib_6$ 0.180315
@data
1:1 2:5 3:7 5:1 6:1 # a # train
1:1 2:2 5:7 6:5 # b # train
3:1 4:7 5:1 # c # train
1:1 2:3 3:7 4:1 6:1 # a # train
1:2 2:1 3:8 # a # train
5:10 6:7 # b # train
3:7 4:10 5:7 # c # train
3:5 4:3 5:8 6:1 # c # train
1:2 2:2 3:5 4:2 6:1 # a# test
2:1 4:1 5:9 6:5 # b # test
2:1 3:4 4:1 5:9 6:2 # c # test
1:1 2:4 3:6 4:1 # a # test

2. Etapa propriu-zisă de selecție. Considerăm ca și intrare fișierul rezultat la etapa anterioara și se va alege (impune) un prag fix. În acest caz pragul poate reprezenta numărul de atribute pe care dorim să îl obținem (de exemplu 100 atribute și atunci selectăm primele 100 de atribute care au obținut cel mai bun câștig informațional) sau pragul poate reprezenta valoarea câștigului informațional, valoare peste care considerăm că v-om păstra atributul (ca și în exemplul de mai sus). În această etapă se parcurge dicționarul de cuvinte (cel din laboratorul Extragerea trăsăturilor) și se vor păstra doar atributele care respectă pragul. După stabilirea noului dicționar (de dimensiune mult mai mică decât dicționarul original) se vor reface toți vectorii de documente astfel încât aceștia să respecte noua dimensiune a dicționarului și pe aceeași poziție în vectorul de document să fie reprezentat frecvența aceluiași cuvânt din dicționar. Vectorii rezultați se vor salva în fișier. De data aceasta vor rezulta 2 fișiere unul care conține doar vectorii marcați ca fiind pentru „Training” și unul care va conține doar vectorii marcați cu „Testing”. Noii vectori vor fi salvați tot în secțiune marcată cu atributul ”@data” Fișierele rezultate dacă considerăm pragul de 0.5 vor fi:

Fișierul de antrenare:

@attribute $atrib_1$ 0.7705
@attribute $atrib_3$ 0.811278
@attribute $atrib_5$ 0.617492

@data
1:1 2:7 3:1 # a
1:1 3:7 # b
2:1 3:1 # c
1:1 2:7 # a
1:2 2:8 # a
3:10 # b
2:7 3:7 # c
2:5 3:8 # c

Fișierul de test:

@attribute $atrib_1$ 0.7705
@attribute $atrib_3$ 0.811278
@attribute $atrib_5$ 0.617492

@data
1:2 2:5 # a
3:9 # b
2:4 3:9 # c
1:1 2:6 # a