PARTEA A III-A. CLASIFICARE

Prediction is very difficult, especially if it's about the future.
Niels Bohr

5. ALGORITMI DE CLASIFICARE. GENERALITĂȚI

5.1. Introducere

Clasificarea datelor reprezintă procesul prin care unui tuplu i se atribuie una sau mai multe etichete dintr-o mulțime de etichete dată. Clasificarea este un proces în două etape. În prima etapă se va construi un clasificator care va descrie un set predefinit de date sau concepte. Această etapă se numeşte şi etapă de învățare sau etapă de antrenare, în care un algoritm de clasificare construieşte clasificatorul învățând pe baza unor tupluri de date (set de antrenament) şi etichetele asociate claselor acestora. Un tuplu X este reprezentat de un vector de atribute n-dimensional $X = (x_1,x_2,...,x_n)$ care reprezintă măsurători efectuate în cadrul tuplului asupra a n atribute $A_1, A_2, ..., A_n$. Se presupune că fiecare tuplu aparține unei clase, care la rândul ei, are o etichetă de clasă. Eticheta clasei este o valoare discretă şi nu este ordonată. În contextul clasificării, tuplurile se numesc exemple, instanțe sau obiecte. Deoarece eticheta clasei este dată pentru fiecare tuplu în etapa de antrenare, acest pas este un caz de învățarea supervizată.

Astfel, prima etapă a procesului de clasificare poate fi văzută ca şi învățarea unei funcții de mapare $y=f(X)$ care poate prezice pentru un tuplu X dat eticheta y a unei clase. Această mapare este reprezentată sub forma unor reguli de clasificare, arbori de decizie sau formule matematice.

A doua etapă în procesul de clasificare o reprezintă utilizarea modelului elaborat în prima etapă, în clasificarea unui set de test. Pentru a măsura eficiența clasificatorului trebuie efectuată o aproximare a acurateței de clasificare a acestuia. Dacă am măsura acuratețea de clasificare pe baza setului de antrenament valorile obținute ar fi foarte optimiste iar efectul de „over-fit" poate apărea (supraînvățare – algoritmul învață unele anomalii existente în setul

de antrenament care nu pot fi aplicate cazului general). Din acest motiv se creează un set de test care constă din tupluri şi clasele asociate acestora, altele decât cele prezente în setul de antrenament. Astfel, acurateţea clasificatorului este calculată ca procentaj al numărului de tupluri corect clasificate de către clasificator. Metrici pentru evaluarea acurateţei clasificatorilor au fost prezentate în secţiunea 2.6.

În secţiunile următoare vom prezenta o taxonomie posibilă pentru algoritmii de clasificare, prezentând pentru fiecare categorie clasificatorii reprezentativi pentru acea categorie.

5.2. Algoritmi stohastici

Cel mai reprezentativ clasificator stohastic este clasificatorul bayesian. El prezice cu o anumită probabilitate apartenenţa la o clasă dată, cum ar fi de exemplu, probabilitatea ca un document să facă parte dintr-o anumită clasă dată. Clasificarea bayesiană se bazează pe teorema lui Bayes, şi este descrisă în secţiunea 5.2.1. Studiile care au comparat algoritmi de clasificare au identificat clasificatorul bayesian simplu cunoscut sub numele de clasificator bayesian naiv (Naïve Bayes Classifier) ca având, în practică, performanţe bune atunci când este aplicat la seturi de date mari, cum este cazul de faţă, unde există multe documente, fiecare reprezentat prin multe atribute.

Clasificatorul bayesian simplu pleacă de la premisa „naivă" că atributele pentru o anumită clasă sunt independente unele de altele. Această prezumţie se numeşte independenţă condiţională de clasă. În cadrul acestui clasificator se porneşte de la această prezumţie, pentru a simplifica calculele.

5.2.1. Clasificarea bayesiană

Fie Y o variabilă pentru o clasă (categorie) care poate lua valorile $\{y1, y2, ..., y_m\}$.

Fie X o instanţă a unui vector cu n atribute $<x_1, x_2, ..., x_n>$ şi x_k o valoare posibilă pentru X şi x_{ij} o valoare posibilă pentru x_i. Pentru clasificarea de tip Bayes calculăm probabilităţile $P(Y = y_i | X = x_k)$ pentru $i = \overline{1,m}$. Asta ar însemna calcularea tuturor probabilităţilor pentru fiecare categorie, pentru fiecare instanţă posibilă din spaţiul de instanţe – ceea ce este foarte greu de calculat pentru un set rezonabil de date.

Practic, pentru a determina categoria lui x_k, trebuie să determinăm pentru fiecare y_i probabilitatea:

$$P(Y = y_i | X = x_k) = \frac{P(Y = y_i)P(X = x_k | Y = y_i)}{P(X = x_k)} \quad (5.1)$$

Probabilitatea $P(X = x_k)$ poate fi determinată, deoarece mulțimea categoriilor este completă și disjunctă. Rezultă imediat relația de echilibru de mai jos:

$$\sum_{i=}^{m} P\ Y = y_i | X = x_k) = \sum_{i=}^{m} P\ Y = y_i)P(X = x_k | Y = y_i) = \quad (5.2)$$

aşadar:

$$P(X = x_k) = \sum_{i=1}^{m} P(Y = y_i)P(X = x_k | Y = y_i) \quad (5.3)$$

Probabilitatea $P(Y = y_i)$ poate fi uşor aproximată având în vedere faptul că dacă n_i exemple din D se regăsesc în y_i atunci $P(Y = y_i) = n_i / |D|$, unde D reprezintă mulțimea documentelor din setul de antrenament.

Probabilitatea $P(X = x_k | Y = y_i)$ trebuie estimată (deoarece există 2^n posibile instanțe pentru a calcula probabilitatea). De aceea, dacă presupunem că atributele unei instanțe sunt independente (condițional independente), atunci:

$$P(X|Y) = P(x_1, x_2, ..., x_n | Y) = \prod_{i=1}^{n} P(x_i | Y) \quad (5.4)$$

Astfel, trebuie să calculăm doar $P(x_i | Y)$ pentru fiecare pereche posibilă "valoare atribut"-"categorie".

Dacă Y şi toate X_i sunt binare, atunci trebuie să calculăm doar $2n$ valori:

$P(x_i = true | Y = true)$ şi $P(x_i = true | Y = false)$ pentru fiecare X_i

$P(x_i = false | Y) = 1 - P(x_i = true | Y)$

față de 2^n valori, dacă nu am presupune independența atributelor.

Practic, dacă setul de date D conține n_k exemple din categoria y_k şi n_{ij_k} din aceste n_k exemple au a j-a valoare pentru atributul x_i pe x_{ij} atunci estimăm că:

$$P(X_i = x_{ij} | Y = y_k) = \frac{n_{ij_k}}{n_k} \quad (5.5)$$

TEXT MINING
Tehnici de clasificare şi clustering al documentelor

Radu G. Creţulescu, Daniel I. Morariu