Data Mining

Lab 6:

Association rules
Regression models

---

1. Association rules

Exemplu (problema cosului de cumparaturi). Se considera un set de tranzactii (T1,T2,…,Tn), fiecare tranzactie continand un set de produse achizitionate. De exemplu:

T1: {paine, lapte, apa}
T2: {paine,carne, apa}
T3: {paine, unt, carne, apa}
T4: {fructe, apa}

Se pune problema identificarii produselor care sunt frecvent cumparate impreuna si a unor reguli de asociere de forma IF paine AND apa THEN carne.

Regulile de asociere sunt de forma IF A THEN B. Termenul A joaca rol de antecedent, iar B rol de consecinta (totusi nu exprima relatii de cauzalitate ci doar de corelare).

Dintr-un set de tranzactii pot fi extrase numeroase reguli – e necesar sa poata fi evaluate relevant lor pentru a fi ierarhizate.

Pentru evaluarea relevantei unei reguli se folosesc cel putin doua marimi:

- Suport (support): supp(A->B) = numarul de tranzactii ce contin pe A si B/ numarul total de tranzactii
- Incredere (confidence): conf(A->B)=numarul de tranzactii ce contin pe A si B/ numarul de tranzactii ce contin pe A

Exemplu: IF paine si apa THEN carne

A={paine, apa}, B={carne}
Supp(A->B)=2/4=0.5
Conf(A->B)=2/3=0.6

Obs: pe langa aceste masuri se folosesc si indicatori ai noutatii regulii (cat este de interesanta sau neobisnuita regula); un exemplu de astfel de indicator este cel denumit lift:

Lift(A->B)=prob(A,B)/(prob(A)prob(B))

(the probability can be estimated as the relative frequency)

The rule is interesting if lift is large. If lift is close to 1 this suggest that A and B are not correlated thus one cannot extract useful association rules A->B

Exemplu: R=IF paine AND carne THEN apa

Conf(R)=2/2=1
Lift(R)=0.5/(0.5*1)=1

Algoritm de extragere a regulilor de asociere din date (algoritmul APRIORI)

Date de intrare: set de tranzactii (fiecare tranzactie contine o lista de entitati)
Parametri de control:

- Prag pentru suport minim (ex: 0.2)
- Nivel minim de incredere (ex: 0.9)

Structura generala algoritm:

Pas 1: identificare subseturi cu support semnificativ (mai mare decat pragul minim) – "frequent itemsets"; identificarea acestor subseturi se bazeaza pe:

- Identificarea subseturilor frecvente cu un singur element (lista L_1)
- FOR k=1,K DO construirea listei L_k cu subseturi frecvente avand k elemente pornind de la lista L_{k-1} (subseturi frecvente cu k-1 elemente)

Pas 2: construirea regulilor prin partitionarea subseturilor identificate la pasul 1 in doua parti (o parte pentru antecedentul regulii si o parte pentru consecinta); se retin doar regulile care au nivelul de incredere mai mare decat pragul

Exercitiul 1.

a) Deschideti in Weka setul de date supermarket.arff
b) Determinati reguli de asociere folosind Associate->Apriori si valorile implicite ale parameterilor
c) Aplicati acelasi algoritm pentru alte valori ale pragului pentru support (lowerBoundMinSupport=0.2) si pentru incredere (minMetric=0.75).

2. Modele de regresie

2.1. Regresie liniara

In modelele liniare dependent dintre variabila(variabilele) prezise si cele predictoare este descrisa printr-o functie liniara de forma Y=WX. Parametrii modelului (elementele vectorului/matricii W) se determina pornind de la date folosind o tehnica de minimizare a sumei patratelor erorilor

Exercitiul 2.

a) Deschideti in Weka fisierul autoPrice.arff
b) Utilizati Class->Functions->SimpleLinearRegression pentru a determina o dependenta liniara simpla intre atributul de iesire (pret) si cel mai relevant dintre atributele de intrare. Analizati valorile corespunzatoare coeficientului de corelatie si erorii Correlation Coeficient si Mean Absolute Error.
c) Utilizati Class->Functions->LinearRegression pentru a determina o dependenta liniara simpla intre atributul de iesire (pret) si cel mai relevant dintre atributele de intrare. Analizati valorile corespunzatoare coeficientului de corelatie si erorii Correlation Coeficient si Mean Absolute Error.

2.2. Regresie neliniara

Exercitiul 3. (tot pentru fisierul autoPrice.arff)

d) Utilizati Class->Functions->MultilayerPerceptron cu valorile implicite ale parametrilor. Analizati valorile corespunzatoare coeficientului de corelatie si erorii Correlation Coeficient si Mean Absolute Error.

e) Utilizati Class->Functions->RBF Network cu valorile implicite ale parametrilor. Analizati valorile corespunzatoare coeficientului de corelatie si erorii Correlation Coeficient si Mean Absolute Error.

f) Identificati in categoria Class->Trees varianta care permite construirea unui arbore de regresie

Exercitiul 4. Aplicati prelucrarile de la Exercitiile 2 si 3 in cazul setului de date autoMPG.arff si analizati diferentele (in particular in ceea ce priveste arborii de regresie).