Lab 7: Data Mining.
Serii temporale
Metode de tip ansamblu

---

1. Serii temporale

Analiza seriilor temporale are ca scop să modeleze şi să explice dependenţa unor date de momente de timp successive. Exemple tipice de serii temporale sunt: temperatura inregistrată zilnic, curs de schimb valutar, preţul unor acţiuni etc.

Principalele prelucrari care pot fi efectuate asupra unei serii de timp sunt:

- *Pre-procesare* (de exemplu, transformarea seriei prin normalizare sau standardizare, completarea valorilor absente prin interpolare, eliminarea zgomotului prin netezire, eliminarea tendinţei prin calcularea diferenţelor dintre elemente succesive etc)
- *Predicţie*: estimarea valorilor ulterioare din serie pe baza valorii curente şi a celor anterioare (folosind un model care descrie dependenţa valorii curente din serie de valorile anterioare).

Un proces de predicţie este caracterizat prin:

- *Intrare:* datele de intrare sunt valori anterioare din serie
- *Iesire:* rezultatul reprezintă valoarea/valorile următoare din serie
- *Model:* un model de regresie care descrie legătura dintre valoarea curentă a seriei şi valorile anterioare (numărul de valori anterioare despre care sa consideră că influenţează valoarea curentă este denumit intârzierea seriei (*time-lag*))

Considerăm seria X_1, X_2,…., X_n şi intârzierea T. Deci valoarea curenta X_i depinde de valorile X_{i-1},X_{i-2},…,X_{i-T}. Prin urmare secventa de valori din serie poate fi transformată într-un alt set de date în care sunt T atribute predictor si un atribut prezis:

Atribute predictor	*Atribut prezis*
X_1 X_2 … X_i … X_T	X_{T+1}
X_2 X_3 … X_{i+1} … X_{T+1}	X_{T+2}
…	
X_{n-T} X_{n-T+1} … X_{n-i} … X_{n-1}	X_n

Folosind acest set de date se poate construi un model de regresie (in aceeaşi manieră ca pentru date care nu sunt temporale). Una dintre principalele dificultăţi este alegerea adecvată a valorii T.

Exercitiul 1.

a) Deschideti fisierul airlines.arff (conţinând nr de pasageri ai unei companii aeriene înregistrat lunar in perioada 1949 – 1960)
b) Construiţi un nou set de date folosind o întârziere T=12. *Indicatie:* utilizaţi Weka pt eliminarea atributului corespunzător datei şi Excel (sau un limbaj de programare) pt construirea noului set de date
c) Aplicaţi un model de regresie pentru noul set de date şi analizaţi rezultatele obţinute

Exercitiul 2.
(doar pt versiune Weka >=3.7.3)

a) Instalati pachetul Time Series Forecasting utilizand Weka GUI Chooser ->Tools->Package manager si selectand pentru instalare timeSeriesForecasting
b) Deschideti fisierul airlines.arff
c) Preziceti urmatoarele 6 valori utilizand unul dintre urmatoarele modele: (i) linear regression; (ii) multilayer perceptron; (iii) random forests. *Indicatie*: selectia modelului se realizeaza utilizand panelul Advanced Configuration->Based Learner

Obs: detalii privind pachetul TimeSeriesForecasting pot fi gasite la http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka.

(pentru Weka 3.8.1) se utilizează Forecast (Basic configuration) şi se specifică doar Number of time units to forecast (6 dacă se doreşte estimarea următoarelor 6 valori).

2. Metode de tip ansamblu (ensemble models)

Sunt meta-modele care se obţin din câteva modele de bază antrenate pe acelaşi set sau pe seturi diferite de antrenare. Există mai multe variante de a construi modele de tip ansamblu:

- Utilizând modele bazate pe algoritmi diferiţi antrenati pe acelaşi set de date (e.g. *bucket of models*)
- Utilizând modele bazate pe acelaşi algoritm dar antrenate pe seturi diferite de date (e.g. *bagging* and *boosting*)
- Utilizând diferite modele şi împărţind setul de date (e.g. *stacking*)

Exercitiul 3. Utilizand Weka Experimenter comparaţi performanţa următoarelor metamodele: Vote, Bagging, Random Forest, AdaBoost si Stacking pt seturile de date: iris.arff, glass.arff

a) Utilizaţi valorile implicite ale parametrilor
b) Imbunătăţiţi comportamentul pt Vote, Bagging si AdaBoost înlocuind clasificatorul de bază cu alt clasificator.