

ANALIZA REGRESIVĂ: CREAREA UNUI MODEL SIMPLU DE CLASIFICARE ÎN JAVA FOLOSIND WEKA

Valeria LEAȘCENCO*

Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Informatică și Microelectronică,
Departamentul Informatică și Ingineria Sistemelor, grupa CR-181, Chișinău, Republica Moldova

*Autorul corespondent: Valeria Leascăncu, leascencovaleria.vl@gmail.com

Rezumat: În articol este prezentată una din metodele de analiză a datelor – analiza regresivă. Ilustrarea acestei metode de analiză a datelor este realizată în baza unui exemplu, care constă în determinarea costului unei plăci de memorie RAM, în dependență de mai mulți factori. Modelul analizei regresive este utilizat pentru a prezice valoarea unei variabile dependente, pe baza valorilor cunoscute a mai multor parametri independenți.

Cuvinte cheie: analiza inteligentă, metoda regresivă, Weka.api, Java.

Introducere

În domeniul tehnologiilor moderne, problema analizei datelor tehnice crește tot mai mult în popularitate. Probabil ați auzit că marile companii precum Google sau Facebook, colectează miliarde de indicatori diferiți despre utilizatorii lor, de unde și apare întrebarea complet logică care se referă la modul în care aceste companii urmează să folosească informația colectată. Un alt exemplu ar fi compania Walmart (domeniul comercial), care folosește cele mai avansate tehnologii pentru analiza datelor, aplicând cu succes rezultatele obținute în scopul dezvoltării afacerilor. Aproape fiecare companie modernă folosește extragerea datelor, iar cele care ignoră o astfel de posibilitate curând pot fi într-un dezavantaj foarte mare.

Analiza inteligentă a datelor

În principiu, analiza inteligentă a datelor reprezintă transformarea volumelor mari de date brute în scheme practice, structuri și reguli. Analiza datelor poate fi divizată în 2 tipuri:

- directă – prognozarea unor indicatori specifici, de exemplu, prognoza valorii de vânzare a unui calculator, pe baza informațiilor despre prețurile calculatoarelor dintr-o anumită categorie (business);
- indirectă – crearea unor grupuri de date sau căutarea unor structuri sau modele specifice într-un set de date existente, de exemplu, determinarea unui grup de studenți. Fiecare colectare a datelor despre studenți implică analiza inteligentă a datelor, precum cadrele universitare încearcă să obțină informații despre fiecare student înmatriculat la universitate, pentru o utilizare practică ulterioară.

Analiza inteligentă a datelor, din punctul de vedere al scopului urmărit în acest articol, a apărut la mijlocul anilor 90, când dezvoltarea tehnologiei informaționale computerizate a atins un nivel destul de înalt, iar costul sistemelor de alimentare și de calcul a scăzut, astfel încât companiile să poată să-și permită de sine stătător să efectueze analiza datelor, fără să recurgă la serviciile altor centre de date.

Desigur că metodele de analiză a datelor nu sunt la fel de simple ca efectuarea unei funcții pe un eșantion de date din careva tabele electronice, dar nu sunt și atât de complicate încât nu ar putea fi utilizate independent. De exemplu, poate fi creat un model de analiză a datelor cu o eficiență de 90%, având doar 10% din cunoștințele unui expert în domeniul analizei datelor.

Principalul scop al analizei inteligente a datelor constă în crearea unui model, care permite interpretarea și utilizarea efectivă a datelor pe care le aveți la moment și acele date pe care le veți obține în viitor.

Weka

Analiza inteligentă a datelor nu este domeniul exclusiv al companiilor mari sau al softului scump. Weka este un produs soft al Universității Waikato, apărut în 1997 și a fost scris pe limbajul de programare Java, ce oferă utilizatorului o interfață grafică pentru lucrul cu fișierele de date și generarea rezultatelor vizuale sub formă de tabele sau grafice. La rândul său, Weka poate fi integrată în orice altă bibliotecă, în propriile aplicații pentru automatizarea datelor din partea serverului, utilizând API-ul standard. În continuare vom analiza un exemplu concret de analiză a datelor prin metoda regresivă.

Metoda regresivă de analiză

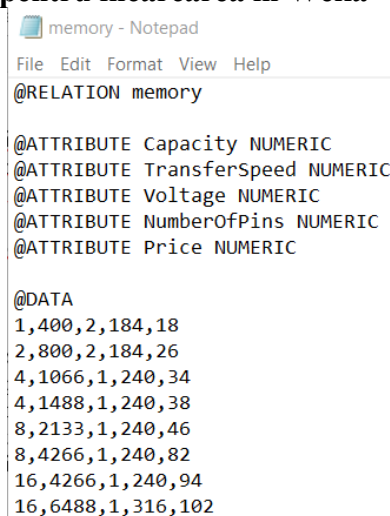
Vom aplica modelul analizei regresive pentru determinarea prețului unei plăci de memorie RAM și vom analiza un exemplu concret. Vom crea un tabel în care vom indica parametrii efectivi ai plăcii de vânzare și vom încerca să estimăm valoarea plăcii memoriei RAM (Tabelul 1).

Tabelul 1

Tabelul cu datele despre obiect

Capacity (GB)	Transfer speed (MHz)	Voltage (V)	Number of pins (buc.)	Price (\$)
1	400	2	184	18
2	800	1	184	26
4	1066	1	240	34
4	1488	1	240	38
8	2133	1	240	46
8	4266	1	240	82
16	4266	1	240	94
16	6488	1	316	102

Crearea setului de date pentru încărcarea în Weka



```

memory - Notepad
File Edit Format View Help
@RELATION memory

@ATTRIBUTE Capacity NUMERIC
@ATTRIBUTE TransferSpeed NUMERIC
@ATTRIBUTE Voltage NUMERIC
@ATTRIBUTE NumberOfPins NUMERIC
@ATTRIBUTE Price NUMERIC

@DATA
1,400,2,184,18
2,800,2,184,26
4,1066,1,240,34
4,1488,1,240,38
8,2133,1,240,46
8,4266,1,240,82
16,4266,1,240,94
16,6488,1,316,102
    
```

Figura 1. Fișierul cu datele despre obiect

Pentru a încărca datele în Weka, acestea trebuie convertite într-un format care poate fi înțeles de acest pachet software, în formatul ARFF (Attribute-Relation File Format), care determină mai întâi tipul de date care se încarcă, apoi indică datele. În fișierul format ARFF se specifică numele și tipul de date pentru fiecare coloană a tabelului, apoi datele reale în rânduri (Figura 1). Modelele de analiză regresivă utilizează doar două tipuri de date: NUMERIC și DATA.

Pentru a introduce fișierul de date în Weka, este necesar să alegem opțiunea *Explorer*. În rezultat se va deschide fereastra *Preprocess*, unde este necesar să verificăm datele. În partea stângă sunt arătați parametrii obiectelor (*Attributes*), care corespund titlurilor coloanelor din tabelul sursă, numărul de obiecte (*Instances*). Dacă vom selecta coloana *TransferSpeed*, în panelul din dreapta se vor afișa date statistice suplimentare despre această coloană (Figura 2).

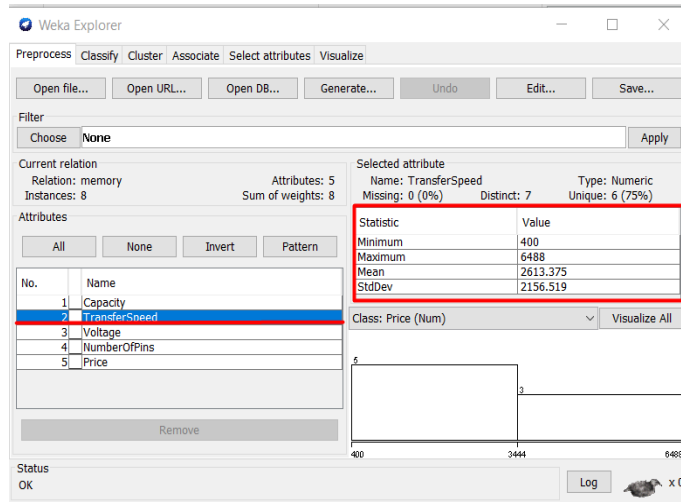


Figura 2. Datele statistice caracteristice unei coloane

Crearea modelului regresiv în Weka

Pentru a crea modelul este necesar să deschidem fereastra *Classify*, după care se selectează tipul modelului pentru analiză, modul de analiză a datelor și ce model să construiască:

1. Facem clic pe butonul *Choose* și deschidem fereastra *functions*.
2. Alegem opțiunea *LinearRegression*.

În așa mod a fost selectat modelul analizei regresive. După alegerea modelului, trebuie să indicăm Weka, care date trebuie folosite pentru crearea ei. În cazul analizei regresive este necesară utilizarea opțiunii *Use training set*. În acest caz Weka va crea modelul pe baza datelor din fișierul ARFF încărcat (Figura 3). Etapa finală a creării modelului constă în selectarea unei variabile dependente (coloana în care se va afla valoarea necunoscută).

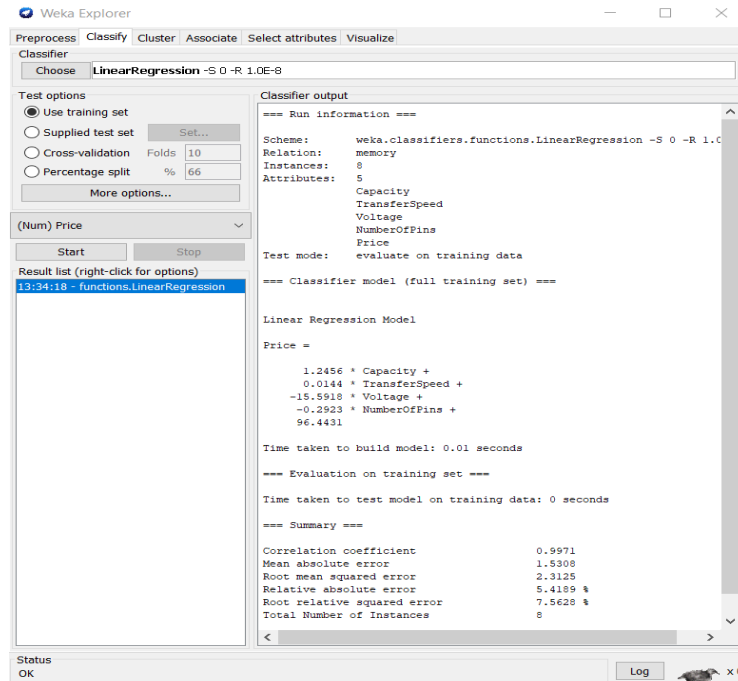


Figura 3. Modelul analizei regresive în aplicația Weka

Interacțiunea limbajului de programare Java și Weka API

Pentru determinarea prețului plăcii de memorie RAM, pentru acest model al analizei regresive vor fi utilizate: limbajul de programare Java, mediul de dezvoltare *Eclipse* și biblioteca Weka API.

```

1 package weka_api; //biblioteca Weka
2 *import java.io.BufferedReader; //citirea efectivă a simbolurilor, masivelor și stringurilor
7 public class weka {
8     public static void main(String args[]) throws Exception {
9         //încărcarea datelor din fisier în obiectul "data"
10        Instances data = new Instances(new BufferedReader(new FileReader("D:\\weka\\memory.arff")));
11        data.setClassIndex(data.numAttributes() - 1);
12        //inițierea și construirea modelului regresiv pe baza datelor din obiectul "data"
13        LinearRegression model = new LinearRegression();
14        model.buildClassifier(data);
15        System.out.println(model);
16        //utilizarea modelului pentru prognozarea prețului
17        Instance mem = data.lastInstance();
18        double price = model.classifyInstance(mem);
19        System.out.println("Memory: (" + mem + "): " + price);
20    }
21 }
22

```

Console Output:

```

<terminated> weka [Java Application] C:\Program Files\Java\jdk-13.0.2\bin\javaw.exe (Mar 14, 2020, 12:43:59 PM)

Linear Regression Model
Price =
  1.2456 * Capacity +
  0.0144 * TransferSpeed +
 -15.5918 * Voltage +
 -0.2923 * NumberOfPins +
  96.4431
Memory: (16,6488,1,316,102): 102.00000008992367

```

Figura 4. Analiza prețului unei plăci de memorie RAM prin intermediul Java

A fost creată o clasă nouă în care au fost introduse toate elementele necesare descrise în cod. Până acum sarcina era rezolvată fără Java, doar prin utilizarea aplicației Weka, în continuare se arată cum sunt introduse datele pentru aflarea și obținerea rezultatului (Figura 4).

Concluzii:

În articol a fost ilustrată analiza prețului unei plăci de memorie RAM, prin intermediul analizei regresive, astfel înțelegând modul de funcționare al modelului regresiv și modul de utilizare al limbajului de programare Java, pentru obținerea rezultatului. Modelul analizei regresive poate arăta tendințele pentru dezvoltarea unui proces mult mai efektiv. Cu ajutorul pachetului Weka este posibilă efectuarea analizei diferitor date cu diverse probleme, precum ar fi: analiza costului sau veniturii pentru diverse produse fie hard, fie soft; analiza timpului de elaborare al unui produs etc. La fel, am demonstrat utilitatea limbajului de programare Java, ce oferă posibilitatea de calcul efectiv, printr-o implementare accesibilă.

Bibliografie:

1. Майкл Абернети *Интеллектуальный анализ данных с помощью программного пакета Weka* [online]. [accesat 26.02.2020]. Disponibil: <https://www.ibm.com/developerworks/ru/library/os-weka1/index.html#artrelatedtopics>
2. Eibe Frank, Mark Hall, Peter Reutemann, Len Trigg *Use Weka in your java code* [online]. [accesat 26.02.2020]. Disponibil: <https://waikato.github.io/weka-wiki/use-weka-in-your-java-code/>