Studiu comparat asupra tehnicilor de data mining utilizate în rezolvarea problemelor de regresie si clasificare

Ec. Valentin MILITARU
Catedra de Informatica Economica, A.S.E. Bucuresti

In today's economic context, the corporations and the business environment in general are producing data in enormous quantity and on a daily basis. Data mining defines the process of extracting the knowledge hidden in large volumes of data and the regression and classification tasks are perhaps the most used in such purpose.
This whitepaper presents four widely used data mining techniques (Naïve Bayes, k-Nearest Neighbor, Neural Networks and Decision Trees) and underlines twelve characteristics that differentiate these techniques. The table at the end of this material summarizes the results of this comparative analysis.
Keywords: *data mining, data analysis, data mining techniques, regression, classification.*

În contextul economic actual, zi de zi, corporatiile si mediul de afaceri în ansamblu produc mari volume de date care sunt utilizate în operatiuni curente. În procesul de analiza, informatia care poate fi extrasa din bazele de date poate fi exploatata în continuare pentru construirea unor modele previzionale, pentru identificarea unor relatii între înregistrarile continute în baza de date, pentru clasificarea acestor înregistrari sau fie si numai pentru realizarea unei descrieri a continutului bazei de date. Tehnicile de data mining permit extragerea informatiilor si realizarea de previziuni pornind de la date istorice.

Clasificarea tehnicilor de data mining

Tehnicile de *data mining* au fost grupate în trei categorii, în functie de tipul de probleme pe care le pot modela:

a) *Clasificarea si regresia* reprezinta cea mai larga categorie de aplicatii, constând în construirea de modele pentru previzionarea apartenentei la un set de clase (clasificare) sau a unor valori (regresie). Exista câteva tehnici dedicate rezolvarii problemelor de clasificare si regresie, dintre care arborii decizionali, tehnica Bayes, retelele neuronale si k-NN se bucura de o larga recunoastere.

b) *Analiza asocierilor si succesiunilor*, denumita uneori analiza cosului de cumparaturi; aceasta tehnica genereaza modele descriptive care evidentiaza reguli de corelatie între atributele unui set de date.

c) *Analiza de tip cluster* este o tehnica descriptiva utilizata pentru gruparea entitatilor similare dintr-un set de date sau în egala masura pentru evidentierea entitatilor care prezinta diferentieri substantiale fata de un grup. Tehnicile de grupare în clustere se bazeaza pe algoritmi din sfera retelelor neuronale, algoritmi demografici, k-NN etc.

Pentru rezolvarea problemelor de clasificare si regresie exista o serie de tehnici, iar pentru fiecare tehnica sunt disponibili mai multi algoritmi. Diferenta dintre clasificare si regresie este aceea ca în primul caz, outputul previzionat este apartenenta la o anumita clasa, în timp ce în al doilea caz, outputul estimeaza valoarea unui atribut. Regresia este utilizata în cazurile în care outputul este definit pe un domeniu foarte larg (chiar infinit) de valori si nu trebuie confundata cu notiunea de "regresie liniara" din matematica.

De remarcat ca o problema de regresie poate fi usor transformata într-una de clasificare si invers. În multe cazuri, instrumentele pot fi utilizate pentru rezolvarea ambelor tipuri de probleme.

- **Tehnica Bayes**[1]**.** Mai putin implementata în aplicatiile de explorare a datelor, tehnica Bayes naiva este o metoda de clasificare care îsi datoreaza numele ministrului britanic Thomas Bayes (1702- 1761). Teoria probabi-

[1] În limba engleza, denumirea tehnicii este însotita de adjectivul "naiv": *Naïve-Bayes*

litatilor a lui Bayes, pe care se bazeaza si tehnica ce-i poarta numele, a fost publicata postum, abia în 1764. Bayes este o tehnica de clasificare cu potential atât predictiv, cât si descriptiv. Ea permite analiza relatiei dintre fiecare variabila independenta si variabila dependenta, prin calcularea unei probabilitati conditionate pentru fiecare din aceste relatii. Când o noua instanta se doreste a fi clasificata, predictia se realizeaza prin combinarea efectelor variabilelor independente asupra variabilei dependente.

Limitele tehnicii. Pentru instantele care apartin setului de date utilizat la calculul probabilitatilor a priori si al celor conditionale, "predictia" atributului-obiectiv este 100% corecta. Însa pentru instante din afara setului de date de instruire, eficienta algoritmului este puternic afectata de prezenta unor probabilitati conditionate egale sau foarte aproape de zero.

O alta limita a algoritmului provine din asumtia ca între atributele independente din setul de date exista (teoretic) o independenta statistica. Aceasta asumtie sta si la originea adjectivului "naiv" din denumirea algoritmului, având în vedere ca independenta statistica de regula nu se verifica si în practica. Algoritmul este limitat din punct de vedere al inputului la date booleene sau categorice. Dincolo de efortul de preprocesare necesar pentru transformarea datelor cu caracter continuu în intervale valorice, asa cum s-a mai mentionat, operatiunea este de multe ori dependenta de experienta si chiar intuitia analistului, factori subiectivi care vor marca rezultatele explorarii.

Avantajele tehnicii. Tinând seama de faptul ca pentru calculul probabilitatilor nu este nevoie decât de o singura parcurgere a setului de date, algoritmul prezinta avantajul important al unei viteze mari de construire a modelului de clasificare.

Ca avantaj semnificativ, algoritmul prezinta capacitatea de a realiza predictii din informatii partiale. În ciuda sensibilitatii la caracteristici slab reprezentate în setul de date, pentru realizarea unei predictii algoritmul nu are obligatoriu nevoie de toate atributele independente, astfel încât cele identificate de analist a fi irelevante pot fi usor eliminate din algoritm. De fapt, chiar daca nu s-ar cunoaste nimic despre atributele independente, analistul tot ar putea face o predictie (fara pretentia de a fi foarte acurata) numai pe baza probabilitatilor a priori.

Modelul obtinut prin aplicarea algoritmului are si un continut descriptiv, care poate fi util analistului. Probabilitatile conditionate aferente fiecarui atribut independent pot fi utilizate în a descrie legatura dintre acestea si atributul-obiectiv.

- **k-NN**. Tehnica (prescurtare a expresiei engleze *k-Nearest Neighbor*) este predictiva de explorare a datelor utilizata cu precadere în probleme de clasificare. Principiul care sta la baza tehnicii este relativ simplu: o instanta noua este clasificata prin analiza "proximitatii" sale (sau gradului de similitudine) cu alte instante dintr-un set de date cunoscut. k-NN este o tehnica folosita în special pentru clasificarea datelor în categorii multiple, însa poate fi aplicata inclusiv pentru previzionarea unui atribut-obiectiv de natura numerica (continua sau discreta), ca rezultat al unor dependente neliniare. Fie un set de date compus din instante care au urmatoarea structura:
  - n atribute numerice independente {Xi, i=1,n};
  - m atribute booleene sau categorice independente {Aj, j=1,m};
  - un atribut-obiectiv Y, reprezentând variabila dependenta a carui valoare va trebui estimata pentru noile instante.

  Pentru a previziona valoarea atributului-obiectiv al unei instante noi, algoritmul cauta în setul de date *k* înregistrari "apropiate" de acea instanta, pentru care se cunosc valorile lui Y. Predictia este data de media valorilor lui Y aferente "vecinilor" identificati în setul de date.

  Aplicarea conceptului în practica ridica urmatoarele probleme:

  i) Prin ce metoda se stabileste relatia de vecinatate dintre doua instante? Care este metrica utilizata la calculul distantelor dintre doua instante?

  ii) Care este valoarea optima pentru *k*? De câte instante similare este nevoie pentru ca media atributelor-obiectiv sa se constituie

într-o predictie cu grad acceptabil de reprezentativitate?
iii) Ce metoda de calcul a mediei va fi utilizata pentru predictia lui Y? Întrebarea este importanta în special pentru cazurile în care atributul-obiectiv nu este de natura numerica, ci booleana sau categorica.
iv) Care din atributele {Xi} si {Aj} sunt cu adevarat reprezentative pentru predictie? Care din aceste atribute identifica cel mai bine "vecinii" instantei de analizat?

Limitele tehnicii. Timpul de calcul este direct proportional cu numarul de instante din setul de date. Din acest motiv, pentru seturi mari de date se impune ca în etapa de preprocesare, din setul initial de date sa se selecteze un subset de instante cu dimensiuni rezonabile. Algoritmul lucreaza eficient în probleme de clasificare atunci când toate clasele aferente atributului-obiectiv au o reprezentare egala ca pondere în setul de date, fapt care face necesara "îmbogatirea" setului de date original. Algoritmul pe care se bazeaza tehnica k-NN permite doar realizarea unei estimari a valorii atributului-obiectiv, fara a produce informatii suplimentare despre instanta supusa analizei, despre structura setului de date ori despre categoriile de clasificare a atributului-obiectiv.
De cele mai multe ori este dificil de stabilit ce tip de functie estimeaza cel mai bine distanta dintre doua instante. Desi din punct de vedere matematic tehnica permite calculul distantelor si pentru atribute categorice si booleene, în astfel de cazuri metrica devine puternic influentata de transformarile aplicate de analist setului de date în preprocesare. De aceea, k-NN este de preferat a fi utilizata mai mult în situatiile în care pentru toate atributele instantelor se poate aplica aceeasi functie de distanta.

Avantajele tehnicii. Tehnica permite clasificarea în multiple clase si modelarea relatiilor neliniare dintre date (în probleme de predictie). Pentru tehnicile care necesita o etapa de învatare a carei output îl constituie un model predictiv, exista riscul ca acest model sa devina desuet în timp, iar predictiile realizate în baza lui sa piarda din reprezentativitate. În cazul k-NN, modelul îl constituie chiar setul de date, care se presupune ca odata supus analizei, este deja în forma sa cea mai recenta. Chiar daca de multe ori pot aparea dificultati în stabilirea unei metrici eficiente, algoritmul este unul dintre putinele care accepta ca input date de natura diferita (continua, categorica, booleana etc.).

- **Retele neuronale**. Tehnica are la baza doua concepte apartinând domeniului inteligentei artificiale. *Neuronul artificial* reprezinta unitatea de baza pentru prelucrarea informatiei în cadrul calculului neuronal. Prin analogie cu neuronul biologic, el a fost definit ca o unitate ce proceseaza inputuri informationale si genereaza outputuri. *Reteaua neuronala artificiala* reprezinta un ansamblu de neuroni artificiali, legati prin conexiuni. Retelele neuronale sunt sisteme dinamice, al caror comportament poate fi caracterizat prin urmarirea starilor la momente diferite de timp. Starea unei retele la un moment dat este definita de ansamblul nivelurilor de activare a neuronilor si de intensitatile conexiunilor dintre neuroni. În plus fata de acesti parametri ajustabili, o retea este definita si de urmatorii parametri ficsi: configuratia conexiunilor si tipul functiilor de activare.

Limitele tehnicii. Retelele neuronale nu opereaza decât direct asupra variabilelor numerice. Drept urmare, orice variabila nonnumerica din setul de date care se doreste analizat va trebui convertita în variabila numerica înainte de utilizarea sa în instruirea retelei. În cazul problemelor complexe, utilizatorul este pus în situatia de a rezolva un compromis, între a creste numarul de neuroni ascunsi, ceea ce poate conduce la o instruire foarte lenta si a accepta o topologie mai simpla, asociata unei solutii mai putin precise. Pentru seturi de date cu numar mare de atribute, folosirea retelelor neuronale devine nefezabila.
Determinarea numarului de neuroni ascunsi, pentru probleme complexe de clasificare, nu se poate face decât experimental, ceea ce pe de o parte creste substantial timpul alocat cautarii modelului optim de clasificare, iar pe de alta parte lasa calitatea rezultatelor analizei sa depinda de nivelul de experienta al utilizatorului. Absenta componentei descriptive

într-un model generat de o retea neuronala face ca evolutia modelului în etapa de instruire sa fie lipsita de transparenta pentru utilizator. Datorita acestei caracteristici, tehnica este deseori comparata cu o "cutie neagra". Totusi, cea mai suparatoare caracteristica a retelelor neuronale este timpul îndelungat necesar pentru o buna instruire, fapt corelat cu necesitatea existentei unui numar relativ mare de instante în setul de instruire.

Avantajele tehnicii. Reteaua odata instruita poate realiza predictii rapide pentru instante noi. Aceasta caracteristica face ca retelele neuronale sa fie utilizate cu succes în probleme care necesita raspuns în timp real. Pâna în prezent, retelele neuronale reprezinta metoda cea mai eficienta de modelare a unor relatii neliniare. Mai mult, aplicatiile de pâna acum au demonstrat aplicabilitatea acestei tehnici în domenii dificil de modelat, precum vederea electronica sau recunoastere vocala. Spre deosebire de celelalte tehnici de data mining, retelele neuronale nu restrictioneaza output-ul la un singur atribut. Folosind o arhitectura de retea potrivita se pot obtine predictii simultane pentru mai multe variabile, ceea ce poate însemna o eficientizare semnificativa a proceselor de explorare a datelor.

- **Arbori decizionali**. Arborele decizional este o tehnica de explorare a datelor cu potential atât predictiv, cât si descriptiv. Denumirea sa provine de la aceea ca rezultatul se prezinta utilizatorului sub forma unui graf de tip arbore. Output-ul major al unui model bazat pe arbori decizionali este arborele în sine. Procesul de instruire care creeaza arborele este numit inductie. Inductia presupune, ca si în cazul retelelor neuronale, parcurgerea de câteva ori a setului de date de instruire, cu deosebirea ca în cazul arborilor, timpul de instruire si implicit numarul de baleieri ale setului de date este mult mai mic decât la retelele neuronale. Mai precis, numarul de parcurgeri ale setului de instruire este egal cu numarul de niveluri în arbore.

Limitele tehnicii. Majoritatea algoritmilor nu folosesc întregul set de date indicat de utilizator pentru inductie. Pentru acesti algoritmi, construirea arborelui presupune transferul instantelor din setul de date de instruire în memoria RAM. Dimensiunea limitata a memoriei face ca programul sa transfere în RAM numai un subset de date, selectat aleator. În consecinta, gradul de reprezentativitate al modelului construit este determinat de capacitatea aplicatie de a selecta un subset reprezentativ pentru întreg setul de inductie. O critica adusa frecvent arborilor decizionali este aceea ca algoritmii de inductie nu iau în considerare la momentul splitarii efectul pe care respectiva separare o are asupra viitoarelor splitari. În plus, toate separarile se fac secvential, ceea ce determina dependenta fiecarei splitari de cele precedente.

Avantajele tehnicii. Majoritatea algoritmilor care construiesc arbori decizionali pot fi aplicati fara restrictii legate de tipul datelor. Desi variabila dependenta trebuie sa fie de natura numerica (în cazul problemelor de regresie) sau categorica (în cazul problemelor de clasificare), pentru majoritatea algoritmilor, variabilele independente pot lua valori în orice domeniu. Tehnica se caracterizeaza prin capacitate de prelucrare a unor seturi de date cu numar mare de atribute. Exista situatii în care o instanta poate fi descrisa printr-un numar relativ mare de atribute, de ordinul sutelor sau chiar miilor. În astfel de situatii, explorarea prin tehnica arborilor decizionali reprezinta singura alternativa, cei mai multi algoritmi fiind capabili sa trateze seturi de date cu peste 1000 de coloane.

Algoritmii de construire a arborilor decizionali necesita un numar redus de parcurgeri a setului de date utilizat în inductie. Consecinta directa a acestei caracteristici functionale este rapiditatea procesului de inductie si aplicarea eficienta asupra seturilor mari de date.

Forma outputului permite nu numai realizarea de previziuni si clasificari, ci si descrierea relatiilor existente între variabilele independente si variabila dependenta. În plus, forma grafica a outputului faciliteaza analiza relatiilor. Exista aplicatii care permit reprezentarea arborelui sub forma unui set de reguli care, pentru arbori de dimensiuni mari, este mai usor de înteles.

Concluzie

În functie de cazul concret, anumite tehnici

de data mining sunt mai eficiente decât altele, existând chiar situatii în care pentru rezolvarea problemei nu exista decât o unica optiune (de exemplu arborii decizionali sunt singura alternativa viabila pentru analiza seturilor de date cu numar mare de variabile, retelele neuronale reprezinta unica solutie pentru probleme în care outputul are o forma vectoriala etc.). Tabelul 1 concentreaza caracteristicile prezentate pe larg în acest material, pentru fiecare din cele patru tehnici de data mining tratate urmarindu-se un set de 12 caracteristici. S-a notat cu semnul "+" situatia în care tehnica satisface criteriul curent si cu "-" situatia contrarie. De departe, arborii decizionali prezinta cele mai multe avantaje, în timp ce retelele neuronale prezinta gradul cel mai mic de flexibilitate.

Tabelul 1 - Analiza comparativa a celor patru tehnici de data mining

Crite riul de comparatie	Denumirea tehnicii			
	Naïve - Bayes	**k-NN**	**Retele neuronale**	**Arbori decizionali**
1. Rapiditate în etapa de instruire	+	-	-	+
2. Rapiditate în aplicarea modelului	+	+	+	+
3. Instruire eficienta pe seturi largi de date	+	-	-	+
4. Operare eficienta pe seturi de date cu nr. mare de atribute	-	-	-	+
5. Capacitate de generare a unor outputuri complexe (mai multe atribute simultan)	-	-	+	-
6. Capacitate de generare a unor outputuri de natura vizuala	-	-	-	+
7. Output cu potential descriptiv	+	-	-	+
8. Utilizare în probleme de predictie	-	+	+	-
9. Utilizare în probleme de clasificare	+	+	+	+
10. Nu comporta restrictii legate de tipul datelor de input	-	+	-	+
11. Solutia (modelul) nu depinde de experienta utilizatorului	+	+	-	+
12. Transparenta modelului fata de utilizator	+	-	-	+
Total	**(+7) (-5)**	**(+5) (-7)**	**(+4) (-8)**	**(+11) (-1)**

Bibliografie

1. *** - *Critical Features of High performance Decision Trees* - Salford Systems, San Diego (U.S.A.), 2001 (www.salford-systems.com)
2. *** - IBM's Data Mining Technology – *Data Management Solutions* - International Business Machines Corporation, 1996 (www.ibm.com)
3. Bodea, Constanta – Nicoleta – *Inteligenta Artificiala si Sisteme Expert* – Editura Inforec, Bucuresti, 1998
4. Brand, Estelle; Gerritsen, Rob – *Classification and Regression* – DBMS Magazine, Data mining Solutions Supplement, Feb. 1998 (www.dbmsmag.com)
5. Brand, Estelle; Gerritsen, Rob – *Decision Trees* – DBMS Magazine, Data mining Solutions Supplement, Feb. 1998 (www.dbmsmag.com)
6. Brand, Estelle; Gerritsen, Rob – *Naïve-Bayes and Nearest Neighbor* – DBMS Magazine, Data mining Solutions Supplement, Feb. 1998 (www.dbmsmag.com)
7. Brand, Estelle; Gerritsen, Rob – *Neural Networks* – DBMS Magazine, Data mining Solutions Supplement, Feb. 1998 (www.dbmsmag.com)