ARBORII DE DECIZIE O PROVOCARE PENTRU CLASIFICAREA ŞI LUAREA DECIZIILOR ÎN ECONOMIE

LAVINIU AURELIAN BĂDULESCU
Universitatea din Craiova; Facultatea de Automatică, Calculatoare şi Electronică, Secţia Inginerie Software; Craiova, str. G. Fotino , nr. 2, bl. b8, 3, 11, 200481, Dolj; mobil 0722276554;
E-mail: *laviniu_aurelian_badulescu@yahoo.com*

Economists have relied on traditional models when trying to describe how a particular binary choice is related to a set of explanatory variables. There are a variety of "classifiers" like decision trees with superior attributes. Decision tree are used to select the best course of action in situations where you face uncertainty. Many business decisions fall into this category.

1. INTRODUCERE

În mod tradiţional, economiştii s-au bizuit pe modele probabilistice şi logice atunci când au încercat să explice cum este asociată o alegere binară particulară cu o mulţime de variabile care o descriu. Astfel de modele au o atracţie larg răspândită prin aceea că ele pot fi derivate, cel puţin intuitiv, din modele economice de luări de decizii individuale, sunt uşor de estimat, iar rezultatele lor sunt uşor de interpretat.

Însă relativ de curând, specialiştii ştiinţei calculatoarelor din domeniul inteligenţei artificiale şi machine learning au pus în evidenţă alte instrumente de clasificare cu caracteristici superioare modelelor probabilistice şi logice. Aceste instrumentele specifice Data Mining includ algoritmi ca: reţelele neuronale şi arborii de decizie. Astfel de algoritmi induc modele puternic neliniare care sunt considerate mai flexibile decât metodele liniare tradiţionale şi în consecinţă, furnizează o potrivire mai bună la datele empirice. Astfel de tehnici Data Mining sunt mai performante în clasificarea unui obiect individual într-o categorie particulară pe baza atributelor sale individuale.[5]

Arborii de decizie sunt folosiţi pentru a selecta cea mai bună direcţie de acţiune în situaţiile în care apare incertitudinea. Multe decizii de afaceri intră în această categorie. De exemplu, un producător trebuie să decidă cât de mare să fie stocul creat înainte de a ştii precis care va fi cererea. O persoană aflată în litigiu trebuie să aleagă între o înţelegere în afara tribunalului sau riscul unui proces. Un jucător la bursă trebuie să decidă să cumpere înainte de a ştii dacă ceea ce a cumpărat poate fi vândut ulterior pentru a obţine un profit. În toate aceste cazuri, cel care ia decizia întâlneşte un necunoscut care pare să îl facă incapabil să aleagă varianta corectă cu siguranţă absolută.

Chiar dacă cel care ia decizia nu cunoaşte ce efect va avea factorul necunoscut, el are de obicei nişte cunoştinţe despre efectele ce pot să apară şi cum e mai probabil să apară fiecare efect. Această informaţie poate fi utilizată pentru a selecta opţiunea care este cel mai

probabil să producă rezultate favorabile. Arborii de decizie fac uşor de aplicat acest tip de analiză.[2]

Algoritmii de arbori de decizie [6] sunt utilizaţi la explorarea setului de date în problemele de afaceri. Aceasta se realizează adesea prin urmărirea predictorilor şi a valorilor care sunt alese la fiecare ramificare a arborelui.

În Figura 1 se evidenţiază câteva tipuri de aplicaţii economice în care se utilizează algoritmi

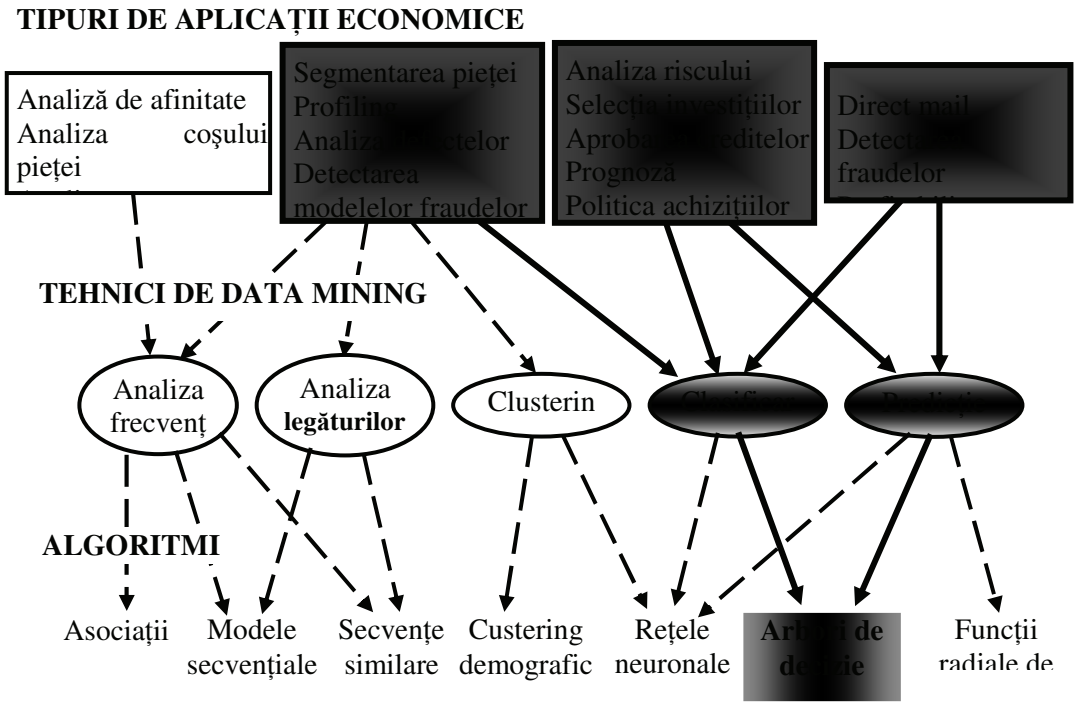


Figura 1.

de Data Mining [3]. Se constată că *algoritmii de arbori de decizie* se folosesc în cazul a două tehnici de Data Mining, respectiv *clasificare* şi *predicţie*. În acelaşi timp aceste două tehnici au utilitate în aproape toate aplicaţiile economice pentru care putem folosi Data Mining. Acest fapt evidenţiază puterea deosebită a algoritmilor de arbori de decizie pentru aplicaţiile economice ce se subsumează aplicării tehnicilor Data Mining.

2. CONSTRUIREA UNUI ARBORE DE DECIZIE DE CALITATE

Să considerăm un arbore de decizie cu M frunze. Acest arbore de decizie corespunde la descompunerea spaţiului caracteristic în M subregiuni nesuprapuse E^1, …, E^M, astfel încât subregiunea E^S corespunde la frunza numărul S (vezi Figura 2). E^S este definit ca produsul cartezian

$$E^S = E_1^S \times E_2^S \times \ldots \times E_n^S, \qquad (1)$$

unde E_j^S este proiecţia lui E^S pe caracteristica numărul j. E_j^S este obţinută în următorul mod. Dacă caracteristica X_j nu este situată pe drumul de la rădăcină la frunza numărul S, atunci E_j^S coincide cu un domeniu al definiţiilor caracteristicii X_j. Altfel, E_j^S este egal cu intersecţia

tuturor subregiunilor caracteristicii X_j, care a fost întâlnită pe drumul de la rădăcină la frunza numărul *S*.
Fie o mulțime a unor observații experimentale

$$Data = \left(x^i , y^i \right), \quad i = 1,\ldots,N . \qquad (2)$$

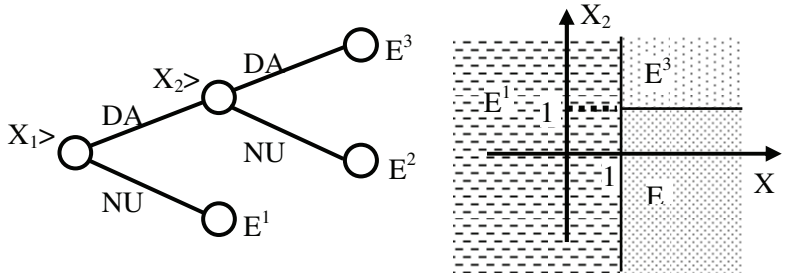
Fiecare din aceste observații aparține (cu privire la *X*) unora dintre subdomeniile considerate, adică $x^i \in E^S$. Vom nota mulțimea observațiilor aparținând lui E^S prin $Data^S$ şi numărul de observații îl vom nota cu N^S. Fie N_i^S notația pentru numărul de observații din $Data^S$ aparținând clasei numărul *i*.

Pentru fiecare scop specific al analizei statistice, există un număr mare de variante diferite de arbori de decizie. Ne punem întrebări de forma: „Care arbore este cel mai bun?" şi „Cum să-l găsim?".

Pentru a răspunde la prima întrebare, vom considera diferite moduri de definire a parametrilor care descriu calitatea unui arbore. Teoretic, putem considera eroarea aşteptată a predicției ca parametru de bază. Totuşi, această valoare poate fi definită numai dacă cunoaştem legea de distribuție probabilistică a variabilelor examinate. De obicei în practică această lege este necunoscută. În consecință, putem estima calitatea numai aproximativ, utilizând mulțimea observațiilor pe care le avem la dispoziție.

Să presupunem că există un arbore de decizie şi un eşantion de obiecte de dimensiune *N*. Este posibil să alegem două tipuri de bază de parametrii pentru descrierea calității unui arbore. Primul tip este reprezentat de parametrii de acurateţe şi al doilea de parametrii de complexitate ai arborelui.

1. *Parametrii de acurateţe* ai arborelui sunt definiți cu ajutorul eşantionului şi caracterizează cât de bine sunt împărțite obiectele din clase diferite (în cazul problemei de



Figura

recunoaştere sau clasificare), sau cât de înaltă este eroarea de predicție (în cazul problemei analizei de regresie).

Numărul relativ (frecvența) greşelilor reprezintă partea obiectelor considerate incorect de un arbore ca făcând parte din altă clasă:

$$\hat{P}_{err} = \frac{N_{err}}{N} \qquad (3)$$

$$N_{err} = \sum_{S=1}^{M} \sum_{\substack{i=1 \\ i \neq \hat{Y}(S)}}^{K} N_i^S \qquad (4)$$

unde *K* este numărul claselor.

Varianța relativă pentru un arbore de regresie poate fi calculată de următoarea formulă:

$$d_{om} = \frac{d_{oc}}{d_0} \qquad (5)$$

$$d_{oc} = \frac{1}{N} \sum_{S=1}^{M} \sum_{i \in Data^S} \left(\hat{Y}(S) - y^i \right)^2 \qquad \text{varianța reziduală} \quad (6)$$

$$d_0 = \frac{1}{N} \sum_{i=1}^{N} \left(y^i - \bar{y} \right)^2 \qquad \text{varianța inițială} \qquad (7)$$

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y^i \qquad (8)$$

2. *Parametrii de complexitate* caracterizează forma arborelui şi nu depind de eşantion. De exemplu, parametrii de complexitate ai unui arbore sunt numărul de frunze ale arborelui, numărul de noduri interne ale sale şi lungimea maximă a unei căi de la rădăcină la o frunză. De asemenea, este posibil să utilizăm lungimea unei căi externe care este definită ca numărul de ramuri ce trebuie adăugate arborelui până la obținerea unui arbore complet.

3. CONCLUZII

Parametrii de complexitate şi acurateţe sunt interconectați: un arbore mai complex, în general, este mai exact, mai corect, adică vom avea acurateţe maximă pentru arborele la care fiecare frunză corespunde la un singur obiect. Însă este de preferat un arbore mai puțin complex. El exprimă dorința de a obține un model mai simplu asupra fenomenului economic cercetat şi facilitează interpretarea ulterioară, explicarea modelului. În plus, din cercetările teoretice rezultă că în cazul unui eşantion de dimensiune mică, în comparație cu numărul caracteristicilor, un arbore prea complex devine instabil, adică furnizează o eroare mai mare pentru noi observații.

Pe de altă parte, este clar că un arbore foarte simplu nu va permite obținerea de rezultate bune pentru predicție. Astfel, la alegerea celui mai bun arbore de decizie, trebuie să facă un compromis între parametrii de acurateţe şi cei de complexitate.

Pentru a obține o astfel de variantă de compromis, este posibil să utilizăm, de exemplu, următorul criteriu pentru calitatea arborelui:

$$Q = p + \alpha M \qquad (9)$$

unde p este un parametru de acurateţe, iar α este un parametru dat. Cel mai bun arbore corespunde la valoarea minimă a criteriului dat. Mai există o variantă utilizată şi anume aceea în care este specificată valoarea maxim admisibilă a complexităţii unui arbore simultan cu căutarea celei mai precise variante de arbore de decizie.[4]

Arborii de decizie reprezintă toate opţiunile şi consecinţele potenţiale într-o manieră care îl face uşor de înţeles şi de comunicat situaţia cu care te întâlneşti. Deplasarea de la stânga la dreapta în arbore, este în general o deplasare înainte în timp. Ceea ce se poate întâmpla ca rezultat al alegerii fiecărei opţiuni este ilustrat sub forma unor ramuri adiţionale.[1]

BIBLIOGRAFIE

1. ***, „Decision Diagrams”, Vanguard Software Corporation, http://www.vanguardsw.com/.
2. ***, „Decision Tree Basics” , Vanguard Software Corporation, http://www.vanguardsw.com/.
3. Baragoin, C., Andersen, C., M., Bayerl, S., Bent, G., Lee, J., Schommer, C., „Mining Your Own Business in Retail Using DB2 Intelligent Miner for Data”, International Technical Support Organization, IBM Corporation, RedBook, San Jose, California, 2001, pp. 28-29.
4. Bryant, R., Samaranayake, V., A., Klinkenberg, R., Wilhite, A., „Alternative Models of Choice: the decision to work”, http://cas.uah.edu/wilhitea/papers/alcohol/alcoholchoice.html.
5. Kamber, M., Winstone, L., Gong, W., Cheng, S., Han, J., „Generalization and Decision Tree Induction: Efficient Classification in Data Mining”, Proc. of 1997 Int'l Workshop on Research Issues on Data Engineering (RIDE'97), Birmingham, England, 1997, p. 111.