INTRODUCERE

Suntem copleşiţi de date - date ştiinţifice, date medicale, date demografice, date financiare, date de marketing. Oamenii nu mai au timp să se uite la aceste date. Atenţia umană a devenit o resursă importantă, astfel încât trebuiesc găsite căi de a analiza datele automat, de a le clasifica automat, de a le sintetiza automat, de a descoperi automat tendinţe în date şi de a caracteriza automat aceste tendinţe. Acest "minerit în date" în vederea găsirii automate de cunoştinţe şi informaţii interesante/noi, este astăzi unul dintre cele mai active şi interesante domenii de cercetare. Cercetătorii din domeniile bazelor de date, statisticii matematice, intelingenţei artificiale şi vizualizării computerizate sunt implicaţi şi contribuie la dezvoltarea acestui domeniu.

Lucrarea de faţă prezintă tehnicile clasice "împrumutate" din statistica matematică de noul domeniu - am numit aici Data Mining; este vorba, mai precis, de tehnici de statistică exploratorie multidimensionale.

Statistica descriptivă permite reprezentarea vie şi asimilabilă a informaţiilor statistice prin simplificare şi schematizare. Statistica descriptivă multidimensională este generalizarea naturală a cazului în care informaţiile privesc mai multe variabile şi/sau dimensiuni.

Trecerea la multidimensional implică însă o schimbare calitativă importantă. Într-adevăr, se spune despre microscop sau despre aparatul radiografic că nu sunt numai intrumente de descriere ci şi instrumente de observaţie, de explorare şi de cercetare. Prin metodele de statistică exploratorie multidimensională realitatea nu este doar simplificată pentru că este complexă, ci şi explorată pentru că este ascunsă. Munca de pregătire şi de codificare a datelor, regulile de interpretare şi validare furnizate de tehnicile furnizate în cazul multidimensional, nu au simplitatea întâlnită în statistica descriptivă elementară. Nu este vorba doar de a prezenta ci şi de a analiza, a descoperi, uneori de a verifica şi dovedi, eventual de a testa anumite ipoteze.

Această lucrare conţine prelegerile ţinute studenţilor de la specializările INFORMATICĂ, şi MATEMATICĂ APLICATĂ ale Facultăţii de Matematică şi Informatică a Universităţii din Bucureşti începând cu anul universitar 1995/1996, în cadrul unor cursuri opţionale organizate anual sau semestrial în funcţie de solicitări.

Numărul metodelor ce permit descrierea şi explorarea tabelelor rectangulare de date statistice (tabele de măsurători-observaţii, tabele de contingenţă, tabele de prezenţă-absenţă, sau tabele de incidenţă) este destul de mare. Metodele reţinute pentru a fi prezentate au fost alese în funcţie de posibilităţile pe care le au de a manipula tabele voluminoase, în funcţie de transpareţa funcţionării lor, în funcţie de calitatea inserţiei în evantaiul metodelor ce sunt în mod real aplicabile şi aplicate.

Două mari familii de metode răspund la aceste exigenţe:

- [capitolul 1]: *metodele factoriale* bazate pe căutarea axelor principale (analiza în componente principale şi analiza corespondenţelor simple şi multiple sunt metodele factoriale cele mai utilizate) care, produc în principal, vizualizări grafice plane sau spaţiale ale obiectelor cercetate;
- [capitolul 2]: *metodele de clasificare* care produc agregări în clase de obiecte sau în familii de clase ierarhizate, obţinute în urma unor calcule algoritmice. Obiectele cercetate sunt grupate, pornind de la vectorii care le descriu, în maiera cea mai puţin arbitrară.

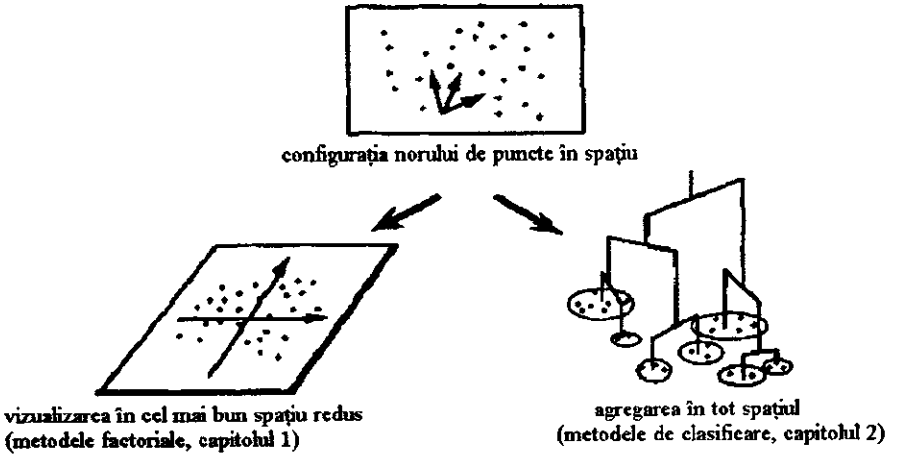

Figura 1 Cele două mari familii de metode ale statisticii exploratorii multidimensionale

Punctele de vedere furnizate de cele două tipuri de metode sunt în esenţă complementare. Vom insista asupra acestei complementarităţi care se manifestă de altfel la mai multe niveluri, fie că este vorba de posibilitatea de a înţelege structuri diverse, fie că este vorba de a ajuta lectura rezultatelor obţinute.

- [capitolul 3]: *metodele explicative uzuale* vor lămuri pe utilizator asupra vocaţiei specifice fiecărei metode (este vorba de analiza discriminantă şi de metodele de segmentare) cât şi asupra legăturilor cu metodele statisticii exploratorii (descrise în primele două capitole). Acest evantai de tehnici acoperă o parte importantă a aplicaţiilor potenţiale ale statisticii.

Nu există totuşi o metodologie generală de articulare în practică a metodelor exploratorii de bază (metodele prezentate în capitolele 1 şi 2) şi metodele explicative uzuale (prezentate în capitolul 3). Fiecare aplicaţie implică, în funcţie de domeniu şi problemă, o muncă originală de codificare şi selecţie a metodelor particulare aplicate. În plus, trebuie să fim conştienţi de faptul că metodele prezentate sunt eficiente în special în cazul datelor nestructurate sau amorfe (în care informaţia a priori asupra acestora este săracă).

Trebuie menționat faptul că există o literatură bogată privind tematica acestei lucrări. Bibliografia atașată constă numai dintr-o selecție a lucrărilor pe care autorul le-a consultat și care pot fi găsite cu ușurință în bibioteci.

Metodele prezentate au un pronunțat carcater matematic-aplicativ. Studenți, practicieni și cercetători din toate disciplinele ce trebuiesc să analizeze și să prelucreze volume mari de date multidimensionale, vor găsi în aceasta lucrare metodele de bază necesare.

Intenția autorului este de a continua dezvoltarea materialului prezentat aici într-o ediție următoare; în consecință observațiile și sugestiile sunt bine venite.

Autorul