Clusterizarea ierarhică cu aplicații în analiza financiară

Dr. Ștefan-Alexandru IONESCU
Academia Română,
Universitatea Româno-Americană, București
ionescusa@gmail.com

Abstract

Analiza cluster poate fi privită ca un instrument care *are ca scop reducerea unor mulțimi de obiecte*, sau chiar de variabile, la un număr mai restrâns de entități informaționale, care sunt clasele sau clusterele. Totuși, deși analiza cluster, privită ca un ansamblu de metode și tehnici de *clasificare a obiectelor*, se aplică în *spațiul variabilelor*, utilizările frecvente ale acestor tehnici de analiză se remarcă pentru clasificarea obiectelor. In aceasta lucrare am aratat cum se pot aplica aceste tehnici in domeniul economico-financiar și cum poate fi detectat numarul de clase în care pot fi împărțite companiile, observând structura latenta existentă.

1. Introducere

Tyron (1939) este primul care folosește termenul de analiză cluster, iar Sokal și Sneath (1963) și Lance și Williams (1967) prezintă primele studii riguroase în acest domeniu. În anii care au urmat, contribuțiile la perfecționarea acestei analize s-au înmulțit și diversificat foarte mult, remarcându-se două importante curente științifice, reprezentate de *școala americană* și de cea *franceză.*

Prin analiza cluster se urmărește, în interiorul unor mulțimi de obiecte sau forme, identificarea de *clase, grupe* sau *clustere* cu elementele cât mai *asemănătoare* în interiorul aceleiași clase (*variabilitate minimă în interiorul claselor)* și cât mai deosebite între ele dacă aceste elemente aparțin unor clase diferite (*variabilitate maximă între clase)*. Rezultă că, analiza cluster permite examinarea *similarităților* și *disimilarităților* dintre obiectele aparținând unei anumite mulțimi, în scopul grupării acestora sub forma unor clase *distincte* între ele și *omogene* în interior. Fiecare obiect din mulțimea analizată este atribuit unei singure clase, iar mulțimea claselor este o mulțime discretă și neordonabilă. Clasele sau grupele sub forma cărora se structurează mulțimile de obiecte se numesc clustere.

Clusterizarea ierarhica este considerata a fi un sistem de ***recunoaștere necontrolată,*** deoarece procesul clasificării pornește fără existența unor informații cu privire la numărul de clase și la apartenența formelor la aceste clase. În acest caz, clasele se construiesc pe măsura creșterii numărului de forme analizate, numărul de clase posibile determinându-se la finalul procesului de recunoaștere. Unii algoritmi de clasificare necontrolată, cum ar fi *algoritmii de partiționare,* presupun fixarea apriorică a numărului de clase în care vor fi împărțite obiectele analizate. Acest lucru nu înseamnă că în mod real este cunoscut numărul de clase, ci doar se face o presupunere cu privire la acest număr.

Sistemele de recunoaştere necontrolată a formelor utilizează principii, metode, proceduri şi tehnici, cunoscute în literatura de specialitate sub denumirea de *tehnici de clasificare*, *clasificare nesupervizată* sau *analiză cluster*.

Analiza cluster, aşa cum o vom denumi în continuare, presupune fixarea formelor sau obiectelor în clustere sau grupe în mod progresiv, fără cunoaşterea apriorică a numărului de clase şi cu respectarea a două *criterii fundamentale*:

a) Fiecare clasă să fie cât mai omogenă, adică să conțină obiecte sau forme cât mai similare în raport cu caracteristicile luate în considerare pentru clasificarea obiectelor;

b) Fiecare clasă trebuie să conțină obiecte clasificate care să difere cât mai mult, din punct de vedere al caracteristicilor de clasificare, de obiectele clasificate în oricare din celelalte clase.

În funcție de caracteristicile procedurilor utilizate, de ipotezele inițiale și de natura rezultatelor obținute, metodele de *clusterizare ierarhică* se împart în clusterizare prin:

- *Agregare* și
- *Divizare.*

Proceduri specifice cunoscute în acest caz, sunt: *metoda agregării simple, metoda agregării complexe, metoda agregării medii, metoda lui Ward, etc.*

În cazul analizei unor cantități mari de date, caracterizate printr-un grad ridicat de eterogenitate, sistemele de recunoaştere necontrolată se utilizează mai mult în scopuri de sistematizare, grupare şi sintetizare informațională. Întrucât aceste tehnici, care se bazează pe utilizarea conceptului de distanță, sunt utile şi eficiente în activitatea de analiză preliminară a datelor, permit organizarea mai eficientă a datelor eterogene, precum şi regăsirea şi interpretarea mai uşoară şi mai consistentă a informațiilor în cadrul unor date astfel structurate.

2. Clusterizarea ierarhica

Metodele de tip ierarhic au ca scop producerea mai multor soluții cluster, numite *ierarhii cluster.* Principala caracteristică este dată de faptul că numărul de clustere nu este cunoscut aprioric și nici nu se sugerează din partea utilizatorului un astfel de parametru.

Ierarhiile cluster sunt structuri cluster cu un număr variabil de clustere, *de tip multinivel* care sunt diferențiate prin numărul de clustere pe care le includ și gradul de agregare al lor. Astfel, având T obiecte, vom avea T soluții cluster, fiecare soluție conținând clustere din ce în ce mai mari, respectiv clustere cu *niveluri de agregare din ce în ce mai ridicate.* O ierarhie cluster are o structură de forma următoare:

nivel 0: $\omega_1^{(0)}, \omega_2^{(0)}, \dots, \omega_{k_0}^{(0)}$

nivel 1: $\omega_1^{(1)}, \omega_2^{(1)}, \dots, \omega_{k_1}^{(1)}$

nivel 2: $\omega_1^{(2)}, \omega_2^{(2)}, \dots, \omega_{k_2}^{(2)}$

...

nivel T-1: $\omega_1^{(T-1)}, \omega_2^{(T-1)}, \dots, \omega_{k_{T-1}}^{(T-1)}$ (1)

unde K_i este numărul de clustere din soluția cluster de la nivelul „*i*”.

Deoarece soluția cluster de tip banal, reprezentată de lista obiectelor supuse clasificării, este prima partiție, rezultă că numărul posibil de soluții dintr-o structură cluster, obținută cu ajutorul algoritmilor ierarhici, va fi mai mic cu 1 decât numărul de obiecte. Acest număr este dat de relația următoare:

$$N_s = T\text{-}1 \quad (2)$$

Alegerea celei mai potrivite soluții cluster, dintre cele *T-1,* se face în funcție de obiectivele urmărite în analiză.

Sunt cunoscute două categorii de algoritmi de clasificare ierarhică:

- *algoritmi de agregare.* În cazul acestor algoritmi, numărul de clustere din prima partiție este egal cu numărul de obiecte, adică $K_0 = T$. De asemenea, numărul de clustere dintr-o partiție de la un anumit nivel este mai mic cu 1 decât numărul de clustere din partiția de la nivelul inferior și mai mare cu 1 decât numărul de clustere din partiția de la nivelul superior, respectiv :

$$K_{T-1} = K_{T-2} - 1 = K_{T-3} - 1 = \cdots = K_2 - 1 = K_1 - 1 = K_0 - 1 \qquad (3)$$

- *algoritmi de dezagregare*. Aceste metode constă practic în aceleași operațiuni folosite de algoritmii aglomerativi, dar în ordine inversă. Astfel, prima partiție considerată, este reprezentată de un singur cluster ce conține toate obiectele, a doua partiție va consta în două clustere, și așa mai departe.

Metodele de clasificare ierarhică sunt considerate *metode euristice,* care cuprind proceduri de clasificare dezvoltate pe baza unei anumite modalitati intuitive de soluționarea unei anumite probleme particulare (euristică).[1]

Printre aceste metode putem menționa *metoda agregării simple, metoda agregării complete, metoda agregării medii, metoda centroidului sau metoda lui Ward.*

Distanța Ward dintre două clustere măsoară *variabilitatea intracluster cumulată*, *indusă de comasarea a două clustere,* la nivelul configurației cluster rezultate. Prin comasarea a două clustere se urmărește obținerea unei omogenități maxime la *nivelul tuturor clusterelor* care aparțin unei configurații date a obiectelor pe clustere.

Rezultă că distanța Ward este singura care ia în calcul minimizarea variabilității intracluster sau, cu alte cuvinte, maximizarea variabilității intercluster, adică a gradului de omogenitate a clusterelor. Trebuie precizat că, gradul de omogenitate a unui cluster se maximizează prin minimizarea sumei totale a pătratelor abaterilor intracluster.

Dacă w_{12} este noul cluster obtinut prin comasarea clusterului w_1 cu w_2, atunci sumele distantelor intra cluster vor fi:

$$\begin{aligned} SSE_{w_1} &= \sum_{i=1}^{T_{w_1}} (y_i - \bar{y}_{w_1})^t (y_i - \bar{y}_{w_1}) \\ SSE_{w_2} &= \sum_{i=1}^{T_{w_2}} (y_i - \bar{y}_{w_2})^t (y_i - \bar{y}_{w_2}) \\ SSE_{w_{12}} &= \sum_{i=1}^{T_{w_{12}}} (y_i - \bar{y}_{w_{12}})^t (y_i - \bar{y}_{w_{12}}) \end{aligned} \qquad (4)$$

Se vor comasa acele două clustere w_1 și w_2 care minimizează creşterea sumei pătratelor erorilor definită ca:

1.Euristicile sunt reguli deduse pe baza unor raționamente teoretice, sau a unor observații statistice.

$$I_{w_{12}} = SSE_{w_{12}} - (SSE_{w_1} - SSE_{w_2}) \quad (5)$$

Ruxanda (2009) consideră etapele analizei cluster pentru clasificarea unei mulțimi de obiecte, ca fiind următoarele:

- *alegerea caracteristicilor* în funcție de care se va face clasificarea;
- *alegerea tipului de măsură* pentru evaluarea proximității dintre obiecte;
- *stabilirea regulilor de formare a claselor sau clusterelor*;
- *construirea claselor*, adică încadrarea obiectelor în clase;
- *verificarea consistenței și semnificației* clasificării;
- *alegerea unui număr optimal de clustere*, în funcție de natura problemei de clasificare și de scopurile care se urmăresc;
- *interpretarea semnificației clusterelor*.

Așadar, prin analiza cluster se încearcă identificarea, în datele inițiale, a unor grupuri, clase sau clustere, în funcție de similaritățile și disimilaritățile dintre obiectele la care se referă respectivele date. În ceea ce privește tehnica utilizată, analiza cluster pentru clasificarea obiectelor evaluează distanțele pentru perechi de obiecte, iar analiza cluster pentru clasificarea variabilelor evaluează distanțele pentru perechi de variabile.

3. Datele folosite în analiză

Au fost selectate 101 firme ce își desfasoara activitatea in Romania. Firmele au fost active și au depus cel puțin o cerere de credit. Acest lucru implică faptul că au depus la 31 decembrie toate declarațiile financiare.

Eșantionul ales este reprezentativ pentru companiile private românești, care nu sunt listate la bursa de valori. Valoarea activelor este cuprinsă între aproximativ 15.000 lei și nu depașește 30 milioane lei. În mod evident, cele mai multe dintre firme sunt de mărime medie, cu active cuprinse intre 1 și 4 milioane lei. Firmele mici și cele mari se regăsesc în proporții asemănătoare în eșantionul selectat.

Așa cum am precizat, datele primare au fost extrase din bilanțurile, conturile de profit și pierdere depuse la sfârșitul anului, precum și din balanțele aferente lunii decembrie.

În primul rând am avut în vedere:

- activele, precum și clasificările acestora;
- datoriile, împărțite de asemenea pe diverse categorii, inclusiv datorii

către bănci și societăți de leasing;
- capitalurile și capitalurile proprii;
- date legate de cifra de afaceri, profit, taxe și impozite.

Ulterior, am prelucrat aceste date și am calculat o serie de rate financiare care oferă un grad de comparabilitate ridicat pentru firme de diferite dimensiuni si din diverse domenii de activitate. Am urmărit în principal să acopăr patru direcții și anume: lichiditate, solvabilitate, activitate și profitabilitate. Din multitudinea de rate existente, am ales sa reținopt dintre ele, și anume:

- Rate de profitabilitate***:*** Rata de rentabilitate a activelor totale ***(ROA),*** Rentabilitatea financiară a capitalului ***(ROE),*** rentabilitatea capitalului angajat ***(ROCE).***
- Rate privind eficiența***:*** Rotația activului total ***(RAT).***
- Rate privind lichiditatea***:*** Lichiditatea curentă ***(CR),*** Lichiditatea imediată ***(QR),*** Lichiditatea efectivă-Cash Ratio ***(CashR).***
- Rate privind solvabilitatea***:*** Solvabilitatea patrimonială ***(SP).***

Pentru prelucrarea datelor a fost folosit pachetul de programe *STATISTICA 8.0*.

4. Rezultate obținute

Metoda de clasificare prezentată este legată de analiza cluster de tip ierarhic. Așa cum am arătat mai sus, prin acest tip de analiză se grupează obiectele, în acest caz-cele 101 firme pe baza măsurării distanțelor sau similarităților dintre acestea. Am luat în considerare firmele descrise de cele 8 variabile prezentate anterior. O astfel de metodă de amalgamare pleacă de la 101 clustere, reprezentate de toate firmele, care urmează să fie comasate treptat, relaxând criteriul de grupare până se ajunge la un singur cluster ce conține toate obiectele. Nu se cere ca input un număr de clustere dorit, gruparea se face natural, iar utilizatorul poate observa numărul de clase care se prefigurează.

În primă fază, am calculat distanțele dintre cele 101 obiecte. Pentru exemplificare, în tabelul 1 sunt prezentate distanțele dintre primele 10 firme.

Distanțele de tip City-block dintre primele 10 obiecte

Tabelul 1

	1	2	3	4	5	6	7	8	9	10
1	0.0000	7.9698	2.3494	2.9692	4.7642	5.4116	7.9730	4.0441	4.8960	7.3244
2	7.9698	0.0000	6.3338	7.0238	7.3226	3.8755	3.0325	7.4861	4.1965	7.9186
3	2.3494	6.3338	0.0000	2.4776	4.0079	3.1626	7.1092	3.2650	3.4434	5.7815
4	2.9692	7.0238	2.4776	0.0000	4.3906	5.5837	7.9746	3.6462	3.2889	7.1048
5	4.7642	7.3226	4.0079	4.3906	0.0000	5.5027	8.3231	3.9546	4.0691	2.7428
6	5.4116	3.8755	3.1626	5.5837	5.5027	0.0000	4.4133	5.1345	3.0615	5.4216
7	7.9730	3.0325	7.1092	7.9746	8.3231	4.4133	0.0000	8.9225	5.1529	8.5541
8	4.0441	7.4861	3.2650	3.6462	3.9546	5.1345	8.9225	0.0000	4.2009	4.8248
9	4.8960	4.1965	3.4434	3.2889	4.0691	3.0615	5.1529	4.2009	0.0000	4.8694
10	7.3244	7.9186	5.7815	7.1048	2.7428	5.4216	8.5541	4.8248	4.8694	0.0000

S-a considerat spațiul 8-dimensional în care am calculat distanțele de tip city-block. Alegerea a fost determinată de faptul că acest tip de distanță nu amplifică diferențele de coordonate prin ridicări la putere, fiind astfel mai robustă în raport cu prezența în date a valorilor aberante.

Distanțele apar sub forma unei matrici simetrice, în care elementul *(i,j)* arată distanța Manhattan dintre firma *i* și firma *j* în spațiul 8-dimensional definit de cele 8 variabile. Evident că elementele ce compun diagonala principală sunt egale cu 0, ele reprezentând distanțe între obiecte pentru care *i=j*. Matricea este simetrică, adică: *d(i,j)=d(j,i).* Astfel, distanța dintre firma *1* și firma *2* este de 7.9698 în spațiul 8-dimensional, distanța dintre firmele *1* și *3* este de 2.3494 în același spațiu, șamd.

Am încercat să folosesc mai multe metode de amalgamare, cea care a dat rezultatele cele mai satisfăcătoare fiind metoda lui Ward. Prin această metodă, se formează clustere, astfel încât la fiecare pas, atribuirea unui obiect la un cluster minimizează varianța din interiorul clusterului.

Programul de amalgamare prin metoda lui Ward

Tabelul 2

Iteration No.	Manhattan Distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6
1	**0.26844**	27	46				
2	**0.425123**	56	76				
3	**0.588868**	57	91				
4	**0.611369**	31	44				
5	**0.679478**	27	46	51			
6	**0.681315**	17	96				
7	**0.745417**	16	77				
8	**0.761216**	60	81				
9	**0.809053**	87	94				
10	**0.811932**	55	57	91			
11	**0.854273**	12	86				
12	**0.937787**	87	94	92			
13	**0.999241**	63	82				
14	**1.011493**	15	53				
15	**1.050489**	3	24				
16	**1.057668**	16	77	99			
17	**1.085942**	38	90				
18	**1.102812**	29	32				
19	**1.145189**	39	66				
20	**1.171946**	34	35				
21	**1.178349**	42	78				
22	**1.258784**	61	75				
23	**1.272284**	47	93				
24	**1.275334**	21	101				
25	**1.288252**	49	62				
26	**1.29384**	22	30				
27	**1.320899**	1	29	32			
28	**1.338401**	22	30	43			
29	**1.342997**	5	95				
30	**1.408685**	8	54				
31	**1.439288**	56	76	87	94	92	
32	**1.458543**	28	36				
33	**1.462615**	58	60	81			

În tabelul 2 am exemplificat primele 33 etape ale agregării. Inițial, există 101 clustere, fiecare conținând una din cele 101 firme. Cea mai mică distanță dintre două firme este de 0.2684397. Primul pas al amalgamării este reprezentat de formarea unui cluster din aceste două obiecte. Astfel, în urma primei iterații, vom avea 100 clustere: unul format din firmele 27 și 46 și alte 99 clustere formate

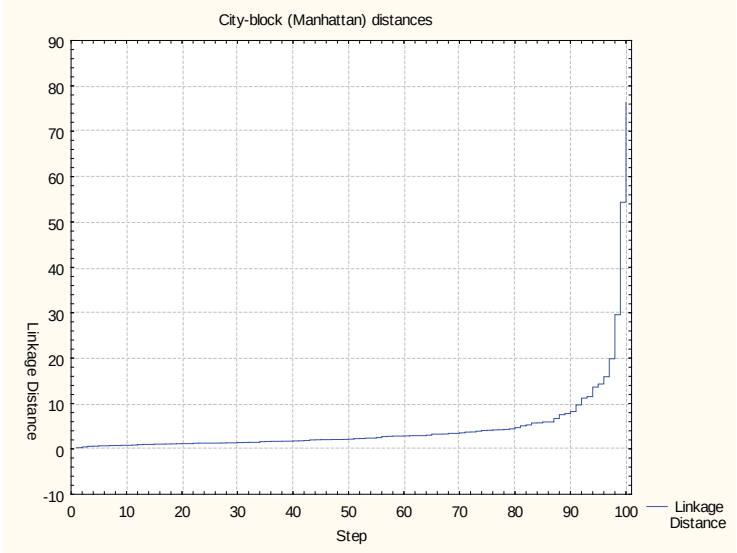
din celelalte 99 firme. Următorul pas este format din gruparea firmelor 56 cu 76 între care există o distanță de 0.4251230. În urma acestei iterații, rămân 99 clustere: cel format la prima iterație (alcătuit din firmele 27 și 46), cel format la cea de a doua iterație (alcătuit din firmele 56 și 76), precum și alte 97 clustere formate din celelalte firme rămase. Procesul continuă asemănător.

Pasul al cincilea reprezintă afectarea obiectului 51 la clusterul deja format la primul pas. Apare astfel un cluster format din 3 firme și anume 27, 46 și 51. La fiecare pas, suma pătratelor abaterilor la nivelul noului cluster format este cea mai mică în comparație cu alte perechi de clustere potențiale. La cea de a 31-a iterație, două clustere formate anterior se unesc într-un cluster mai mare. Astfel, distanța de 1.439288 dintre clusterul format la iterația 2 (alcătuit din firmele 56 și 76) și cel format la iterația 12 (alcătuit din firmele 87, 94, 92) permite comasarea acestora într-un nou cluster ce va conține toate aceste 5 firme. În urma iterației 100, toate cele 101 firme vor forma un singur cluster.

Distanțele din prima coloană a tabelului 2 sunt reprezentate pe axa *Oy* în figura 1. Pe axa *Ox*, apar cele 100 iterații. În dreptul primei iterații se pornește cu un punct, la nivelul 0.2684397 pe *Oy*. În dreptul iterației 2, se trasează un segment de dreaptă, paralelă cu axa *Oy*, între valorile 0.2684397 și 0.425123 și așa mai departe până la ultima iterație. Pentru fiecare caz, extremitatea superioară a segmentului de dreaptă corespunzător iterației *i* se unește cu extremitatea inferioară a segmentului de dreaptă corespunzător iterației *i+1*.

Graficul distanțelor de agregare

Figura 1



Dendrograma amalgamării pentru metoda lui Ward

Figura 2

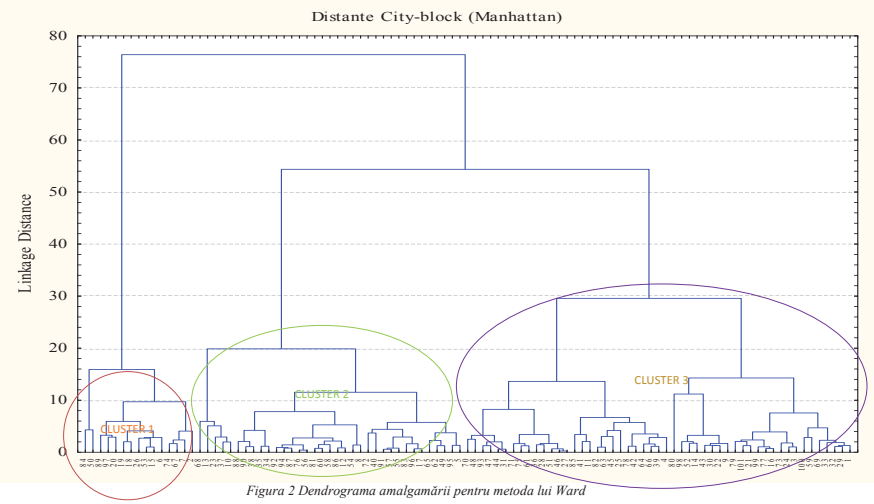


Figura 2 Dendrograma amalgamării pentru metoda lui Ward

Acest grafic poate fi foarte folositor, sugerând vizual unde ar trebui să se întrerupă natural procesul de clusterizare. Pe măsură ce se înaintează către dreapta, distanța dintre obiecte crește (lungimea segmentelor de dreaptă devine mai mare), se formează clustere mai mari, iar varianța intra-cluster este mai mare. În primă fază se observă o evoluție lentă, până la pasul 80, creșterea distanței fiind foarte mică. Urmează creșteri mai însemnate ale distanțelor până la pasul 98, ultimile 2 etape constând în alipirea unor obiecte ce au distanțe foarte mari. Dacă distanța dintre obiectele comasate la primul pas este de 0.2684397, distanța dintre obiectele comasate la pasul 100 este de 76.4076, adică de 285 ori mai mare. Deoarece distanța de amalgamare de la pasul i este mai mare decât distanța de amalgamare de la pasul *i-1* (oricare ar fi *i*), putem spune despre metoda aleasă că îndeplinește condiția de monotonicitate și este ultrametrică. Distanța poate fi un criteriu optim în stabilirea numărului de clustere ce urmează a fi reținute.

Formarea a 3 clustere naturale reiese și din figura 2, unde este prezentat arborele ierarhic. De la etapa 98 la etapa 99, distanța aproape că se dublează, reprezentând o alipire nenaturală. Sugerez astfel reținerea a 3 clustere după cum sunt marcate în figura 2.

5.Concluzii

Analiza cluster se deosebeşte fundamental de procedurile de natură statistică, prin faptul că nu se bazează şi nu presupune îndeplinirea apriorică a niciunei ipoteze specifice. Rencher (2002) consideră că analiza cluster constituie un important şi eficient instrument de *analiză exploratorie,* al cărui scop este acela de a crea aşa numitele *taxonomii sau tipologii,* bazate pe analiza *asemănărilor* şi *deosebirilor* existente între obiectele unei mulţimi date.

Analiza cluster este utilă în orice proces de analiză a datelor, nu numai în cele care necesită o clasificare. De exemplu, în cazul unui proces de analiză ce vizează un set de date de dimensiuni foarte mari, atât din punct de vedere al obiectelor analizate, cât şi din punct de vedere al caracteristicilor acestora, sintetizarea şi structurarea informaţiei poate fi făcută prin instrumente adecvate. Astfel, pentru identificarea unor categorii, clase sau grupe informaţionale pe o mare cantitate de informaţii brute, poate fi folosită cu succes analiza cluster.

Analiza cluster permite deducerea legilor evoluţiei unor populaţii de fenomene, precum şi a principiilor procesului de cunoaştere, prin:

- *definirea unor scheme de clasificare formală* şi a unor *tipologii*, pentru cunoaşterea şi înţelegerea mai bună a realităţilor complexe;
- *identificarea unor modele statistico-matematice* pentru înţelegerea, sintetizarea şi simplificarea mulţimilor complexe şi eterogene de fenomene şi procese;
- *definirea mai corectă şi mai completă a caracteristicilor fundamentale* ale unor populaţii de fenomene şi procese;
- *deducerea unor măsuri numerice adecvate pentru caracterizarea dimensiunilor populaţiilor* de fenomene şi pentru evidenţierea modificărilor care au loc în structura acestora;
- *identificarea unor entităţi individuale care sunt reprezentative* pentru clase şi categorii complexe de fenomene şi procese.

Recunoaştere:

Aceast articol a beneficiat de suport financiar prin proiectul *„Rute de excelenţă academică în cercetarea doctorală şi post-doctorală – READ”*, contract nr. POSDRU/159/1.5/S/137926, Beneficiar Academia Română, proiect cofinanţat din Fondul Social European prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007-2013.

Bibliografie

1. Aggarwal, C., & Yu, P. (2000). Finding generalized projected clusters in high dimensional spaces. *Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00)*, (pp. 70-81). Dallas, USA.
2. Arabie, P., Hubert, L., & De Soete, G. (1996). *Clustering and Classification.* New York, USA: World Scientific.
3. Back, A.D.; Weigend, A.S. Discovering Structure in Finance Using Independent Component Analysis; (1998) *Advances in Computational Management Science* Volume 2, 1998, pp 309-322
4. Beil, F., Ester, M., & Xu, X. (2002). Frequent term-based text clustering. *Proc. 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02)*, (pp. 436-442). Edmonton, Canada.
5. Bradley, P., Fayyad, U., & Reina, C. (1998). Scaling clustering algorithms to large databases. *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, (pp. 9-15). New York, USA.
6. Chen, K.H. and Shimerda, T.A. An Empirical Analysis of Useful Financial Ratios, (1981), *Financial Management* Vol. 10, No. 1 (Spring, 1981), pp. 51-60
7. Dieckmann, S., Plank, T., Default Risk of Advanced Economies: An Empirical Analysis of Credit Default Swaps during the Financial Crisis, (2012), *Review of Finance* (2012) 16 (4):903-934.doi: 10.1093/rof/rfr015
8. Hastie, T; Tibshirani, R; Friedman, J (2009). "14.3.12 Hierarchical clustering". *The Elements of Statistical Learning* (PDF) (2nd ed.). New York: Springer. pp. 520–528. ISBN 0-387-84857-6. Retrieved2009-10-20.
9. Jain, A.K, (1999,). Data Clustering: A Review,. *ACM Computing Surveyes (CSUR), 31*, 264-323
10. Kaufmann, L., & Rousseuw, P. (2005). *Finding Groups inData: An Introduction to Cluster Analysis.* New York, USA: John Wilwy & Sons.
11. Lance, G., & Williams, W. (1967). A general theory of classificatory sorting strategies. *Computer Journal, 9*
12. Liu, B., Xia, Y., & Yu, P. (2001). Clustering through decision tree construction. *Proc. 2000 ACMCIKM Int. Conf. Information and KnowledgeManagement (CIKM'00)*, (pp. 20-29). McLean, USA.
13. Sokal, R., & Sneath, P. (1963). *Principles of numerical taxonomy.* San Francisco, USA: W.H. Freeman Co.
14. Rencher, A. (2002). *Methods of Multivariate Analysis.* New York, USA: John Wiley & Sons.
15. Ruxanda, G. (2001). *Analiza Datelor.* București: ASE.
16. Ruxanda, G. (2010). Construirea, estimarea și implantarea software a metodelor matematice. *Cercetarea științifică în ASE*.
17. Tyron, R. (1939). *Cluster Analysis.* Ann Arbor, USA: Edwards Brothers
18. Zhang, et al. "Graph degree linkage: Agglomerative clustering on a directed graph." 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012