(19) OFICIUL DE STAT
PENTRU INVENŢII ŞI MĂRCI
Bucureşti

ROMÂNIA

(11) RO 128168 A2
(51) Int.Cl.
G06F 17/00 (2006.01)

(12) # CERERE DE BREVET DE INVENŢIE

(21) Nr. cerere: **a 2011 00279**

(22) Data de depozit: **30.03.2011**

(41) Data publicării cererii:
28.02.2013 BOPI nr. **2/2013**

(71) Solicitant:
• **ANESCU GEORGE DORI**,
STR. CETATEA DE BALTĂ NR. 12, BL. 27, SC. C, AP. 59, SECTOR 6, BUCUREŞTI, B, RO

(72) Inventatori:
• **ANESCU GEORGE DORI**,
STR. CETATEA DE BALTĂ NR. 12, BL. 27, SC. C, AP. 59, SECTOR 6, BUCUREŞTI, B, RO

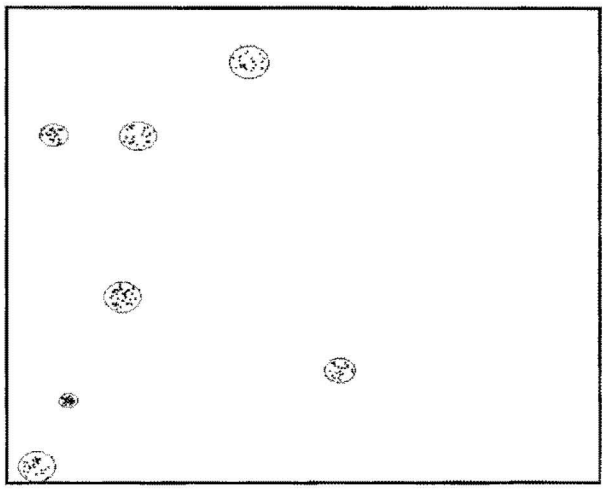
(74) Mandatar:
ROMINVENT S.A.,
STR. ERMIL PANGRATTI NR.35, SECTOR 1, BUCUREŞTI

(54) # METODĂ RAPIDĂ DE IDENTIFICARE A AGLOMERĂRILOR

(57) **Rezumat:**

Invenţia se referă la o metodă de identificare a seturilor de aglomerare formate din puncte de date relaţionate, dintr-un interval Euclidian multidimensional. Metoda este denumită MMCC (Moving Mass Center Clustering - Identificare de Aglomerări prin Deplasarea Centrului de Masă), deoarece în bucla interioară a algoritmului, de fiecare dată când unele puncte de date sunt eliminate pe baza unui criteriu de raport de distanţe, centrul de masă al punctelor de date rămase se deplasează mai aproape de o posibilă aglomerare. În bucla exterioară, prin eliminarea punctelor de date ale unei aglomerări identificate sau a punctelor de date ale unui subdomeniu cu o densitate mare, dar care nu satisfac criteriile de identificare de aglomerări, se creează o oportunitate de a fi identificată o nouă aglomerare într-o nouă execuţie a buclei interioare. Rezultatele finale ale algoritmului metodei MMCC constau dintr-un set de aglomerări globulare cu centre şi raze cunoscute.

Revendicări: 16
Figuri: 5



Setul de aglomerari identificate C

Fig. 4

RO 128168 A2

Descriere

1. Domeniul Invenției

[0001] Invenția de față se referă la metode pe bază de calculator și sisteme de calcul pentru lucrul cu și analiza seturilor de date cu scopul identificării aglomerărilor unui set de obiecte în sub-seturi de obiecte similare. Identificarea de Aglomerări (Eng., Clustering) sau Analiza de Aglomerări (Eng., Cluster Analysis) este o metodă de învățare nesupervizată, și o tehnică obișnuita pentru analiza statistică a datelor utilizată în multe domenii, inclusiv în Învățarea Automată (Eng., Machine Learning), Minarea de Date (Eng., Data Mining), Recunoașterea de Modele (Eng., Pattern Recognition), Analiza de Imagini (Eng., Image Analysis), regăsirea de informație, optimizare și bioinformatica. Pe lângă termenul de aglomerare, există un număr de termeni cu semnificație similară, inclusiv clasificare automată, taxonomie numerică, botriologie și analiză tipologică. Scopul unei metode de identificare a aglomerărilor este atribuirea unui set de date format din obiecte la sub-seturi (denumite aglomerări) astfel încât obiectele din aceeași aglomerare sunt similare într-un anumit sens. Similaritatea a două elemente este determinată pe baza definiției unei măsuri de distanță.

2. Descrierea Domeniului Înrudit

[0002] Sunt cunoscute mai multe tipuri de metode de identificare a aglomerărilor cum ar fi metode de partiționare (Eng., partitioning), metode de ierarhizare (Eng., hierarchical), metode pe bază de densitate (Eng., density-based), și metode pe bază de rețea (Eng., grid-based). Cu toate acestea în practică, există unele dezavantaje asociate cu fiecare dintre metodele de identificare a aglomerărilor menționate. Un dezavantaj asociat cu mulți algoritmi de identificare a aglomerărilor este cerința de a specifica, înainte de execuția algoritmului (ca un parametru număr întreg), numărul de aglomerări care urmează sa fie produse din setul de puncte de date de intrare. Dacă nu există posibilitatea de a cunoaște valoarea adecvată în prealabil, aceasta trebuie sa fie

determinată pe parcursul execuției algoritmului, în acest mod fiind creată o problemă suplimentară pentru care s-au dezvoltat un număr de tehnici de abordare.

[0003] Mai întâi, în ceea ce privește metodele de identificare a aglomerărilor pe bază de partiționare, în mod caracteristic acestea determină toate aglomerările în același timp, dar pot fi de asemenea utilizate ca algoritmi de diviziune în metodele de identificare a aglomerărilor pe bază de ierarhizare. Funcția criteriu pe care algoritmul de identificare a aglomerărilor încearcă să o minimizeze poate sublinia structura locală a datelor, la fel ca și cum ar atribui aglomerări la vârfuri din funcția de densitate de probabilitate, sau din structura globală. În mod caracteristic criteriile globale implica minimizarea unei masuri de ne-similitudine în eșantioanele din fiecare aglomerare, în același timp maximizând ne-similitudinea diferitelor aglomerări. De exemplu în metode de identificare a aglomerărilor denumită K-means funcția criteriu este distanța pătratică medie a punctelor de date față de centroizii de aglomerare cei mai apropiați acestora. Un număr dat de centroizi de aglomerare este selectat (aleatoriu) și fiecare punct de date este atribuit la cel mai apropiat centroid de aglomerarea (utilizând o măsura de distanță, de exemplu, distanța Euclidiană). După atribuire fiecare centroid de aglomerare este deplasat la media punctelor de date asociate la acesta, și pașii de atribuire de puncte de date și de deplasare de centroizi de aglomerare sunt repetați până când centroizii de aglomerare converg la poziții fixe. Deși algoritmii de identificarea de aglomerări prin partiționare sunt eficienți în viteza de identificarea a aglomerărilor, rezultatul identificării de aglomerări este instabil, pot apare diferite tipuri de aglomerări atunci când numărul de centroizi sau pozițiile inițiale ale centroizilor sunt modificate. O buna inițializare a centroizilor de aglomerare poate fi de asemenea esențială, unele aglomerări pot chiar rămâne vide dacă centroizii acestora sunt plasați inițial departe față de distribuția de date. Alte neajunsuri ale algoritmilor de identificare de aglomerări prin partiționare constau în faptul ca aceștia nu sunt adecvați pentru tratarea de aglomerări ne-globulare de diferite forme, dimensiuni, și densități și faptul ca punctele de date de zgomot nu sunt eliminate prin filtrare. Pe lângă K-means alți algoritmi reprezentativi de identificare de aglomerări prin partiționare sunt K-medoids, PAM (Partition Around Medoids – partiție în jurul medoizilor), CLARA (Clustering Large

Applications - identificarea de aglomerări în aplicații mari), CLARANS (CLARA randomizat), c-means fuzzy, etc.

[0004] În privința metodelor de identificare a aglomerărilor pe bază de ierarhizare, abordarea acestora este de a găsi aglomerări succesive utilizând aglomerări stabilite anterior. Setul de puncte de date este organizat într-o structură ierarhică de tip arbore care este construită printr-o abordare aglomerativă („de jos în sus") sau printr-o abordare divizivă („de sus în jos"). Algoritmii aglomerativi încep cu fiecare element ca o aglomerare separată și le unesc pe acestea în aglomerări succesive din ce în ce mai mari. Algoritmii divizivi încep cu întregul set și continuă sa-l divizeze în aglomerări succesive din ce în ce mai mici. Cu toate acestea, algoritmii de identificare de aglomerări ierarhici convenționali trebuie să compare similitudinea datelor în timpul combinării sau descompunerii, ceea ce poate cauza cu ușurință o durata mare de timp de execuție. Complexitatea temporală și spațială limitează sever dimensiunea seturilor de date care pot fi procesate de către algoritmii de identificare de aglomerări ierarhici. Exemple reprezentative de algoritmi pentru abordarea aglomerativă sunt BRICH, CURE, ROCK, etc., și un algoritm reprezentativ pentru abordarea divizivă este CHAMELEON.

[0005] În ceea ce privește metodele de identificare de aglomerări pe bază de densitate, în cazul acestora o aglomerare este privită ca o regiune în care densitatea obiectelor de date depășește un prag pre-stabilit. Acestea sunt capabile sa descopere oricare aglomerări de formă arbitrară în conformitate cu densitatea de date a unei regiuni prin localizarea regiunilor de densitate înaltă care sunt separate una de alta prin regiuni de densitate scăzută. Căutarea poate fi extinsă de la zona identificată inițial, și alte zone care îndeplinesc criteriile de densitate pot fi combinate, astfel încât sa formeze rezultatul de identificare de aglomerări. Algoritmii de identificare de aglomerări pe bază de densitate sunt capabili să identifice figuri neregulate și să elimine prin filtrare date de zgomot în mod eficient. Principalul dezavantaj al algoritmilor de identificare de aglomerări pe bază de densitate este legat de necesitatea de a calcula toate sau un număr mare de distante dintre perechi de puncte de date care formează setul de date.

Pentru seturi de puncte de date mari și număr de dimensiuni mare timpul de execuție poate deveni excesiv. Un alt neajuns este în legătură cu dificultatea de a trata aglomerări cu densități variabile datorită faptului că parametrii metodei sunt definiți în general ca parametri globali pentru spațiul problemei. Algoritmi pe bază de densitate reprezentativi sunt DBSCAN (Density Based Spatial Clustering of Applications with Noise - identificare de aglomerări spațiale pe bază de densitate pentru aplicații cu zgomot), IDBSCAN (Improved DBSCAN - DBSCAN îmbunătățit), GDBSCAN (Generalized DBSCAN - DBSCAN generalizat), OPTICS (Ordering Points To Identify the Clustering Structure - ordonare de puncte cu scopul identificării structurii de aglomerare), HOP, DENCLUE, etc.

[0006] În final, în privința metodelor de identificare a aglomerărilor pe bază de rețea, abordarea consta mai întâi în a rezuma setul de date cu ajutorul unei reprezentări de rețea, și apoi a uni celulele rețelei cu scopul de a obține aglomerări. Identificarea de aglomerări pe bază de rețea este considerată adecvată în mod particular pentru a trata seturi de date masive. Viteza de identificare de aglomerări a operațiunii pe bază de rețea convenționale este rapidă datorită faptului ca unitatea de aglomerare minimă este o rețea. Cu toate acestea, rețelele sub formă de dreptunghi pot genera rezultate de identificare de aglomerări imprecise sau modele cu margini în zigzag. Algoritmi reprezentativi pentru metodele de identificare a aglomerărilor pe bază de rețea sunt STING (STatistical INformation Grid – rețea de informație statistică), WaveCluster, CLIQUE (CLustering în QUEst - identificare de aglomerări în căutare), etc.

[0007] În conformitate cu stadiul tehnicii prezentat în general mai sus, există mai multe neajunsuri asociate cu metodele și algoritmii de identificare de aglomerări convenționali. Există o nevoie în domeniu de metode și de tehnici de identificare de aglomerări îmbunătățite capabile sa trateze seturi de date mari intr-un spațiu cu un număr mare de dimensiuni cu limitări de timp și spațiu acceptabile, care în același timp furnizează rezultate stabile și fiabile. De asemenea metodele îmbunătățite ar trebui sa fie capabile sa trateze în mod eficient datele cu zgomot și să localizeze în mod fiabil aglomerări cu forme complexe (ne-globulare) și diferite dimensiuni și densități.

3. Rezumatul Invenției

[0008] În conformitate cu invenția de față, este furnizată o metodă pe bază de calculator fundamentală de identificare a seturilor de aglomerare formate din puncte de date relaționate dintr-un interval Euclidian multi-dimensional. Metoda este denumită Identificare de Aglomerări prin Deplasarea Centrului de Masă (MMCC - Moving Mass Center Clustering) deoarece în bucla interioară a algoritmului de fiecare data când unele puncte de date sunt eliminate pe baza unui criteriu de raport de distanțe, centrul de masă al punctelor de date rămase se deplasează mai aproape de o posibila aglomerare. În bucla exterioară, prin eliminarea punctelor de date ale unei aglomerări identificate sau a punctelor de date ale unui sub-domeniu cu o densitate mare, dar care nu satisface criteriile de identificare de aglomerări, se creează o oportunitate de a fi identificată o nouă aglomerare într-o nouă execuție a buclei interioare. Rezultatele finale ale algoritmului fundamental al metodei MMCC constau dintr-un set de aglomerări globulare cu centre și raze cunoscute.

[0009] În conformitate cu o primă aplicație concretă a invenției de față, metoda pe bază de calculator fundamentală poate fi adaptată pentru aplicații în Minare de Date (Eng., Data Mining). Este necesară o fază de pre-procesare constând în standardizare și scalare cu scopul de a transforma setul de date la o formă care poate fi tratata de către metodă.

[0010] În conformitate cu o a doua aplicație concretă a invenției de față, metoda pe bază de calculator fundamentală poate fi adaptată pentru aplicații de Învățare Automată (Eng., Machine Learning). Metoda are nevoie de toate datele de antrenament inițial disponibile în același timp cu scopul de a efectua procesul de antrenament care constă în identificarea categoriilor de date.

[0011] În conformitate cu o a treia aplicație concretă a invenției de față, metoda pe bază de calculator fundamentală poate fi adaptată pentru aplicații în Analiză de

Imagini (Eng., Image Analysis) pentru care formele, dimensiunile și densitățile aglomerărilor identificate sunt importante.

[0012] În conformitate cu o a patra aplicație concretă a invenției de față, metoda pe bază de calculator fundamentală poate fi adaptată la Optimizarea prin Identificare de Aglomerări (Eng., Clustering Optimization), unde ponderile asociate cu punctele de date sunt importante.

[0013] În conformitate cu un aspect al invenției de față, invenția poate lua forma unui sistem de calcul general sau special care implementează oricare dintre metodele invenției.

[0014] În conformitate cu un aspect suplimentar al invenției de față, invenția poate lua forma unui mediu care poate fi citit de calculator având instrucțiuni care pot fi executate de calculator, cum ar fi soft, incorporate pe acesta pentru efectuarea oricărora dintre metodele invenției.

[0015] În conformitate cu invenția de față metoda dezvăluită de identificare de aglomerări pe bază de calculator fundamentală împreună cu variantele sale adaptate are avantaje față de metodele de identificare de aglomerări convenționale în privința vitezei de execuție, a stabilității, a fiabilității, a capacității de gestionare a zgomotului, a capacității de a trata seturi de date mari în spatii de dimensiuni mari și a capacității de a localiza aglomerări de diverse forme, dimensiuni și densități. De asemenea nu este nevoie să se specifice numărul de aglomerări care trebuie sa fie produse.

4. Scurta Descriere a Desenelor

[0016] Cuvântul „exemplar" este utilizat aici cu semnificația de servind ca un exemplu, caz, sau ilustrare. Oricare aplicație concretă sau proiect descris aici ca „exemplar" nu trebuie sa fie considerat în mod necesar ca preferat sau avantajos față de alte posibile aplicații concrete sau proiecte.

[0017] Caracteristicile și natura invenției de față vor deveni mai evidente din descrierea detaliată prezentată mai jos atunci când este considerată împreună cu desenele anexate, în care:

[0018] Fig. 1 prezintă grafic un set exemplar de puncte de date utilizat pentru ilustrarea funcționării algoritmului fundamental MMCC.

[0019] Fig. 2 prezintă grafic un instantaneu al stării setului exemplar de puncte de date din Fig. 1 după o iterație a buclei interioare a algoritmului fundamental MMCC.

[0020] Fig. 3 prezintă grafic un instantaneu al stării setului exemplar de puncte de date din Fig. 1 după o iterație a buclei exterioare a algoritmului fundamental MMCC.

[0021] Fig. 4 prezintă grafic setul de aglomerări identificate din setul de puncte de date din Fig. 1 la sfârșitul execuției algoritmului fundamental MMCC.

[0022] Fig. 5 este o reprezentare schematica a unui sistem în conformitate cu invenția de față.

5. Descriere Detaliata a Invenției

[0023] Documentul de față dezvăluie o nouă metodă de identificare de aglomerări de date denumită Identificare de Aglomerări prin Deplasarea Centrului de Masa (MMCC - Moving Mass Center Clustering). MMCC este o metodă de identificare de aglomerări de date pe bază de densitate capabilă să identifice toate aglomerările formate de către un set dat de puncte de date în conformitate cu un set de parametri de metodă.

[0024] Diversele aspecte și aplicații concrete ale invenției sunt descrise mai în detaliu mai jos.

5.1. Descrierea Problemei de Identificare de Aglomerări

[0025] Entitățile fizice din lumea reală care sunt clasificate de către metoda invenției de față sunt denumite generic obiecte. Datele de intrare ale metodei MMCC constau dintr-un set de obiecte, fiecare obiect având asociat un set de caracteristici, și în abordarea invenției de față fiecare caracteristică este presupusă ca măsurabilă de către o cantitate numerică asociată.

[0026] În practică nu întotdeauna există o cantitate numerică asociată la fiecare caracteristică a unui obiect, de exemplu în aplicațiile de minare de date unde baze de date cu diverse tipuri de date au nevoie sa fie analizate. O caracteristică dintr-o bază de date (denumită de asemenea câmp de tabel sau atribut) ar trebui sa corespundă la o dimensiune spațială din algoritmul MMCC. Problema consta în faptul ca unele seturi de date conțin pe lângă caracteristicile cantitative și caracteristici care țin de categorii sau binare și este dificil să se găsească o măsura de similitudine globală. Este necesară o fază de standardizare și rescalare cu scopul de a deduce o măsură de similitudine unificată. Există tehnici cunoscute de către persoanele cu calificare în domeniu pentru transformarea caracteristicilor care țin de categorii și a celor binare în caracteristici cantitative.

- Dacă caracteristica care ține de categorii are categoriile sale ordonate aceasta este denumită variabilă de rang, și poate fi ușor convertită la o caracteristică cantitativă prin atribuirea de valori numerice echidistante diferite la fiecare dintre categoriile acesteia și păstrând ordinea din punct de vedere numeric

- Dacă o caracteristică care ține de categorii are categoriile sale disjuncte, adică nici o entitate nu poate să fie clasificată în mai mult de una dintre acestea, și nu este ordonată, adică, caracteristicile pot fi numai comparate unele cu altele dacă coincid sau nu, aceasta este denumită o caracteristică

nominală. O caracteristică nominală poate fi convertită la un set de caracteristici cantitative, fiecare categorie definind o variabilă cantitativa *0-1* asupra entităților cu *1* corespunzând la prezența acesteia și *0* corespunzând la absența acesteia.

- O caracteristică calitativă este denumită ca fiind binară dacă are două categorii care pot fi considerate ca răspunsul Da (*TRUE*) sau Nu (*FALSE*) la o întrebare. O caracteristică binară poate fi convertită într-o caracteristică cantitativa prin conversia categoriei sale *TRUE* în *1* și a categoriei sale *FALSE* în *0*.

[0027] Presupunând că toate caracteristicile obiectelor din setul de date sunt numerice, fiecare obiect de date poate fi reprezentat de către un punct de date dintr-un interval al unui spațiu Euclidian multi-dimensional. Atunci problema de identificare de aglomerări poate fi formulată sau transformată în general la următoarea formulare:

Fiind dat un set de *N* puncte de date $P = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ într-un interval al spațiului vectorial Euclidian *n*-dimensional $D \subset \mathbf{R}^n$:

$$D = [x_{1,l}, x_{1,u}] \times [x_{2,l}, x_{2,u}] \times \ldots \times [x_{n,l}, x_{n,u}] \qquad (5.1)$$

$x_{j,l}, j = 1, \ldots, n$ fiind limitele inferioare și $x_{j,u}, j = 1, \ldots, n$ fiind limitele superioare care corespund la fiecare dimensiune,

$$\begin{aligned} \mathbf{x}_i &= (x_{i,1}, x_{i,2}, \ldots, x_{i,n}) \\ x_{1,l} &\le x_{i,1} \le x_{1,u}, \\ x_{2,l} &\le x_{i,2} \le x_{2,u}, \\ &\ldots \\ x_{n,l} &\le x_{i,n} \le x_{n,u}, \\ i &= 1, 2, \ldots, N \end{aligned}$$

a-2011-00279--
30-03-2011

problema constă în a determina aglomerările din setul de puncte de date *P* pe baza unui set de criterii pre-definite. Fig. 1 prezintă grafic un set exemplar de puncte de date utilizat pentru ilustrarea problemei de identificare de aglomerări tratată de către algoritmul fundamental MMCC.

5.2. Descriere Detaliată a Algoritmului Fundamental MMCC

[0028] În conformitate cu invenția de față, este fumizat un algoritm de identificare de aglomerări fundamental aplicat la seturi de puncte de date relaționate. Criteriul principal pentru identificarea unei aglomerări în algoritmul MMCC este variația de densitate într-un sub-domeniu hiper-sferic generic al domeniului problemei $D' \subseteq D$, sub-domeniul *D'* fiind determinat pe parcursul execuției algoritmului. Densitatea poate fi definită ca masa în unitatea de volum, considerând ca punctele au o cantitate de tip masă asociată cu acestea:

$$\rho(D') = \frac{\sum_{j \in I'} w_j}{V(D')} \qquad (5.2)$$

unde *I'* este mulțimea de indici ai punctelor de date din setul *P* localizate în interiorul hiper-sferei *D'*, $w_j \geq 0$ este ponderea asociată punctului $\mathbf{x}_j$ și *V* este funcția de volum. Ponderea este analogul masei atunci când se modelează sisteme fizice cu puncte de masă, în special în mecanica clasică. Semnificația ponderii depinde de problemă, de exemplu în probleme de optimizare (atunci când sunt aplicate metode de optimizare prin identificarea aglomerărilor) ponderea este relaționată cu valorile funcției obiectiv. Dacă nu există o cantitate de tip pondere asociată cu puncte de date (așa cum este cazul în majoritatea situațiilor întâlnite în practică, de exemplu în probleme de minare de date sau în probleme de analiză de imagine), se poate presupune ca toate ponderile sunt egale cu unitatea 1, și ecuația de mai sus se simplifică la:

$$\rho(D') = \frac{\sum_{j \in I'} 1}{V(D')} = \frac{|I'|}{V(D')} \qquad (5.3)$$

unde $|I'|$ este cardinalul mulțimii I' (sau cu alte cuvinte numărul de puncte de date din setul P localizate în interiorul sub-domeniului hiper-sferă D').

[0029] Alte criterii date ca parametri ai metodei sunt: numărul minim de puncte de date necesar pentru a forma o aglomerare nc_{min}, raportul de densitate minim r_{min} dintre densitatea punctelor de date dintr-un sub-domeniu hiper-sferic D' și densitatea medie a domeniului problemei ρ_{avg} pentru ca punctele de date din D' să poată fi considerate ca formând o aglomerare, raportul de eliminare de puncte K, și numărul de iterații consecutive n_{jmp} utilizate pentru detecția unui salt de creștere în densitate. Densitatea medie în domeniul problemei poate fi evaluată ca:

$$\rho_{avg} = \frac{\sum_{i=1}^{N} w_i}{\prod_{j=1}^{n} (x_{j,u} - x_{j,l})} \qquad (5.4)$$

[0030] Algoritmul are o buclă exterioară și o buclă interioară. În bucla exterioară, la sfârșitul unei iterații, punctele de date identificate ca formând o nouă aglomerare sunt eliminate total din setul curent de puncte de date P_{out}, în timp ce în bucla interioara punctele de date sunt eliminate numai temporar, fiind adăugate înapoi la începutul unei noi iterații a buclei exterioare.

[0031] În conformitate cu invenția de față pașii algoritmici ai metodei MMCC fundamentale sunt după cum urmează:

Pasul 1. Se inițializează datele problemei prin citirea informației în legătură cu setul de puncte de date P și a informației în legătură cu domeniul D. De asemenea se inițializează parametrii metodei nc_{min}, r_{min}, K și n_{jmp}. O buna selecție a valorilor

parametrilor nc_{min} și r_{min} depinde de problema analizată și câteodată algoritmul are nevoie sa fie re-executat cu parametrii modificați cu scopul de a îmbunătăți calitatea rezultatelor. Parametrul *K* depinde de problema analizata dar de asemenea și de funcția distanță utilizată și de dimensiunea spațiului *n*. O alegere buna pentru n_{jmp} descoperită în mod experimental este $n_{jmp} = 3$.

Pasul 2. Se inițializează setul de aglomerări identificate *C* la setul vid, se inițializează numărătorul de iterație al buclei exterioare $k_1 = 0$, se inițializează setul curent de puncte de date din bucla exterioară $P_{out}^{(k_1)} = P_{out}^{(0)} = P$ și se pornește bucla exterioară.

Pasul 3. Se verifică condiția de terminare a buclei exterioare:

$$|P_{out}^{(k_1)}| = 0 \qquad (5.5)$$

unde funcția |.| este funcția cardinal a mulțimii, adică se verifică dacă setul $P_{out}^{(k_1)}$ devine vid. Dacă condiția de terminare este satisfăcută atunci se termină bucla exterioară și se execută procesarea finală descrisă la *Pasul 16.*

Pasul 4. Se inițializează numărătorul de iterație al buclei interioare $k_2 = 0$, se inițializează setul curent de puncte de date din bucla interioară $P_{in}^{(k_2)} = P_{in}^{(0)} = P_{out}^{(k_1)}$ și se pornește bucla interioară.

Pasul 5. Se verifica condiția de terminare a buclei interioare:

$$|P_{in}^{(k_2)}| \leq nc_{min} \qquad (5.6)$$

adică dacă numărul de puncte de date ramase în setul $P_{in}^{(k_2)}$ devine mai mic sau egal cu numărul minim de puncte de date care pot forma o aglomerare. Dacă condiția de

terminare este satisfăcută atunci se salvează toate punctele de date din $P_{in}^{(k_2)}$ într-un set de eliminare $E_{in}^{(k_2)}$ și se termină bucla interioară.

Pasul 6. Se determină centrul de masă al punctelor de date aflate curent în $P_{in}^{(k_2)}$. Cu scopul de a simplifica notația considerăm mulțimea de indici ai punctelor de date din $P_{in}^{(k_2)}$, $I_{in}^{(k_2)} = \{i \in N : \mathbf{x}_i \in P_{in}^{(k_2)}\}$. Atunci centrul de masă este dat de:

$$\mathbf{x}_c^{(k_2)} = \frac{\sum_{i \in I_{in}^{(k_2)}} w_i \mathbf{x}_i}{\sum_{i \in I_{in}^{(k_2)}} w_i} \tag{5.7}$$

Notă: pentru eficiență numerică, în coduri de programare care implementează metoda MMCC, formula de mai sus nu va fi implementată în mod direct, și în loc de aceasta de fiecare dată când un punct de date este eliminat din $P_{in}^{(k_2)}$ (așa cum va fi explicat la *Pasul 9.*) noul centru de masă este calculat în mod recursiv în conformitate cu

$$\mathbf{x}_c^{(k_2)} = \frac{\mathbf{x}_c^{(k_2)} W - \mathbf{x}' w'}{W - w'}$$

$$W = W - w'$$

unde $W = \sum_{i \in I_{in}^{(k_2)}} w_i$ prin definiție și se presupune că punctul de date $\mathbf{x}'$ cu ponderea corespunzătoare w' este cel eliminat din $P_{in}^{(k_2)}$.

Pasul 7. Se determină punctul $\mathbf{x}_{i_{max}}^{(k_2)} \in P_{in}^{(k_2)}$ care este la cea mai mare distanță față de centrul de masă $\mathbf{x}_c^{(k_2)}$:

$$||\mathbf{x}_i - \mathbf{x}_c^{(k_2)}|| \leq ||\mathbf{x}_{i_{max}}^{(k_2)} - \mathbf{x}_c^{(k_2)}||, \ \mathbf{x}_i \in P_{in}^{(k_2)} \tag{5.8}$$

unde funcția ||.|| este o funcție matematică de tip normă definita pe spațiul Euclidian *n*-dimensional, $\|\cdot\| : \mathbf{R}^n \rightarrow \mathbf{R}$.

Pasul 8. Se consideră sub-domeniul $D_{in}^{(k_2)}$ format de către hiper-sfera cu centrul în $\mathbf{x}_c^{(k_2)}$ și raza $\|\mathbf{x}_{i_{max}}^{(k_2)} - \mathbf{x}_c^{(k_2)}\|$ care conține toate punctele de date din $P_{in}^{(k_2)}$ și se estimează densitatea în acest sub-domeniu ca:

$$\rho_m^{(k_2)} = \frac{\sum_{i \in I_{in}^{(k_2)}} w_i}{V(D_{in}^{(k_2)})} \qquad (5.9)$$

Pasul 9. Se elimină din $P_{in}^{(k_2)}$ toate punctele de date $\mathbf{x}_i$ care satisfac inegalitatea:

$$\frac{\|\mathbf{x}_i - \mathbf{x}_{i_{max}}^{(k_2)}\|}{\|\mathbf{x}_i - \mathbf{x}_c^{(k_2)}\|} < K, \ \mathbf{x}_i \in P_{in}^{(k_2)} \qquad (5.10)$$

Se salvează toate punctele de date eliminate la acest pas într-un set $E_{in}^{(k_2)}$. Se salvează de asemenea cu punctele de date eliminate centrul $\mathbf{x}_c^{(k_2)}$ și raza $\|\mathbf{x}_{i_{max}}^{(k_2)} - \mathbf{x}_c^{(k_2)}\|$ ale hiper-sferei curente. În unele aplicații (în special în aplicații de analiză de imagine, așa cum va fi explicat mai târziu) setul $E_{in}^{(k_2)}$ poate fi ordonat în ordinea crescătoare a distanței $\|\mathbf{x}_i - \mathbf{x}_c^{(k_2)}\|$. Pentru acest scop poate fi aplicat un algoritm de sortare prin inserție, unde poziția de inserție este determinată pentru fiecare punct de date nou adăugat prin aplicarea unui algoritm de căutare binară asupra punctelor de date deja existente în $E_{in}^{(k_2)}$.

Pasul 10. Se incrementează numărătorul buclei interioare $k_2 = k_2 + 1$ și se determină noul set:

$$P_{in}^{(k_2)} = P_{in}^{(k_2-1)} \setminus E_{in}^{(k_2-1)} \tag{5.11}$$

Fig. 2 prezintă un instantaneu al stării setului exemplar de puncte de date din Fig. 1 după execuția unui număr de iterații ale buclei interioare a algoritmului MMCC.

Pasul 11. Se continuă execuția la *Pasul 5.*

Pasul 12. Se determina limitele noii aglomerări. Dacă se notează cu $k_2^{(k_1)}$ valoarea maxima atinsa de către indicele k_2 înainte ca bucla interioară sa se termine, atunci cu scopul de a determina limitele noii aglomerări se calculează setul de rapoarte:

$$\{\frac{\rho_{in}^{(0)}}{\rho_{in}^{(n_{jmp})}}, \frac{\rho_{in}^{(1)}}{\rho_{in}^{(n_{jmp}+1)}}, \ldots \frac{\rho_{in}^{(k_2^{(k_1)}-n_{jmp})}}{\rho_{in}^{(k_2^{(k_1)})}}\} \tag{5.12}$$

și se ia în considerație valoarea maxima din set, sa presupunem ca aceasta este $\frac{\rho_{in}^{(m)}}{\rho_{in}^{(m+n_{jmp})}}$, pentru un indice m între *0* și $k_2^{(k_1)} - n_{jmp}$. Apoi se consideră hiper-sfera $D_{in}^{(m+n_{jmp})}$ ca frontieră a noii aglomerări (de asemenea o alta valoare de indice dintre m și $m + n_{jmp}$ poate fi selectată ca indice de frontieră pe baza unor diferite criterii, de exemplu valoarea maximă a densității, indicele median, etc., dar s-a descoperit experimental faptul că $m + n_{jmp}$ este o alegere buna).

Pasul 13. Se verifică dacă criteriul de densitate este satisfăcut:

$$\frac{\rho_{in}^{(m+n_{jmp})}}{\rho_{avg}} \geq r_{min} \tag{5.13}$$

Dacă criteriul de densitate este satisfăcut atunci punctele de date din setul $P_{in}^{(m+n_{jmp})}$ formează o noua aglomerare în conformitate cu criteriile metodei și aceasta este salvată în setul de aglomerări identificate *C*. Fig. 3 prezintă un instantaneu al stării

setului exemplar de puncte de date din Fig. 1 după ce o iterație a buclei exterioare a algoritmului MMCC cu identificarea unei noi aglomerări este terminată.

Notă: la acest pas setul $P_{in}^{(m+n_{jmp})}$ nu mai există, dar acesta poate fi reconstruit pe baza uniunii seturilor de puncte de date eliminate:

$$P_{in}^{(m+n_{jmp})} = E_{in}^{(m-n_{jmp})} \cup \ldots \cup E_{in}^{(k_2^{(k_1)}} \tag{5.14}$$

Pasul 14. Se incrementează numărătorul buclei exterioare $k_1 = k_1 + 1$ și se determină noul set exterior de puncte de date prin elirninarea din setul exterior de puncte de date a setului $P_{in}^{(m+n_{jmp})}$:

$$P_{out}^{(k_1)} = E_{in}^{(0)} \cup \ldots \cup E_{in}^{(m+n_{jmp}-1)} = P_{out}^{(k_1-1)} \setminus P_{in}^{(m+n_{jmp})} \tag{5.15}$$

Pasul 15. Se continua execuția la *Pasul 3.*

Pasul 16. După execuția pașilor de algoritm descriși mai sus se obțin în setul *C* toate aglomerările identificate în conformitate cu criteriile metodei cu fiecare aglomerare identificată limitată grosier într-o hiper-sferă cu centrul cunoscut și raza cunoscuta. Toate celelalte puncte de date din setul inițial de puncte de date *P* sunt eliminate ca fiind considerate date de zgomot. Fig. 4 prezintă setul de aglomerări identificate ale setului de puncte de date exemplar din Fig. 1 la sfârșitul execuției algoritmului MMCC fundamental. Punctele de date din aglomerări conținute în hiper-sfere care se intersectează se unesc cu scopul de a forma super-aglomerări. Pentru acest scop noi centre de masă și raze de hiper-sfere sunt calculate pentru super-aglomerări. De exemplu, presupunând că două super-sfere identificate cu centre c_1, c_2 și raze r_1, r_2 se intersectează, condiția de intersecție putând fi matematic exprimată ca

$$|r_1 - r_2| < d < r_1 + r_2 \tag{5.16}$$

unde d este distanța dintre centrele celor două hiper-sfere, $d = \|\mathbf{c}_1 - \mathbf{c}_2\|$, se poate demonstra faptul că super-hiper-sfera minimă care conține cele două hiper-sfere care se intersectează are raza:

$$r = \frac{d + r_1 + r_2}{2} \tag{5.17}$$

si centrul

$$\mathbf{c} = \lambda \mathbf{c}_1 + (1 - \lambda)\mathbf{c}_2 \tag{5.18}$$

unde

$$\lambda = \frac{1}{2}\left(1 + \frac{r_1 - r_2}{d}\right) \tag{5.19}$$

Dacă există mai mult de două hiper-sfere care se intersectează, acestea pot fi adăugate iterativ la noua super-hiper-sferă calculată prin repetarea procedurii descrise mai sus pentru două hiper-sfere.

[0032] Deși, așa cum a fost prezentată mai sus, metoda MMCC nu este proiectată ca un algoritm ierarhic, aceasta poate localiza ierarhii de aglomerări în mod implicit atunci când creșterea de densitate când se deplasează centrul de masă de la o aglomerare exterioară la o aglomerare interioară este semnificativă. În cazurile în care este localizată mai întâi aglomerarea interioară, aglomerarea exterioară va fi localizată într-o iterație ulterioară a buclei exterioare a algoritmului, dar și reciproca este de asemenea posibilă, adică, numai aglomerarea exterioară este localizată și aceasta conține aglomerări interioare nelocalizate. În ultimul caz algoritmul poate fi îmbunătățit astfel încât o nouă analiză de identificare de aglomerări MMCC este începută pentru domeniul aglomerării exterioare. Domeniul noii analize de identificare de aglomerări MMCC este în acest caz hiper-sfera care conține aglomerarea exterioara și densitatea medie este densitatea aglomerării exterioare. Această nouă analiza de identificare de

aglomerări MMCC poate fi executată în totală independență față de analiza de identificare de aglomerări MMCC curentă și există o oportunitate de implementare a acesteia ca un proces de calcul paralel pe mașini cu procesare paralelă.

[0033] Algoritmul fundamental MMCC prezentat mai sus produce ca rezultat final un set *C* de aglomerări globulare cu centre și raze cunoscute. Din perspectiva acestor rezultate algoritmul MMCC descris mai sus poate fi considerat o metodă de partiționare, deși rezultatele sale se bazează pe analiza de densitate. Sub forma prezentată mai sus metoda poate fi utilizata în aplicații de clasificare de date, având avantajele, comparativ cu alți algoritmi de partiționare (de exemplu, cei din familia K-means), că aceasta poate gestiona în mod eficient datele de zgomot, și că nu are nevoie de specificarea în avans a numărului de aglomerări pe care trebuie sa le identifice.

[0034] Algoritmi adaptați pentru domenii de aplicație specifice pot fi obținuți prin modificarea algoritmului fundamental prezentat mai sus, așa cum va fi dezvăluit în continuare.

[0035] Metoda dezvăluită este denumită Identificare de Aglomerări prin Deplasarea Centrului de Masă (MMCC - Moving Mass Center Clustering) deoarece în bucla interioară de fiecare data când unele puncte de date sunt eliminate, pe baza criteriului de raport de distanțe descris la *Pasul 9.*, centrul de masă al punctelor de date ramase se deplasează mai aproape de o posibilă aglomerare. În bucla exterioară, prin eliminarea punctelor de date ale unei aglomerări identificate sau a punctelor unui sub-domeniu detectat cu densitate mare (chiar dacă acesta nu este o aglomerare în conformitate cu criteriile metodei) se creează o oportunitate de identificare a unei noi posibile aglomerări la execuția următoare a buclei interioare, deoarece la execuția următoare a buclei interioare centrul de masă converge la un punct diferit față de cele la care acesta a convers la execuțiile anterioare ale buclei interioare.

5.3. Considerații legate de Distanță

[0036] O funcție de distanță (metrica) *d* pe un spațiu vectorial real *X* este definită în matematică ca o funcție $d: X \times X \rightarrow \mathbf{R}$ ($\mathbf{R}$ fiind mulțimea numerelor reale) care satisface pentru toate x, y, z din *X* condițiile:

1. $d(x, y) \geq 0$ (ne-negativitate)
2. $d(x, y) = 0$ dacă și numai dacă $x = y$ (identitatea punctelor indistinctibile)
3. $d(x, y) = d(y, x)$ (simetrie)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (inegalitatea triunghiului)

prima condiție fiind o consecință a celorlalte trei.

[0037] Dacă funcția de distantă pe spațiul vectorial *X* satisface suplimentar proprietățile:

5. $d(x, y) = d(x + a, y + a), a \in X$ (invarianță la translație)
6. $d(\alpha x, \alpha y) = |\alpha| d(x, y), \alpha \in \mathbf{R}$ (omogenitate)

atunci o funcție normă $\|.\|: X \rightarrow \mathbf{R}$ poate fi definită prin $\|x\| = d(x, 0)$, și în acest caz spațiul vectorial real *X* este denumit spațiu vectorial real normat.

[0038] Pentru $X = \mathbf{R}^n$ spațiul Euclidian *n*-dimensional, unele funcții distanță și funcții normă induse de către acele funcții distanță au o importanță particulară, toate fiind generalizate sub conceptul de distanță Minkovski definită ca:

$$\|x - y\|_p = \left(\sum_{j=1}^{n} |x_j - y_j|^p \right)^{\frac{1}{p}}, \quad p \in \mathbf{R}, \quad p > 0 \qquad (5.20)$$

Cazuri particulare sunt distanța Manhattan (*p*=1):

$$\|\mathbf{x}-\mathbf{y}\|_1 = \sum_{j=1}^{n} |x_j - y_j| \qquad (5.21)$$

Distanța Euclidiană (p=2):

$$\|\mathbf{x}-\mathbf{y}\|_2 = \left[\sum_{j=1}^{n} (x_j - y_j)^2\right]^{\frac{1}{2}} \qquad (5.22)$$

și distanța Chebyshev ($p = \infty$):

$$\|\mathbf{x}-\mathbf{y}\|_\infty = \max_{1 \le j \le n} |x_j - y_j| \qquad (5.23)$$

[0039] Conceptul de distanță poate fi extins la alte tipuri de date decât datele numerice pe alte tipuri de spatii vectoriale decât spațiul vectorial Euclidian, de exemplu distanța Hamming dintre șiruri de biți sau șiruri de caractere, dar pentru a fi capabili să extindem metoda MMCC la alte tipuri de date există de asemenea necesitatea de a defini conceptele de volum și densitate, ceea ce în general reprezintă sarcini teoretice dificile.

5.4. „Blestemul Dimensionalității"

[0040] Deși pentru metoda MMCC oricare definiție de distanță (și normă indusă de distanță) poate fi aplicată, distanța Minkovski are o importanță particulară, în special pentru aplicații de minare de date și pentru metodele de optimizare prin identificarea aglomerărilor, unde valorile lui n sunt mari și se întâmpină probleme datorită creșterii exponențiale a volumului asociată cu adăugarea de dimensiuni suplimentare la spațiul vectorial. Această problemă este cunoscută în domeniile optimizării, învățării automate și minării de date sub numele plastic de „blestemul dimensionalității".

[0041] Cu scopul de a obține o înțelegere a problemei „blestemului dimensionalității" avem nevoie sa comparăm în spațiul Euclidian *n*-dimensional cu *n* mare, volumul unei hiper-sfere cu raza unitate:

$$B_n(1) = \{\mathbf{x} \in \mathbf{R}^n : ||\mathbf{x}||_2 \leq 1\} \tag{5.24}$$

cu volumul unui hiper-cub cu laturi de lungime 2:

$$C_n(1) = \{\mathbf{x} \in \mathbf{R}^n : ||\mathbf{x}||_\infty \leq 1\} \tag{5.25}$$

Avem:

$$V(B_n(1)) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \tag{5.26}$$

unde $\Gamma(x)$ este cunoscuta funcție Gamma $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$, și

$$V(C_n(1)) = 2^n \tag{5.27}$$

și poate fi demonstrat că

$$\lim_{n\to\infty} \frac{V(B_n(1))}{V(C_n(1))} = \lim_{n\to\infty} \frac{\pi^{\frac{n}{2}}}{2^n\Gamma(\frac{n}{2}+1)} = 0 \tag{5.28}$$

astfel încât pe măsură ce dimensiunea *n* a unității de spațiu crește, hiper-sfera devine un volum nesemnificativ relativ la cel al hiper-cubului. Aceasta înseamnă că aproape tot spațiul unitar de dimensiune mare este „departe"' față de centru, sau că acesta constă aproape în întregime din „colțurile" hiper-cubului, cu aproape fără „mijloc".

[0042] Considerațiile de mai sus sugerează un remediu la problema „blestemului dimensionalității", avem nevoie să utilizăm o distanță care aduce hiper-sfera aproape de hiper-cub, adică avem nevoie să utilizam valori mari ale lui *p* în distanța Minkovski pentru *n* mare.

5.5. Volumul Hiper-Sferei Minkovski

[0043] Definim hiper-sfera Minkovski prin aplicarea distanței Minkovski:

$$B_{n,p}(r) = \{\mathbf{x} \in \mathbf{R}^n : ||\mathbf{x}||_p \leq r\} \tag{5.29}$$

În metoda MMCC avem nevoie sa calculăm volumul hiper-sferei Minkovski care va fi utilizat pentru evaluarea densității. Volumul hiper-sferei Minkovski este dat de către integrala:

$$V(B_{n,p}(r)) = \int_{||\mathbf{x}||_p \leq r} dx_1 dx_2 \dots dx_n = \int_{\left(\sum_{j=1}^{n} |x_j|^p\right)^{\frac{1}{p}} \leq r} dx_1 dx_2 \dots dx_n \tag{5.30}$$

Dacă introducem pentru domeniul *n*-dimensional notația

$$D_{n,p}(1) = \{\mathbf{x} \in \mathbf{R}^n : \left(\sum_{j=1}^{n} x_j^p\right)^{\frac{1}{p}} \leq 1,\ x_1 \geq 0, x_2 \geq 0, \dots x_n \geq 0\} \tag{5.31}$$

atunci din considerații de simetrie și omogenitate integrala de mai sus poate fi simplificată la

$$V(B_{n,p}(r)) = r^n V(B_{n,p}(1)) = 2^n r^n \int_{D_{n,p}(1)} dx_1 dx_2 \dots dx_n = 2^n r^n I_1(n,p) \tag{5.32}$$

unde

$$I_1(n,p) = \int_{D_{n,p}(1)} dx_1 dx_2 \ldots dx_n \qquad (5.33)$$

prin definiție. $I_1(n,p)$ este o integrală dificilă pentru a fi calculată analitic, dar numeric aceasta poate fi tratată prin aplicarea metodelor Monte-Carlo, în special a metodelor Quasi-Monte Carlo (QMC) pentru o precizie mai mare.

[0044] Metodele de integrare QMC aplică șiruri de numere cvasi-aleatoare care caută să acopere golul dintre flexibilitatea generatoarelor de numere pseudo-aleatoare și avantajele pe care rețelele regulate le aduc preciziei de integrare. Exemple de șiruri cvasi-aleatoare cunoscute în domeniu sunt: șirurile Halton, șirurile Sobol, șirurile Niederreiter și șirurile de hiper-cub Latin.

[0045] Poate fi demonstrat teoretic faptul că pentru un spațiu Euclidian *n*-dimensional metodele Quasi-Monte Carlo dau o eroare deterministă de ordinul $O(\frac{(\log N_s)^{n-1}}{N_s})$, N_s fiind numărul de puncte de eșantionare. Se observă că chiar și prin aplicarea metodelor Quasi-Monte Carlo, odată cu creșterea lui *n* numărul de puncte de eșantionare N_s necesare pentru o precizie bună a integralei $I_1(n,p)$ crește imens. Din fericire poate fi demonstrată o relație de recurență foarte utilă pentru $I_1(n,p)$:

$$I_1(n,p) = I_1(n-2,p) I_2(n-2,p) \qquad (5.34)$$

unde $I_2(m,p)$ este definită prin:

$$I_2(m,p) = \int_{D_{2,p}(1)} (1 - x_1^p - x_2^p)^{\frac{m}{p}} dx_1 dx_2 \qquad (5.35)$$

Pentru *n* impar avem, prin extinderea relației de recurenta de mai sus:

$$I_1(n,p) = I_2(n-2,p)I_2(n-4,p)\ldots I_2(3,p)I_2(1,p)I_1(1,p) = I_2(n-2,p)I_2(n-4,p)\ldots I_2(3,p) \quad (5.36)$$

considerând faptul că $I_1(1,p) = 1$. În mod similar, pentru *n* par avem:

$$I_1(n,p) = I_2(n-2,p)I_2(n-4,p)\ldots I_2(4,p)I_2(2,p)I_1(2,p) \quad (5.37)$$

În acest mod se poate reduce calculul integralelor $I_1(n,p)$ numai la domenii bidimensionale, și pentru domenii bidimensionale putem obține o precizie acceptabilă pentru metodele Quasi-Monte Carlo cu un număr relativ scăzut de puncte de eșantionare în conformitate cu eroarea determinista $O(\frac{\log N_s}{N_s})$.

[0046] Valorile lui $I_1(n,p)$ pentru diferite valori ale lui *n* și *p* pot fi precalculate și încărcate în tablouri statice de către programele de calculator care implementează metoda MMCC, în acest mod economisind timp de calcul și îmbunătățind eficienta programelor de calculator.

5.6. Considerații în privința Eficienței

[0047] În multe metode de identificare de aglomerări, în special metode pe bază de densitate și metode ierarhice o problemă de eficiență este legată de numărul mare de distanțe pereche dintre punctele de date care este necesar să fie calculate. În general numărul de distanțe pereche este de ordinul $O(N^2)$, *N* fiind numărul de puncte de date, ceea ce devine prohibitiv pentru seturi mari de puncte de date și dimensiuni mari. În această secțiune voi demonstra faptul că pentru algoritmul MMCC numărul de distanțe dintre punctele de date care este necesar sa fie calculate este de ordinul $O(N)$, adică liniar cu *N*.

[0048] Cu scopul de a obține o estimare a complexității temporale a algoritmului MMCC putem presupune faptul ca rata de eliminare de puncte de date a buclei

exterioare a algoritmului MMCC este r_1 $(0 \leq r_1 \leq 1)$ și că rata de eliminare de puncte de date a buclei interioare a algoritmului MMCC este r_2 $(0 \leq r_2 \leq 1)$.

Rata r_1 este dependentă de dimensiunile aglomerărilor din setul de puncte de date P, o limită inferioară pentru r_2 fiind $\frac{nC_{typical}}{N}$.

Rata r_2 este puternic dependentă de parametrul raport de eliminare de puncte K, valori mari pentru K generând valori mari pentru r_2. Pentru un spațiu bidimensional (n=2, plan) rezultate bune sunt obținute pentru $K \leq 2$, dar odată cu creșterea dimensiunii n valorile bune pentru K cresc exponențial, variația exponențială fiind relaționată cu creșterea exponențială a volumului hiper-sferei Minkovski n-dimensionale.

[0049] Cu scopul de a simplifica evaluarea putem presupune ratele r_1 și r_2 constante pe durata execuției algoritmului MMCC. Sa notam cu $N^{(k_1,k_2)}$ numărul de puncte de date care rămân pentru procesare la pasul buclei exterioare k_1 și la pasul buclei interioare k_2. Avem:

$$N^{(k_1,k_2)} = N^{(k_1,0)}(1-r_2)^{k_2} = N^{(0,0)}(1-r_1)^{k_1}(1-r_2)^{k_2} = N(1-r_1)^{k_1}(1-r_2)^{k_2} \qquad (5.38)$$

Numărul de evaluări de distanțe la fiecare pas al buclei interioare este $2N^{(k_1,k_2)} - 1 \approx 2N^{(k_1,k_2)}$ deoarece avem nevoie să calculăm distanțele de la fiecare punct la centrul de masă și la punctul localizat la distanța maximă față de centrul de masă. în acest mod numărul total de distanțe poate fi evaluat ca:

$$\sum_{k1,k2} N^{(k_1,k_2)} \leq \sum_{k1=0}^{\infty}\sum_{k2=0}^{\infty} N^{(k_1,k_2)} = 2N\sum_{k1=0}^{\infty}(1-r_1)^{k_1}\sum_{k2=0}^{\infty}(1-r_2)^{k_2} = \frac{2N}{r_1 r_2} = O(N) \qquad (5.39)$$

ceea ce demonstrează dependenta liniara a complexității temporale a algoritmului MMCC față de numărul de puncte de date N.

6. Exemple de Aplicații ale Invenției

[0050] Metoda MMCC prezintă complexitate temporală și spațială liniară cu dimensiunea datelor, este capabilă să localizeze aglomerări de forme complexe cu dimensiuni și densități care variază, tratează în mod eficient datele de zgomot și furnizează stabilitate și fiabilitate. Datorită lipsei unei probleme de optimizare asociată cu metoda în legătură cu minimizarea unei funcții obiectiv globale MMCC evită dificultatea legată de încercarea de rezolvare a unei probleme de optimizare combinatorială dificile. Nu există nevoia de a specifica, ca un parametru al metodei, numărul de aglomerări care trebuie sa fie produse, MMCC fiind capabila sa localizeze toate aglomerările în conformitate cu parametrii metodei care sunt în principal în legătură cu densitatea.

[0051] În această secțiune sunt evidențiate avantajele algoritmului de identificare a aglomerărilor MMCC față de alte metode și algoritmi convenționali de identificare a aglomerărilor prin prezentarea de diverse aplicații concrete exemplare suplimentare și aplicații ale invenției prezente în unele domenii de aplicație generice. Metoda MMCC combină caracteristicile metodelor de identificare de aglomerări partiționale și pe bază de densitate în același timp eliminând unele dintre neajunsurile pe care metodele menționate le prezintă. În același timp metoda MMCC este capabilă de clasificare ierarhică a datelor în mod similar cu metodele ierarhice de identificare a aglomerărilor.

6.1. Minare de Date

[0052] În conformitate cu o primă aplicație concretă a invenției de față, metoda MMCC poate fi aplicată la aplicații de Minare de Date (Eng., Data Mining).

[0053] Analiza de Identificare de Aglomerări așa cum este utilizată în minarea de date este un instrument pentru găsirea de modele și regularități în date. O dificultate poate apare în aplicarea metodei MMCC în minarea de date deoarece aceasta este proiectată pentru a trata date cantitative (adică cu valori reale). O caracteristica dintr-o

bază de date (denumită de asemenea câmp de tabel sau atribut) va corespunde la o dimensiune spațială în algoritmul MMCC. Există o problema atunci când unele seturi de date conțin pe lângă caracteristicile cantitative, de asemenea caracteristici de categorii și binare și este dificil sa se găsească pentru acestea o măsură de similitudine globală. Există o nevoie de o fază de pre-procesare care constă în standardizarea și rescalarea datelor cu scopul de a stabili o măsură de similitudine unificată. Există tehnici cunoscute în domeniu pentru transformarea caracteristicilor de categorii și binare în caracteristici cantitative.

[0054] Unele avantaje ale metodei MMCC comparativ cu alte metode de identificare de aglomerări convenționale din domeniul minării de date constă în faptul ca aceasta este rapidă, stabilă, fiabilă și capabilă sa elimine zgomotul. Datorita stabilității sale aceasta poate fi re-executata din timp în timp cu scopul de a identifica noi aglomerări formate, fără a pierde nici una dintre aglomerările deja localizate.

6.2. Învățarea Automată

[0055] În conformitate cu o a doua aplicație concretă a invenției, metoda MMCC poate fi aplicată la aplicații de Învățare Automata (Eng., Machine Learning).

[0056] În Învățarea Automată problema centrală a Analizei de Identificare de Aglomerări este predicția mai degrabă decât adaptarea de model sau identificarea de modele. Învățarea Automata tinde sa considere datele ca un instrument pentru a învăța cum să se prezică categorii pre-specificate sau nou create. Entitățile sunt considerate ca venind cate una la un moment de timp astfel încât mașina poate învăța în mod adaptiv printr-o modalitate supervizată.

[0057] Problema principala este că în cazul algoritmului MMCC, așa cum acesta este proiectat, este nevoie ca toate datele de antrenament să fie disponibile în același timp cu scopul de a identifica categoriile de date. După procesul de antrenament inițial, pe durata căruia aglomerările sunt identificate, noile date pot fi cu ușurință categorizate

în conformitate cu aglomerările existente și zgomotul poate fi eliminat prin filtrare cu ușurință deoarece pentru fiecare aglomerare exista asociate un centru și o raza. Centrul și raza pentru fiecare aglomerare pot fi actualizate în mod dinamic prin adăugarea de date noi la aglomerare în conformitate cu un criteriu de densitate. De asemenea algoritmul de identificare de aglomerări poate fi re-executat pentru re-antrenament din timp în timp cu scopul de a identifica noi aglomerări formate și de a îmbunătăți calitatea învățării (noile date de zgomot nu sunt eliminate, acestea sunt numai etichetate ca atare).

6.3. Analiza de Imagine

[0058] În conformitate cu o a treia aplicație concretă a invenției de față, metoda MMCC poate fi aplicată în aplicații de Analiză de Imagine (Eng., Image Analysis).

[0059] În analiza de imagine este importantă posibilitatea de a aplica analiza de identificare de aglomerări cu scopul de a identifica obiecte de diferite dimensiuni, forme și densități. Pentru acest scop un număr de algoritmi de identificare de aglomerări pe bază de densitate sunt deja dezvoltați și sunt aplicați cu diverse rezultate, de exemplu cei din familia de algoritmi DBSCAN (DBSCAN, IDBSCAN, GDBSCAN, etc.). Principala problema cu care acești algoritmi se confruntă este în legătură cu nevoia de a calcula distantele dintre perechile de puncte de date din seturi mari de puncte de date. Sunt aplicate diverse îmbunătățiri cu scopul de a reduce numărul de distante calculate, și în acest mod, a crește eficienta metodelor. O altă problemă importantă pe care aceștia o întâmpină este dificultatea de a trata aglomerările cu densități variabile în domenii largi de valori datorită faptului ca parametrii metodei sunt definiți în mod global pentru spațiul de date. Metoda MMCC poate ajuta metodele pe bază de densitate existente pentru a trata eficient problemele menționate.

[0060] Comparativ cu alți algoritmi convenționali de identificare de aglomerări pe bază de densitate, algoritmul MMCC are unele caracteristici avantajoase din construcție. Mai întâi prin identificarea de aglomerări grosiere conținute în hiper-sfere cu

centre şi raze cunoscute acesta oferă un set bun de puncte de început pentru extinderea aglomerărilor identificate la aglomerări mai mari de forme, dimensiuni şi densităţi arbitrare. Punctele de date din hiper-sfere pot fi analizate statistic cu scopul de a determina proprietăţile statistice ale distribuţiei variabilelor *k*-dist (distanţa de la un punct la cel mai apropiat *k*-lea vecin). Pentru acest scop, iniţial este necesar sa fie calculate numai distantele dintre perechi de puncte de date care aparţin la aglomerările globulare identificate. Valoarea lui *k* nu poate fi mai mare decât dimensiunea aglomerării globulare, o valoare *k*=4 fiind considerată ca o alegere rezonabila pentru seturi de date bidimensionale de către algoritmii din familia DBSCAN. Analiza statistica este utilizată cu scopul de a determina parametrii algoritmului de tip DBSCAN care va fi utilizat mai departe, dar principala diferenţă consta în faptul că de această dată parametrii sunt locali pentru aglomerarea curentă, nu globali aşa cum au fost înainte în cazul algoritmilor convenţionali. După aceea punctele de date cu proprietăţi statistice similare din seturile de puncte de date eliminate $E_{in}^{(k_2)}$ pot fi analizate pentru adăugare la aglomerare, seturile de puncte de date eliminate fiind analizate în ordine inversă faţă de cea a creaţiei lor. Aşa cum a fost explicat la *Pasul 9.* al algoritmului MMCC, ca un al doilea avantaj din construcţie al metodei MMCC, pentru aplicaţii de analiză de imagine setul $E_{in}^{(k_2)}$ poate fi ordonat în ordinea crescătoare a distantei de la punctele de date la centrul de masa $||\mathbf{x}_i - \mathbf{x}_c^{(k_2)}||$. Aceasta este ordinea în care punctele de date eliminate sunt analizate pentru adăugare la aglomerare. Pentru acest scop este nevoie numai de distanţele dintre punctele de date din setul de eliminare analizat în mod curent $E_{in}^{(k_2)}$ şi un număr de seturi de eliminare analizate anterior $E_{in}^{(k_2+1)} \ldots E_{in}^{(k_2+n_e)}$, n_e fiind un parametru de metodă suplimentar. Procesul continuă în sens invers până când este găsit un set de eliminare care nu poate contribui cu nici un punct de date la aglomerare. La sfârşitul unei iteraţii a buclei exterioare toate punctele de date din aglomerarea extinsă sunt eliminate, în acest mod contribuind la o rata de eliminare mai mare şi la o eficienţă mai bună a algoritmului.

[0061] Algoritmul MMCC modificat pentru analiza de imagine, aşa cum a fost prezentat mai sus este capabil sa localizeze aglomerări amestecate de diferite densităţi,

ceea ce este o sarcină dificila pentru alte metode pe bază de densitate convenționale. Pentru acest scop acesta poate fi modificat astfel încât odată ce o aglomerare globulară este localizată, o noua analiza de identificare de aglomerări va fi pornită pentru domeniul aglomerării globulare localizate cu scopul de a localiza sub-aglomerări de densitate înaltă (densitatea medie pentru noua analiza de identificare de aglomerări fiind densitatea aglomerării globulare curente). Acest proces poate fi dezvoltat în mod recursiv pe multe nivele pana când nu mai pot fi localizate noi sub-aglomerări de densitate mai înaltă. Odată ce toate sub-aglomerările pe diferite nivele sunt localizate, metode pe bază de densitate specializate, așa cum au fost menționate mai sus, pot fi aplicate pentru a extinde sub-aglomerările localizate în ordine inversă față de ordinea în care au fost localizate (care este o ordine naturală în procesele recursive).

6.4. Optimizare prin Identificarea de Aglomerări

[0062] În conformitate cu o a patra aplicație concretă a invenției de față, metoda MMCC poate fi aplicată în aplicații de Optimizare prin Identificarea de Aglomerări (Eng., Clustering Optimization).

[0063] Metodele de Optimizare prin Identificarea de Aglomerări au fost utilizate pentru creșterea eficienței în selecția punctelor de început în metode de optimizare globală cu puncte de început multiple pentru funcții numerice cu mai multe variabile. Metodele de acest tip au în mod obișnuit trei pași:

(a) eșantionarea de puncte în regiunea de interes;
(b) transformarea punctelor eșantionate pentru a obține puncte grupate în vecinătăți de puncte de extrem local;
(c) utilizarea unei tehnici de identificare de aglomerări pentru a obține aglomerările.

După ce aglomerările sunt identificate cu succes, punctele de minim local (și astfel de asemenea minimul global) pot fi determinate cu ușurința prin aplicarea unei metode de optimizare locală pentru fiecare aglomerare.

[0064] Metoda MMCC dezvăluită aici poate fi utilă la cel de-al treilea pas al metodei generale de optimizare prin identificarea de aglomerări prezentată mai sus în special, deoarece aceasta este o metodă pe bază de densitate, și utilizează ponderile asociate cu punctele de date cu scopul de a evalua centrele de masă și densitățile locale. Ponderile asociate cu punctele de date pot lua valori pozitive intre *0* și *1* calculate prin scalarea inversă a valorilor funcției obiectiv pentru un domeniu de căutare între o valoare de funcție minima și o valoare de funcție maxima. În acest mod este posibil sa se identifice posibilele minime locale și regiunile de atracție ale acestora, chiar și atunci când punctele de eșantionare sunt generate prin aplicarea de metode de eșantionare uniforme, cum ar fi metode Quasi-Monte Carlo cunoscute. Regiunile de atracție identificate la un nivel pot fi rafinate la nivele mai adânci prin generarea de puncte de eșantionare locale mai dense. Regiunile de atracție identificate pot fi analizate în mod independent și prezintă o oportunitate de execuție în paralel a metodei.

6.5. Sistem de Calcul

[0065] Invenția de față este o metodă pe bază de calculator și în conformitate cu un aspect al invenției, aceasta poate lua forma unui sistem de calcul de scop general sau de scop special care implementează metoda.

[0066] Fig. 5 reprezintă o reprezentare schematică a unui sistem în conformitate cu invenția de față. Sistemul 1 include un procesor 2, o memorie 3, un mediu care poate fi citit de calculator 4 și un dispozitiv de ieșire 5. Sistemul poate lua forma unui, sau poate include un, calculator de scop general sau de scop special. Procesorul 2 poate lua orice forma și memoria 3 poate fi separata sau integrată cu acesta. Dispozitivul de

ieșire 5 poate fi un afișaj, o imprimată, un dispozitiv de înregistrare sau de stocare, sau poate lua alte forme. Mediul care poate fi citit de calculator 4 poate lua orice formă și de preferință are instrucțiuni executabile de către calculator, cum ar fi soft, încorporate pe acesta pentru efectuarea unuia sau mai multora dintre algoritmii metodei MMCC descrise aici. Instrucțiunile executabile pe calculator sunt efectuate de către procesor. Mediul 4 poate fi integrat cu restul sistemului, sau poate fi separat de acesta. Invenția include suplimentar un mediu care poate fi citit pe calculator, cum ar fi un mediu de stocare magnetic sau optic (de exemplu, ROM, RAM, CD-uri, discuri dure, etc.), având un program, așa cum este descris aici, concretizat pe mediu. Datele care trebuie sa fie procesate de către metodă sau sistem pot fi introduse de către un utilizator, furnizate ca ieșire de la un alt dispozitiv sau sistem, sau pot fi stocate pe un mediu care poate fi citit de calculator sau de un dispozitiv de calcul.

[0067] În conformitate cu un alt aspect al invenției de față, aceasta poate lua forma unui mediu care poate fi citit de calculator având instrucțiuni care pot fi citite de calculator, cum ar fi soft, concretizate pe acesta pentru efectuarea metodei pe bază de calculator MMCC.

6.6. Remarci Finale privind Aplicațiile Invenției

[0068] Trebuie sa se înțeleagă faptul ca scopul invenției de față se extinde la toate aplicațiile practice ale identificării de aglomerări. Aceste aplicații includ, dar nu sunt limitate la, recunoașterea de modele, predicția de serii temporale, teoria învățării, astrofizică, aplicații medicale inclusiv de imagine și de procesare de date, partiționare de rețele, comprimare de imagine, strângere de date prin satelit, gestionare de baze de date, minare de baze de date, analiză de baze de date, recunoaștere automata a țintei și recunoaștere de vorbire și de text.

[0069] Descrierea anterioară a aplicațiilor concrete dezvăluite este furnizată pentru a permite oricărei persoane cu calificare în domeniu să realizeze sau sa utilizeze invenția de față. Numeroase variațiuni suplimentare și modificări ale aplicațiilor concrete

vor fi evidente imediat pentru cei cu calificare în domeniu, și principiile generice definite aici pot fi aplicate și la alte aplicații concrete fără a se îndepărta de la scopul revendicărilor. Astfel dezvăluirea de față nu este intenționată a fi limitată la exemplele și aplicațiile concrete prezentate aici.

Revendicări

1. Metodă pe bază de calculator de identificare a aglomerărilor unui set de puncte de date relaționate dintr-un interval Euclidian multi-dimensional, denumită Identificare de Aglomerări prin Deplasarea Centrului de Masa (MMCC - Moving Mass Center Clustering), algoritmul fundamental al metodei MMCC cuprinzând următorii pași:

Pasul 1. se citește setului de puncte de date al problemei *P* și informația legată de domeniului *D*, se inițializează parametrii de identificare de aglomerări nc_{min}, r_{min}, *K* și n_{jmp};

Pasul 2. se inițializează setul de aglomerări identificate *C* la mulțimea vidă, se inițializează numărătorul de iterație al buclei exterioare $k_1 = 0$, se inițializează setului curent de puncte de date din bucla exterioara $P_{out}^{(k_1)} = P_{out}^{(0)} = P$ și se pornește bucla exterioară;

Pasul 3. se verifică condiția de terminare a buclei exterioare

$$|P_{out}^{(k_1)}| = 0$$

adică se verifică dacă setul $P_{out}^{(k_1)}$ devine vid și dacă condiția de terminare este satisfăcută atunci se termină bucla exterioară și se execută procesarea finală descrisă la *Pasul 16.*;

Pasul 4. se inițializează numărătorul de iterație al buclei exterioare $k_2 = 0$, se inițializează setului curent de puncte de date din bucla interioară $P_{in}^{(k_2)} = P_{in}^{(0)} = P_{out}^{(k_1)}$ și se pornește bucla interioară;

Pasul 5. se verifica condiția de terminare a buclei interioare

$$|P_{in}^{(k_2)}| \leq nc_{min}$$

adică dacă numărul de puncte de date rămas în setul $P_{in}^{(k_2)}$ devine mai mic decât sau egal cu numărul minim de puncte de date care pot forma o aglomerare, și dacă condiția de terminare este satisfăcută atunci se salvează toate puncte de date din $P_{in}^{(k_2)}$ într-un set de eliminare $E_{in}^{(k_2)}$ și se termină bucla interioară;

Pasul 6. se determina centrul de masa al punctelor de date aflate în mod curent în $P_{in}^{(k_2)}$:

$$\mathbf{x}_c^{(k_2)} = \frac{\sum_{i \in I_{in}^{(k_2)}} w_i \mathbf{x}_i}{\sum_{i \in I_{in}^{(k_2)}} w_i}$$

Unde $I_{in}^{(k_2)} = \{i \in N : \mathbf{x}_i \in P_{in}^{(k_2)}\}$ este setul de indici ai punctelor de date din $P_{in}^{(k_2)}$;

Pasul 7. se determina punctul $\mathbf{x}_{i_{max}}^{(k_2)} \in P_{in}^{(k_2)}$ care este la cea mai mare distanță față de centrul de masă $\mathbf{x}_c^{(k_2)}$

$$||\mathbf{x}_i - \mathbf{x}_c^{(k_2)}|| \leq ||\mathbf{x}_{i_{max}}^{(k_2)} - \mathbf{x}_c^{(k_2)}||. \ \mathbf{x}_i \in P_{in}^{(k_2)}$$

Pasul 8. se consideră sub-domeniul $D_{in}^{(k_2)}$ format de către hiper-sfera cu centru în $\mathbf{x}_c^{(k_2)}$ și raza $||\mathbf{x}_{i_{max}}^{(k_2)} - \mathbf{x}_c^{(k_2)}||$ care conține toate punctele de date din $P_{in}^{(k_2)}$ și se estimează densitatea acestui sub-domeniu ca

$$\rho_{in}^{(k_2)} = \frac{\sum_{j \in I_{in}^{(k_2)}} u_j}{V(D_{in}^{(k_2)})}$$

Pasul 9. se elimină din $P_{in}^{(k_2)}$ toate punctele de date $\mathbf{x}_i$ care satisfac inegalitatea

$$\frac{\|\mathbf{x}_i - \mathbf{x}_{i_{max}}^{(k_2)}\|}{\|\mathbf{x}_i - \mathbf{x}_c^{(k_2)}\|} < K, \ \mathbf{x}_i \in P_{in}^{(k_2)}$$

și se salvează toate puncte de date eliminate la acest pas într-un set $E_{in}^{(k_2)}$, și odată cu punctele de date eliminate sunt de asemenea salvate centrul $\mathbf{x}_c^{(k_2)}$ și raza $\|\mathbf{x}_{i_{max}}^{(k_2)} - \mathbf{x}_c^{(k_2)}\|$ ale hiper-sferei $D_{in}^{(k_2)}$;

Pasul 10. se incrementează numărătorul buclei interioare $k_2 = k_2 + 1$ și se determină noul set

$$P_{in}^{(k_2)} = P_{in}^{(k_2-1)} \setminus E_{in}^{(k_2-1)}$$

Pasul 11. se continuă execuția algoritmului la *Pasul 5.*;

Pasul 12. se calculează setului de rapoarte

$$\{\frac{\rho_{in}^{(0)}}{\rho_{in}^{(n_{jmp})}}, \frac{\rho_{in}^{(1)}}{\rho_{in}^{(n_{jmp}+1)}}, \ldots \frac{\rho_{in}^{(k_2^{(k_1)}-n_{jmp})}}{\rho_{in}^{(k_2^{(k_1)})}}\}$$

unde $k_2^{(k_1)}$ este valoarea maximă atinsă de către k_2 înainte ca bucla interioară să fie terminată, și considerarea valorii maxime pentru acest set $\frac{\rho_{in}^{(m)}}{\rho_{in}^{(m+n_{jmp})}}$, pentru un indice m

intre 0 și $k_2^{(k_1)} - n_{jmp}$, și apoi se consideră ca frontieră a noii aglomerări a hiper-sferei $D_{in}^{(m+n_{jmp})}$, sau se selectează o alta valoare de indice între *m* și *m* + n_{jmp} ca indice de frontieră pe baza unor diferite criterii, de exemplu valoarea maximă a densității, indicele median, etc.;

Pasul 13. se verifică dacă criteriul densității este satisfăcut

$$\left| \frac{\rho_{in}^{(m+n_{jmp})}}{\rho_{avg}} \geq r_{min} \right|$$

și dacă criteriul densității este satisfăcut atunci, se salvează punctele de date din setul $P_{in}^{(m:m+n_{jmp})}$ ca o nouă aglomerare identificată în conformitate cu parametrii metodei în setul de aglomerări identificate *C*;

Pasul 14. se incrementează numărătorul buclei exterioare $k_1 = k_1 + 1$ și se determină un nou set exterior de puncte de date

$$\left| P_{out}^{(k_1)} = E_{in}^{(0)} \cup \ldots \cup E_{in}^{(m+n_{jmp}-1)} = P_{out}^{(k_1-1)} \setminus P_{in}^{(m:m+n_{jmp})} \right|$$

Pasul 15. se continua execuția algoritmului de la *Pasul 3.*;

Pasul 16. se unesc punctele de date de aglomerare conținute în hiper-sfere care se intersectează în super-aglomerări, pentru acest scop calculându-se noi centre de masă și raze de hiper-sferă pentru super-aglomerări prin repetarea iterativă a procedeului pentru două hiper-sfere, în care în cazul a două hiper-sfere care se intersectează cu centrele c_1, c_2 și razele r_1, r_2, super-hiper-sfera minimă care conține cele două hiper-sfere care se intersectează are raza

$$r = \frac{d + r_1 + r_2}{2}$$

și centrul

$$\mathbf{c} = \lambda \mathbf{c_1} + (1 - \lambda)\mathbf{c_2}$$

unde *d* este distanța dintre centrele celor două hiper-sfere, $d = ||\mathbf{c_1} - \mathbf{c_2}||$, și

$$\lambda = \frac{1}{2}\left(1 + \frac{r_1 - r_2}{d}\right)$$

2. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 1 în care algoritmul este îmbunătățit astfel încât o nouă analiză de identificare de aglomerări MMCC este pornită pentru domeniile aglomerărilor localizate cu scopul de a localiza aglomerările interioare, domeniul noii analize de identificare de aglomerări MMCC fiind în acest caz hiper-sfera care conține aglomerarea exterioară și densitatea medie fiind densitatea aglomerării exterioare, în care această nouă analiză de identificare a aglomerărilor MMCC este executată total independent față de analiza de identificare a aglomerărilor curentă MMCC, fiind posibil ca aceasta sa fie implementată ca un proces de calcul paralel pe mașini de procesare paralelă.

3. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 1 în care metoda este adaptată pentru aplicații în Minarea de Date (Eng., Data Mining), adaptarea constând într-o fază de standardizare și rescalare de date cu scopul de a stabili o măsură de similitudine unificată.

4. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 1 în care metoda este adaptată pentru aplicații în Învățarea Automată (Eng., Machine Learning), adaptarea constând în analiza tuturor datelor de antrenament inițial în același timp în scopul identificării categoriilor de date.

5. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 1 în care metoda este adaptată pentru aplicații în Analiza de Imagine (Eng., Image Analysis), adaptarea constând în adăugarea unui pas de procesare suplimentar la *Pasul 9.* al algoritmului fundamental în conformitate cu Revendicarea 1, în care procesarea suplimentară consta în ordonarea setului $E_{in}^{(k_2)}$ în ordinea crescătoare a distanței $\|x_i - x_i^{(k_2)}\|$ prin utilizarea unui algoritm simplu de sortare prin inserție în care poziția de inserție este determinată pentru fiecare nou punct adăugat prin aplicarea unui algoritm de căutare binară.

6. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 5 în care metoda este mai departe adaptată pentru aplicații în Analiza de Imagini, adaptarea constând în adăugarea unui pas de procesare suplimentar după localizarea unei noi aglomerări conținute intr-o hiper-sferă cu centru și rază cunoscute la *Pasul 13.* al algoritmului fundamental în conformitate cu Revendicarea 1, unde la acest pas de procesare suplimentar punctele de date din hiper-sfere sunt analizate statistic cu scopul de a determina proprietățile statistice ale distribuției unei variabile *k*-dist cu *k* ales în mod adecvat, unde analiza statistică este utilizată cu scopul de a determina parametrii algoritmului de tip DBSACN care este utilizat mai departe, unde după aceea punctele de date cu proprietăți statistice similare din seturile de puncte de date eliminate $E_{in}^{(k_2)}$ sunt adăugate la aglomerare, ordinea de analiză a punctelor de date pentru adăugare fiind cea de după aplicarea sortării în conformitate cu Revendicarea 4 și seturile de puncte de date eliminate fiind analizate în ordine inversă față de ordinea în care au fost create, unde procesul este continuat până când este găsit un set de eliminare care nu poate contribui cu nici un punct de date la aglomerare.

7. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 6 în care în timpul analizei punctelor de date din seturile de puncte de date eliminate $E_{in}^{(k_2)}$, cu scopul de a decide dacă acestea sunt adăugate la aglomerare, fiind calculate numai distanțele dintre punctele de date din setul de eliminare analizat în

mod curent $E_{in}^{(k_2)}$ și un număr de seturi de eliminare analizate anterior $E_{in}^{(k_2+1)}, \ldots E_{in}^{(k_2+n_e)}$, n_e fiind un parametru de metodă suplimentar.

8. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 1 în care metoda este adaptată pentru aplicații în Optimizarea prin Identificare de Aglomerări (Eng., Clustering Optimization), adaptarea constând în asocierea la punctele de date de ponderi având valori pozitive intre 0 și 1 calculate prin scalarea inversă a valorilor funcției obiectiv între valoarea minimă a funcției și valoarea maximă a funcției pe un domeniu de căutare.

9. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 8 în care punctele de date sunt generate prin aplicarea de metode uniforme Quasi-Monte Carlo.

10. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 9 în care metoda de optimizare prin identificarea de aglomerări este aplicată în mod recursiv la regiunile de aglomerare identificate în mod curent prin generarea de puncte de eșantionare locală mai dense ca o metodă de rafinare a căutării pentru puncte de minim local.

11. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu revendicarea 10 în care metoda de rafinare este aplicată la regiunile de aglomerare identificate în mod curent prin utilizarea de metode de calcul paralele pe mașini de procesare paralelă.

12. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu oricare dintre revendicările de la 1 la 11 în care distanța pentru spațiul Euclidian multi-dimensional R^neste distanța Minkovski

$$\|\mathbf{x}-\mathbf{y}\|_p = \left(\sum_{j=1}^{n} |x_j - y_j|^p\right)^{\frac{1}{p}}, \quad \mathbf{x},\mathbf{y} \in \mathbf{R}^n, \quad p \in \mathbf{R}, \quad p > 0$$

13. Metodă de identificare de aglomerări pe bază de calculator în conformitate cu oricare dintre revendicările de la 1 la 7 extinsă la spatii altele decât spațiul Euclidian multi-dimensional unde noțiunile de distanță, volum și densitate pot fi definite în mod adecvat.

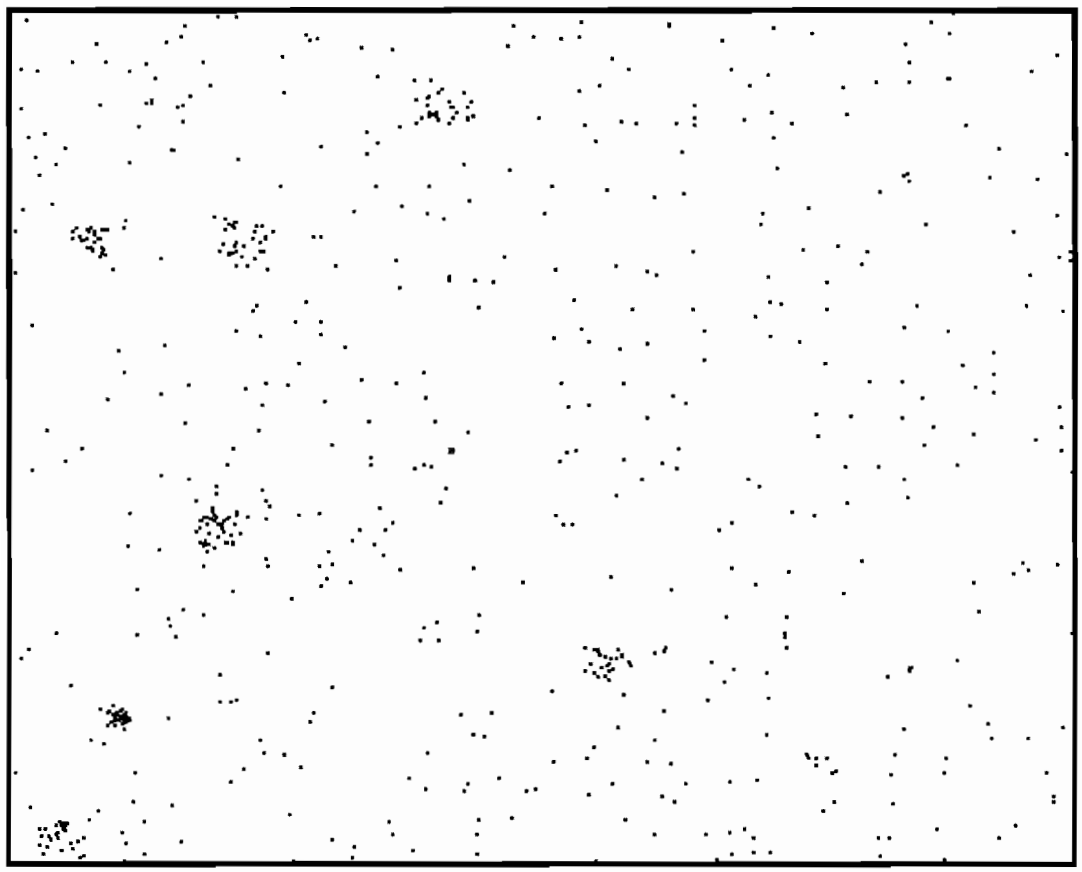
14. Sistem de calcul care implementează oricare dintre metodele din revendicările de 1 la 13, în care:

- sistemul include un procesor, o memorie, un mediu care poate fi citit de calculator și un dispozitiv de ieșire;
- sistemul poate lua forma unui, sau poate include un, calculator de scop general sau special;
- procesorul poate lua orice formă și memoria poate fi separată sau integrată cu acesta;
- dispozitivul de ieșire poate fi un afișaj, o imprimantă, un dispozitiv de înregistrare sau de stocare, sau poate lua alte forme;
- mediul care poate fi citit de calculator poate lua orice formă și de preferință are încorporate pe acesta instrucțiuni care pot fi executate de calculator, cum ar fi soft;
- instrucțiunile care pot fi executate de calculator sunt efectuate de către procesor;
- mediul poate fi integrat cu restul sistemului, sau poate fi separat de acesta.
- datele care trebuie sa fie procesate de către metode pot fi introduse de către un utilizator, furnizate ca o ieșire de la un alt dispozitiv sau sistem, sau pot fi stocate pe un mediu care poate fi citit de calculator sau de un dispozitiv de calcul;

caracterizat prin aceea că instrucțiunile încorporate pe mediul care poate fi citit de calculator implementează unul sau mai mulți algoritmi ai metodelor MMCC descrise aici.

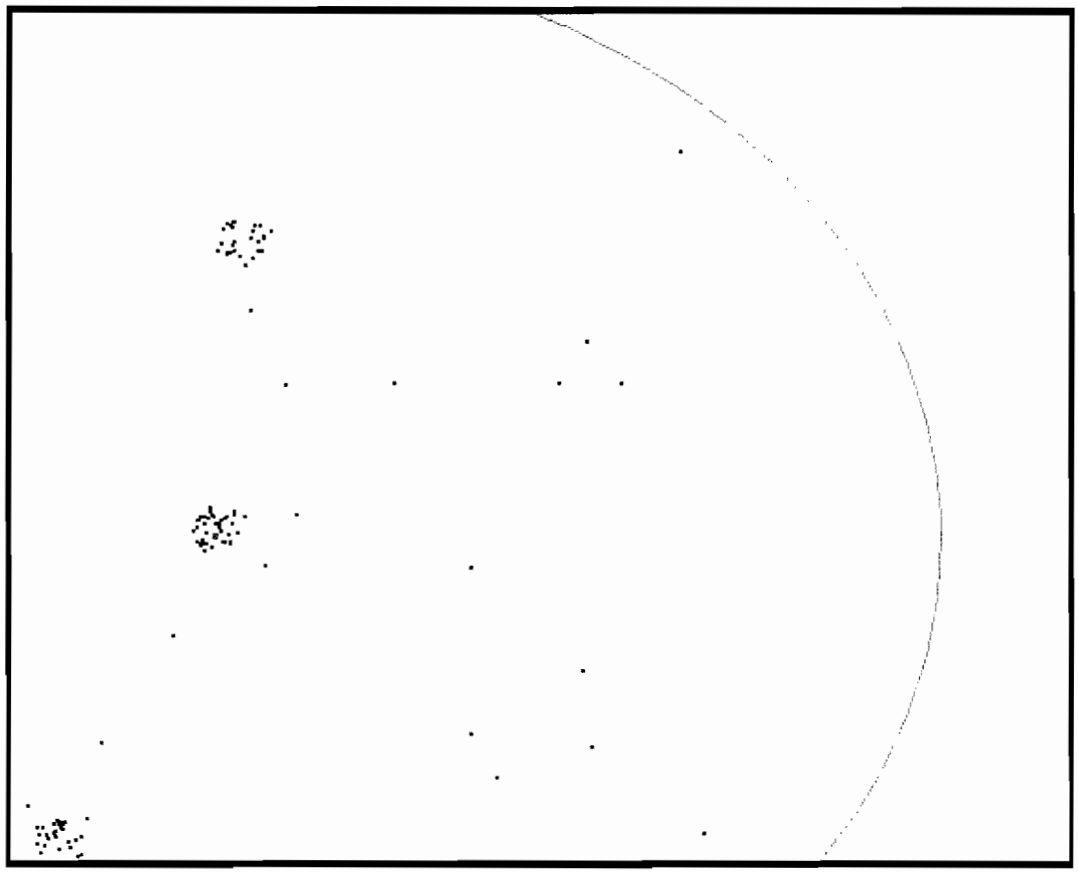
15. Sistemul de calcul în conformitate cu revendicarea 14 în care sistemul include un mediu care poate fi citit de calculator având un program de calculator care implementează oricare dintre metodele din revendicările de la 1 la 13 încorporat pe mediu.

16. Mediu care poate fi citit de calculator având instrucțiuni care pot fi citite de calculator încorporate pe acesta pentru efectuarea oricăreia dintre metodele pe bază de calculator din revendicările de la 1 la 13.



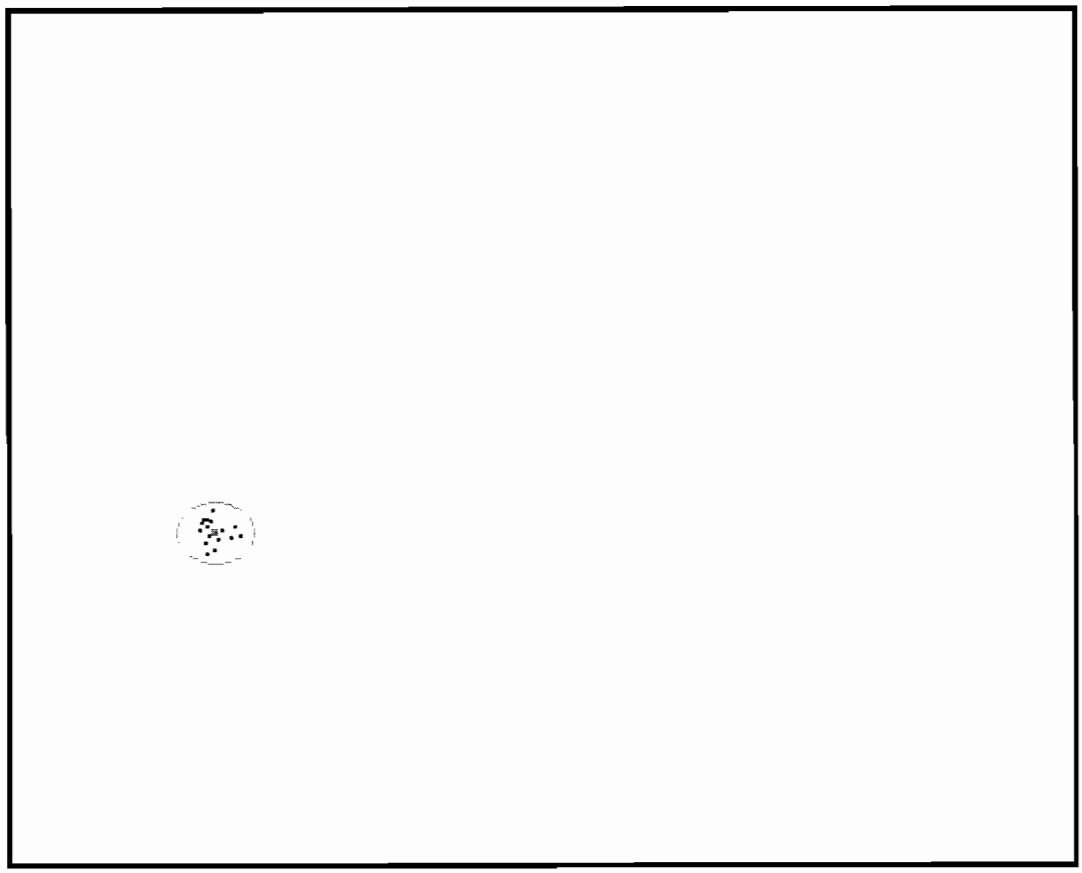
Exemplu de set inițial de puncte de date P

Fig. 1



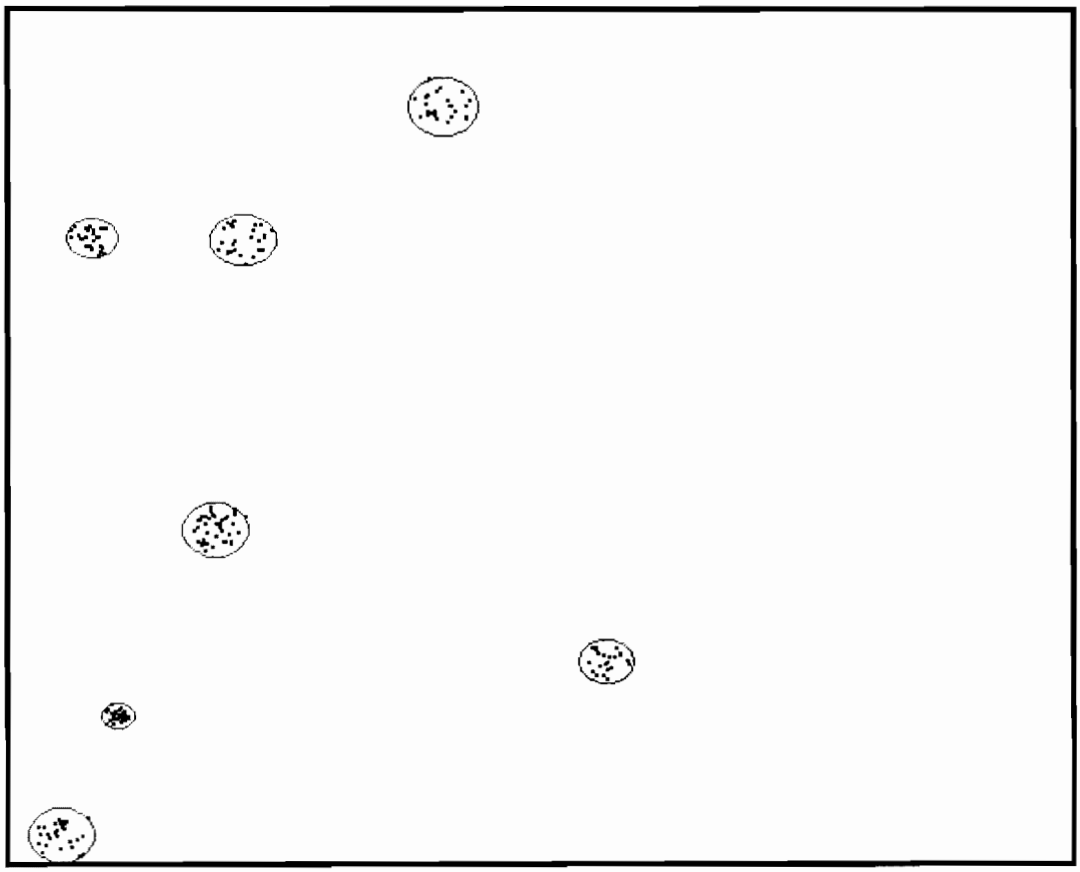
Instantaneu al setului de puncte de date după un număr de iterații ale buclei interioare

Fig. 2



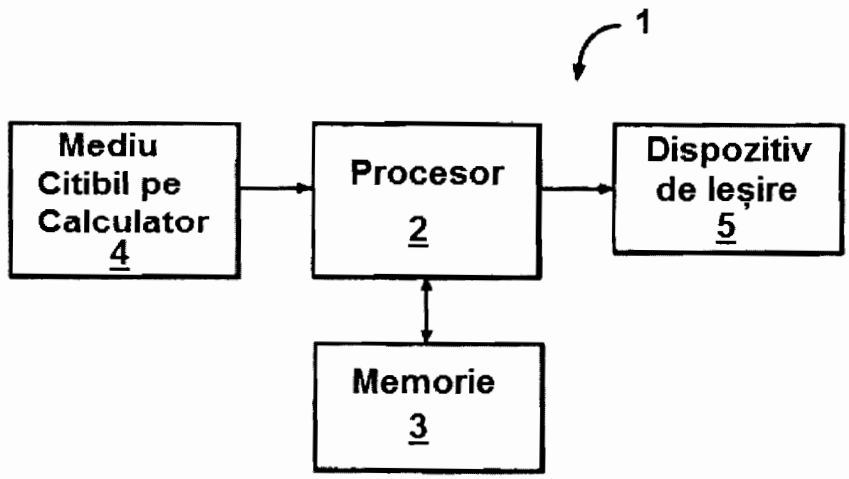
Identificarea unei noi aglomerari

Fig. 3



Setul de aglomerari identificate C

Fig. 4



Sistem de calcul general care implementează metoda MMCC

Fig. 5