Proiect cercetare Lab2

Articole:

* Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm [[1]](#Bookmark1)
* A new hierarchical clustering algorithm [[2]](#Bookmark1)
* Cluster quality based performance evaluation of hierarchical clustering method [[3]](#Bookmark1)
* Principal component analysis: a review and recent developments [[4]](#Bookmark1)
* A Study on Multiple Linear Regression Analysis [[5]](#Bookmark1)

Rezumat:

[[1]](#Bookmark1) **Relevanta:** Imbunatatirea algoritmului k-means clustering prin prisma caruia se poate efectua gruparea de date dintr-un set de date. Este prezentata o metoda ce optimizeaza algoritmul prin reducerea numarului de calcule de distanta dintre fiecare instanta prezenta in setul de date si centrul grupurilor create.

**Structura:**

Abstract: Este introdus algoritmul k-means clustering.

Introducere: Se pune accentul pe importanta algoritmului in diferite domenii si incercari precedente de a il optimiza.

Algoritmul original: Este explicat atat algoritmul clasic cat si dezavantajele acestuia.

Algoritmul imbunatatit: Se prezinta schimbarile propuse asupra algoritmului.

Rezultate experimentale: Algoritmul modificat este rulat pe trei seturi de date de diferite marimi (iris, glass si letter) si se pot vedea beneficiile aduse.

Concluzie: Este rezumat articolul in scurt si se mentioneaza reducerea complexitatii de la O(nkt) la O(nk) fara a pierde din acuratete.

[[2]](#Bookmark1) **Relevanta:** Este prezentat un algoritm experimental de clustering bazat pe conceptul si formula de distanta Euclidiana cu scopul de a obtine acuratete superioara fata de alti algoritmi de grupare.

**Structura:**

Abstract: Se prezinta scopul algoritmilor de grupare.

Introducere: Este prezentata o privire de ansamblu asupra importantei algoritmilor de grupare in special a celui ierarhic.

Grupare: Se subliniaza scopul unui algoritm de grupare precum si ce ar presupune un algoritm ideal.

Grupare ierarhica: Sunt explicate conceptele de agglomerative si divisive hierarchical clustering precum si avantajele acestui tip de grupare (irerhic).

Grupare partitionata: Este explicat algoritmul k-means.

Noul algoritm de grupare ierarhic: Se propune un nou algoritm si se explica pasii acestuia

Experimente: Algoritmul propus este rulat in paralel cu k-means pe doua seturi de date diferite.

Concluzie: Articolul este rezumat, mentionand acuratetea crescuta a noului algoritm precum si timpul mai lung de rulare pe care acesta il prezinta. Sunt discutate potentiale imbunatatiri pe viitor.

[[3]](#Bookmark1) **Relevanta:** Articolul propune analiza metodelor de grupare in functie de urmatorii factori: masura de coeziune, silhouette index si timpul de calcul. Principala metoda testata reprezinta gruparea

Irearhica (agglomerative) din cauza importantei acesteia in diferite domenii. Aceasta este testata pe diferite marimi de seturi de date.

**Structura:**

Inroducere: Se explica cei trei factori (cat de apropiate sunt intrarile intr-un grup si verificare apartenentei unei intrari intr-un grup si se mentioneaza cativa algoritmi de grupare.

Articole relatate: Sunt prezentate articole relatete si metode folosite anterior.

Metode de grupare: Este prezentata o explicatie detaliata asupra algoritmilor de grupare irerahica precum si metode de dizolvare a grupurilor formate cu scopul de a testa performanta si acuratetea acestora.

Design experimental: Este prezentata configuratia folosita de autori pentru dirijarea testelor.

Metodologie si rezultate: Sunt sumarizate rezultatele obtinute, accentul punandu-se pe relatia dintre marimea setului de date si cele trei metrice propuse anterior.

Concluzie: Se conclude faptul ca metricii de coeziune si de timp de rulare sunt metode valide de testare a algoritmilor de grupare ierarhici; in schimb silhouette index nu pare a fi potrivit pentru scopul articolului. Se propun experimente alternative.

[[4]](#Bookmark1) **Relevanta:** Articolul acopera in detaliul analiza de componenta principala ce ajuta la sumarizarea seturilor de date fara a pierde variatia datelor precum si metodele de analiza robust si simbolic si functional.

**Structura:**

Introducere: Este explicat motivul pentru care analiza de com ponenta principala este adesea folosita pentru analizarea seturilor de date.

Metoda clasica: Se prezinta bazele matematice pe care analiza este bazata si cum aceasta transforma datele. Este propus un exemplu practic asupra dintilor fosilizati.

Problemele intalnite: Sunt prezentate probleme ce afecteaza matricea de corelatie si covarianta precum analiza datelor de tipuri diferite.

Biplots: Se prezinta constructia si interpretarea grafica biplots ce este folosita pentru vizualizarea analizei efectuate.

Centrari diferite: Se prezinta cele trei centrari: dupa coloana (standard), necentrata, dupa linie/dublu-centrata.

Cazul n<p: Este prezentat cazul in care sunt mai multe variabile decat observatii precum si alte considerente practice si teoretice.

Adaptari ale analizei de componenta principala: Sunt prezentate conceptele de analiza functionala si de simplificare a componentei principale . Este propus un exemplu bazat pe presiunea subacvatica.

Analiza de componenta principala robusta: Este prezentata aceasta variatie a analizei care a fost conceputa cu scopul de a diminua efectul valorilor aberante.

~~Analiza de componenta principala simbiotica:~~

Concluzie: Se conclude ca analiza de coomponenta principala ramane de importanta inalta la analiza seturilor de date din cauza flexibilitatii si variatiei demonstrate.

[[5]](#Bookmark1) **Relevanta:**  Studiul este concentrat asupra analizei liniare de regresie asupra unui exemplu ce prezinta variabile independente si o variabila dependenta. Contextul este reprezentat de un set de date axat pe notele obtinute de 240 de studenti in cadrul universitar.

Metodologia include teste statistice cu scopul de a calcula regresia liniara si normala.

**Structura:**

Introducere: Este efectuata prezentarea setului de date ales.

Recenzie literara: Sunt prezentate studii precedente in domeniu.

Metodologie: Se prezinta utilizarea regresiei liniare cu scopul de a evalua impactul unui subset de variablile independente asupra rezultatului total.

Rezultat: Analiza de regresie este interpretata de catre autori si comparata cu rezultate precedente din alte studii.

Concluzie: Este rezumat rezultatul cheie si se fac diferite recomandari asupra cercetarilor viitoare.

Referinte/Citatiie

Formatul referintelor este consistent peste articolele selectate, acestea fiind specificate sub categoria lor independenta la sfarsitul articolului indexate cu paranteze patrate ([x]) unde x merge de la 1 la n. Pentru a trimite referinta la acestea pe parcursul articolelor sunt prezente aceleasi notatii cu x inlocuit cu numarul referintei citate la finalul acesteia.

Numarul de referinte: 15 [[1]](#Bookmark1) 11 [[2]](#Bookmark1) 16 [[3]](#Bookmark1) 51 [[4]](#Bookmark1) 5 [[5]](#Bookmark1)

Numarul de citatii: 443 [[1]](#Bookmark1) 47 [[2]](#Bookmark1) 26 [[3]](#Bookmark1) 8933 [[4]](#Bookmark1) 1533 [[5]](#Bookmark1)

Bibliografie:

[1] Shi Na, Liu Xumin and Guan Yong, “Research on k-means Clustering Algorithm: An

Improved k-means Clustering Algorithm”, Proc. ㅤ of the 3rd International Symposium

on Intelligent Information Technology and Security Informatics, 2010

[2] Zahra Nazari, Dongshik Kang, M. Reza Asharif, Yulwan Sung and, Seiji Ogawa, “A new

hierarchical clustering algorithm”, Proc. ㅤof the International Conference on Intelligent

Informatics and Biomedical Sciences (ICIIBMS), 2015

[3] Nisha and Puneet Jai Kaur, “Cluster quality based performance evaluation of hierarchical

clustering method”, Proc. ㅤof the 1st International Conference on Next Generation

Computing Technologies (NGCT), 2015

[4] Jolliffe IT, Cadima J. 2016 Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A 374: 20150202. http://dx.doi.org/10.1098/rsta.2015.0202

[5] Gülden Kaya Uyanık and Neşe Güler, “A Study on Multiple Linear Regression Analysis”,Proc. of the 4th International Conference on New Horizons in Education, 2013