Autor referat: Sebastian Stefan
242

R2. Recenzia unui articol din domeniul Învățării Nesupervizate

Stephen J. Redmond, Conor Heneghan (2007), 'A method for initialising the K-means clustering algorithm using kd-trees', Pattern Recognition Letters, Vol. 28, No. 8, pp.965-973
Accesibil online:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.454.9638&rep=rep1&type=pdf

Subiectul tratat și semnificația acestuia
Tema de cercetare pe care am să o abordez este legată de aplicarea unui algoritm de Învățare Nesupervizată, un domeniu care constă în detectarea de grupuri similare de obiecte dintr-un set de date[5]. Algoritmul pe care l-am ales este K-means, care deși a apărut în urmă cu peste 50 de ani, rămâne cea mai populară metodă în domeniul clusteringului și în prezent[5]. Articolul prezentat în acest rezumat oferă o metodă de îmbunătățire a unui punct slab al algoritmului clasic K-means, și anume selecția punctelor de pornire. Algoritmul propus este unul original și contribuie la optimizarea celei mai răspândite metode de Învățare Nesupervizată, motiv pentru care îl consider de o importanță majoră pentru domeniu.

În continuare voi face o sinteză a lucrării pe secțiuni:

Titlul
Titlul este descriptiv și sumarizează conținutul lucrării, anume intenția autorilor de a introduce o nouă metodă de inițializare a algoritmului K-means folosind structura de date kd-tree.

Rezumatul
Rezumatul sumarizează conținutul articolului și natura experimentului: propunerea unui nou algoritm de inițializare a punctelor de start în metoda K-means. Algoritmul pornește de la experimentele lui Katsavounidis[2] pe care le extinde cu informații de kd-densitate extrase folosind structura de date kd-tree[4]. Motivația este dezvoltarea unui algoritm mai performant decât exista în cercetarea din domeniu la acel moment.

Introducerea
Tema de cercetare abordată este una teoretică, domeniul de clustering al datelor. O introducere în domeniu este oferită făcând referință la sinteza lui A. K. Jain[1]. Este prezentat pe scurt algoritmul vizat, K-means, și este identificată o problemă: lipsa unei metode eficiente de inițializare a algoritmului, aceasta reprezentând nișa de cercetare a acestui articol. După o trecere în revistă a metodelor curente de inițializare este descrisă contribuția lucrării, un algoritm nou pentru această problemă.

Corpul lucrării
Tipul lucrării este teoretic, dezvoltarea unui nou algoritm. Articolul este structurat astfel: 1.Introducere (Clustering, K-means clustering, abordări asemănătoare), 2.Methods (descrie metodele folosite în dezvoltarea algoritmului: kd-trees, kd-density initialization), 3. Experimental design (prezintă câteva seturi de date reale și sintetice pe care va fi evaluat algoritmul, alături de un set de metrici), 4.Results (o comparație între algoritmul propus și alte metode din cercetare pe

seturile de date introduse la secțiunea 3), 5.Discussion and conclusions (o discuție pe rezultatele obținute).

Concluziile

Principalul rezultat obținut este un algoritm nou și eficient(atât ca performanță cât și ca timp de computație) pentru problema inițializării algoritmului K-means.

Algoritmul propus a fost comparat cu alte 3 metode de inițializare(metodele lui Forgy, Bradley și Fayyad, Katsavounidis), pe 36 de seturi de date. Acest algoritm obține rezultate mai bune din punct de vedere al măsurii Distortion[3] pentru peste 30 din seturile de date față de fiecare din celelalte metode. În celelalte cazuri obține rezultate comparabile. Din punct de vedere al timpului de computație, algoritmul se comportă similar cu 2 din cele 3 metode, și mult mai bine decât a treia, cea a lui Forgy, fiind cam de 10 ori mai rapid.

Algoritmul reprezintă în prezent una din cele mai eficiente metode de rezolvare a problemei prezentate anterior. O nouă direcție de cercetare deschisă de acest articol ar fi optimizarea algoritmului pentru lucrul cu seturi mari de date, de exemplu prin paralelizare.

Bibliografie

[1] Jain, A. K., Murty, M. N., Flynn, P. J., (1999), 'Data clustering: a review', ACM Computing Surveys Vol.31, No. 3, pp.264-323

[2] Katsavounidis, Ioannis & Kuo, C.-C. Jay & Zhang, Zhen. (1994), 'A New Initialization Technique for Generalized Lloyd Iteration', Signal Processing Letters, IEEE., Vol. 1, pp.144 - 146

[3] MacQueen, J. (1967), 'Some methods for classification and analysis of multivariate observations', Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, Calif., pp.281-297

[4] Jon Louis Bentley, (1975), 'Multidimensional binary search trees used for associative searching', Communications of the ACM, ACM New York, NY, USA, pp.509-517

[5] Jain, A. K., (2010), 'Data clustering: 50 years beyond K-means', Pattern Recognition Letters, Vol. 31, No.8, pp.651-666