Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)
SAINTEKS 2020
ISBN: 978-602-52720-7-3
Februari 2020
Hal 156-158

Analisis Algoritma Dbscan Dalam Menentukan Parameter Epsilon Pada Clustering Data Numerik

Fahmi Izhari

Sains dan Teknologi, Universitas Pembangunan Pancabudi Medan
E-mail: fahmi_izhari@dosen.pancabudi.ac.id

Abstrak−Algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* merupakan salah satu algoritma *clustering* yang berbasis numerik, pada algoritma ini digunakan data numerik sebagai pengujiannya. Algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* memiliki kelemahan yaitu sulitnya menentukan nilai *Epsilon* yang sesuai agar diperoleh hasil *clustering* yang baik. Pada algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)*, nilai *epsilon* dihitung berdasarkan banyak data dari keseluruhan data yang dijui. Pada penelitian ini dilakukan modifikasi terhadap algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* dengan melakukan penentuan nilai *epsilon* , hasil yang diperoleh pada penelitian nilai *Euclidean Distance* yang diperoleh lebih baik bila dibandingkan dengan hasil yang diperoleh dari algoritma algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* biasa.

Kata Kunci: ***DBSCAN, Clustering, Euclidean Distanc, Epsilon.***

1. PENDAHULUAN

Clustering merupakan sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum. Ada dua pendekatan utama yang digunakan dalam mengembangkan metode *clustering* yaitu *clustering* dengan pendekatan partisi dan *clustering* dengan pendekatan hirarki. *Clustering* dengan pendekatan partisi atau sering disebut dengan *partition-based clustering* adalah pengelompokkan data dengan memilah-milah data yang dianalisis ke dalam *cluster-cluster* yang ada. (Poteras, et al. 2014).

Clustering merupakan bagian dari metode pembelajaran tak terawasi *(unsuperviced learning)* karena tidak memerlukan pendefinisian *cluster* terlebih dahulu (Nisha dan Kaur, P.J. 2015). Pada *clustering* pengukuran kemiripan antar objek dilakukan dengan mengukur jarak untuk setiap pasang objek. Pengukuran ini dapat dilakukan dengan metode *Euclidean Distance*, *Manhattan Distance* dan *Minkowski Distance*.

Hampir semua algoritma pengelompokan terkenal membutuhkan parameter input yang sulit untuk ditentukan tetapi memiliki pengaruh yang signifikan terhadap hasil pengelompokan. Selain itu, untuk banyak set data nyata bahkan tidak ada pengaturan parameter global yang hasil algoritma pengelompokan menggambarkan struktur pengelompokan intrinsik secara akurat. DBSCAN adalah algoritma dasar untuk teknik clustering berbasis kepadatan. Penelitian ini memberikan survei algoritma pengelompokan berbasis kepadatan dengan algoritma ditingkatkan yang diusulkan yang secara otomatis memilih parameter input bersama dengan implementasinya dan perbandingan dengan algoritma DBSCAN yang ada. Hasil percobaan menunjukkan bahwa algoritma yang diusulkan dapat mendeteksi kelompok kepadatan bervariasi dengan bentuk dan ukuran yang berbeda dari sejumlah besar data yang mengandung noise dan outlier, hanya membutuhkan satu parameter input dan memberikan output yang lebih baik daripada algoritma DBSCAN. (Gaonkar, 2012)

DBSCAN merupakan algoritma dasar untuk teknik clustering berbasis kepadatan. Salah satu keuntungan menggunakan teknik-teknik ini adalah bahwa metode tidak memerlukan jumlah cluster untuk diberikan sebelumnya atau mereka tidak membuat asumsi tentang kepadatan atau varian dalam cluster yang mungkin ada dalam kumpulan data. Ini dapat mendeteksi kelompok yang berbagai bentuk dan ukuran dari sejumlah data besar yang mengandung noise dan outlier. (Glory, 2012)

2. METODE PENELITIAN

Clustering adalah salah satu topik penelitian yang penting dalam bidang *machine learning* dan data *mining*. *Clustering* telah berkembang menjadi teknik yang populer dalam bidang pengenaalan pola, pemrosesan citra dan data *mining* (Aranganayagi, 2007). Teknik *clustering* klasik seperti metode *k-means*, melakukan partisi data menjadi *k cluster* (MacQueen, 1967) dan sangat peka terhadap nilai awal dari masing-masing pusat *cluster* (Cuietal, 2015).

Menurut Tan (2006) *clustering* adalah mengelompokkan objek (data) yang didasarkan hanya pada informasi yang terdapat dalam objek tersebut dan hubungan antar objek tersebut. Pengelompokan data tersebut biasanya dilakukan berdasarkan kesamaan nilai antar data (Xia et al. 2008).

Prinsip dasar dari *clustering* adalah mengukur jarak atau kemiripan antar objek pada suatu basis data. *Clustering* termasuk dalam metode pembelajaran tak terawasi *(unsuperviced learning)* (Nisha, 2015). *Clustering* bertujuan agar objek-objek pada satu kelompok adalah hanya terdiri dari objek-objek yang memiliki kemiripan satu sama lain dan berbeda dengan objek pada kelompok yang lain.

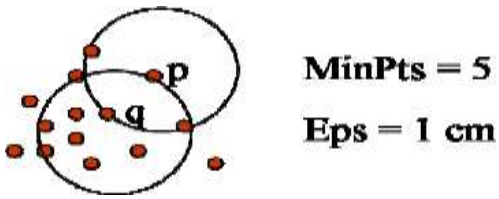
2.1 DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)*

DBSCAN adalah salah satu algoritma *clustering density-based*. Algoritma memperluas wilayah dengan kepadatan yang tinggi ke dalam *cluster* dan menempatkan *cluster* irregular pada database spasial dengan *noise*. Metode ini mendefiniskan

ISBN: 978-602-52720-7-3
Februari 2020
Hal 156-158

cluster sebagai *maximal set* dari titik-titik yang *density-connected*. DBSCAN memiliki 2 parameter yaitu *Eps* (radius maksimum dari *neighborhood*) dan *MinPts* (jumlah minimum titik dalam *Eps-neighborhood* dari suatu titik). Ide dasar dari *density-based clustering* berkaitan dengan beberapa definisi baru:

1. *Neighborhood* dengan radius *Eps* dari suatu obyek disebut *Epsneighborhood* dari suatu obyek tersebut
2. Jika *Eps-neighborhood* dari suatu obyek mengandung titik sekurangkurangnya jumlah minimum, *MinPts*, maka suatu obyek tersebut dinamakan *core object*
3. Diberikan set obyek *D*, obyek *p* dikatakan *directly density-reachable* dari obyek *q* jika p termasuk dalam *Eps-neighborhood* dari *q* dan *q* adalah *core objek*.



Gambar 1. *Eps-neighborhood* Arthur (2010)

2.2 Metode Euclidean Distance

Euclidean Distance atau jarak *Euclidean* adalah perhitungan jarak dari dua buah titik dalam *Euclidean space*. *Euclidean space* diperkenalkan oleh *Euclid*, seorang matematikawan dari Yunani sekitar tahun 300 B.C.E. untuk mempelajari hubungan antara sudut dan jarak. *Euclidean* ini berkaitan dengan *Teorema Phytagoras* dan biasanya diterapkan pada 1, 2 dan 3 dimensi.

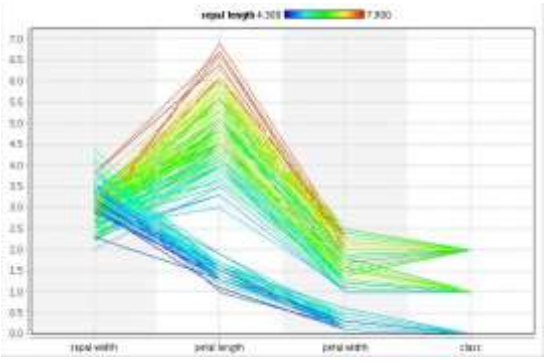
Jarak Euclidean adalah jarak yang diukur lurus dari titk koordinat yang satu ke titik koordinat yang lain. Meskipun cara ini kurang realistis, tetapi pada umumnya sering digunakan karena cara ini mudah dimengerti dan mudah dimodelkan. Aplikasi dari jarak *Euclidean* pada umumnya bisa kita jumpai pada beberapa model konveyor, sistem transportasi dan distribusi.

2.3 Identifikasi Masalah

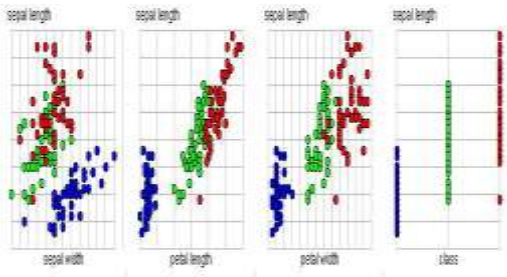
Dari latar belakang yang telah di jelaskan sebelumnya hampir semua algoritma pengelompokan membutuhkan parameter input yang sulit untuk ditentukan tetapi memiliki pengaruh yang signifikan terhadap hasil, menemukan ukuran dalam bentuk *cluster* dan efisien bahkan untuk set data besar. DBSCAN akan mendeteksi *cluster* serta menentukan parameter epsilon secara otomatis dengan cara yang akurat untuk menemukan parameter input dan menemukan cluster dengan kepadatan yang berbeda-beda.

3. ANALISA DAN PEMBAHASAN

Pada tahap ini dilakukan pengujian kinerja dari algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)*, Gambar 4.1 menunjukkan grafik hasil *clustering* yang diukur berdasarkan nilai *Euclidean Distance* yang diperoleh dari algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)*



Gambar 1. Grafik Hasil Pengujian pada *Iris Dataset*



Gambar 2. Plot Cluster Hasil Pengujian pada *Iris Dataset*

ISBN: 978-602-52720-7-3
Februari 2020
Hal 156-158

Pada tahap ini dilakukan pengujian kinerja dari algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* terhadap penyebaran cluster yang dapat dilihat Gambar 4.2 menunjukkan plot hasil *clustering* yang diukur berdasarkan nilai *Euclidean Distance* yang diperoleh dari algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)*.

4. KESIMPULAN

Penerapan nilai *Epsilon* dan pada algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* dapat dilakukan untuk memperoleh hasil *clustering* yang lebih baik. Metode penentuan nilai *Epsilon* pada algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise)* akan berpengaruh langsung terhadap jumlah *clustering* yang dihasilkan.

REFERENCES

[1] Aranganayagi, S & Thangavel, T. 2007. Clustering Categorical Data using Silhouette Coefficient as a Relocating Measure. Proceedings of 2007 Internasional Conferenceon Computational Intelligenceand Multimedia Aplication. pp.13-17
[2] Cui, X & Wang, F. 2015. An Improved Method for K-Means Clustering. Proceedings of 2015 International Conferenceon Computational Intelligence and Communication Networks (CICN). pp.756-759
[3] Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A Density-Based Algorithm for Discovering Clusters, 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996
[4] Fayyad, U., Shapiro, G.P & Smyth, P. 1996. From Data Mining to Knowledge Discovery in Database. AI Magazine: pp. 37-53
[5] Glory H. Shah, C. K. Bhensdadia, Amit P. Ganatra. 2012. An Empirical Evaluation of Density-Based Clustering Techniques. International Journal of Soft Computing and Engineering (IJSCE)
[6] Gorunescu, F. 2011. Data Mining : Concepts, Models and Techniques. Springer: Berlin.
[7] Gothai, E & Balasubramanie, P. 2010. Performance Evaluation of Hierarchical Clustering Algortihms. Proceedings of The Internasional Conference on Communication and Computational Intelligence - 2010. pp. 457-460.
[8] Han, J. & Kamber, M. 2006. Data Mining: Concepts and Techniques. 2nd Edition. Elsevier: San Francisco.
[9] MacQueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Statistics and Probability. Universityof California Press, Berkeley. California.
[10] Manisha Naik Gaonkar & Kedar Sawant. 2012. AutoEpsDBSCAN : DBSCAN with Eps Automatic for Large Dataset. Goa College of Engineering, Computer Department, Ponda-Goa, Goa College of Engineering, Computer Department, Ponda-Goa.
[11] Nisha & Kaur. P.J. 2015. Cluster Quality Based Performance Evaluation of Hierarchical Clustering Method. Proceedings of 2015 1st Internatinal Conference on Next Computing Technologies. pp. 649-653.
[12] Poteras, C.M., Mihăescu, M.C. & Mocanu, M. 2014. An optimized version of the kmeans clustering algorithm. Proceedings of the 2014 Federated Conference on Computer Science and Information Systems, pp. 695–699.
[13] Rokach, L & Maimon. O. 2005. Data Mining and Knowledge Discovery Handbook. Springer: Tel Aviv.
[14] Tan, P.N., Steinbach, M & Kumar, V. 2006, Introduction to Data Mining (Vol. 1), Pearson Addison Wesley: Boston.
[15] Xiaojuan Hu1, Lei Liu1, Ningjia Qiu2, Di Yang2 and Meng Li3. 2017. A MapReduce-based improvement algorithm for DBSCAN. Journal of Algorithms & Computational Technology