PENGELOMPOKKAN POKOK BERDASARKAN KAEDAH KETUMPATAN

Mohamed Ashraff Dzureil Hadi
Shahnorbanun Sahran

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Pengelompokkan pokok berdasarakan kaedah ketumpatan adalah satu kajian yang dibangunkan untuk menyelidik data LiDAR (Light Detection and Ranging) dataset yang disediakan oleh Pusat Angkasa Universiti Kebangsaan Malaysia. Dengan melakukan kajian terhadap data sebenar, banyak maklumat yang dapat diekstrak untuk kegunaan pada masa hadapan. Justeru itu, hasil kajian ini boleh dijadikan sebagai platform kepada penyelidik lain dalam menyelidik data sebenar yang berkaitan dengan dataset LiDAR. Metodologi yang digunakan dalam kajian ini ialah model *Waterfall*. Perisian yang digunakan untuk pembangunan kajian adalah Python. Maklumat ketinggian titik awan dalam data LiDAR dihasilkan berdasarkan lapisan pada dataset untuk menghasilkan kajian lanjut. Kajian ini memfokuskan pencarian pengelompokan dengan menggunakan kaedah pembelajaran mesin DBSCAN.

1 PENGENALAN

Projek ini berfokus kepada kajian tentang pengelompokan data yang merupakan tugas penting dalam pembelajaran mesin dan pengecaman corak. Sejak beberapa tahun yang lepas penggunaan *Airborne Scanner System* yang menghasilkan 3 dimensi persekitaran maya point cloud kepada permukaan topologi hutan banyak digunakan dengan menggunakan LiDAR (Light Detection and Ranging) bagi memudahkan pengesanan pokok didalam hutan. LiDAR merupakan teknologi kawalan pengesanan yang menggunakan cahaya dalam bentuk laser untuk mengukur pembolehubah jarak ke bumi dan ianya merupakan kawalan pengesanan aktif. Namun daripada LiDAR sahaja, ia tidak dapat melakukan pengecaman pokok dengan maklumat yang diperlukan sepenuhnya. Oleh itu, kawasan algoritma pengelompokan k-means, mean-shift dan DBSCAN akan dipilih untuk membangunkan model pembelajaran mesin yang digunakan untuk mencari hasil kajian yang akan dijalankan. Pengesktrakan data pokok secara terus melalui LiDAR akan

digunakan dengan kaedah pengelompokan untuk kemanfaatan hasil kajian. Secara umumnya, kaedah yang akan dimplementasi ini khususnya memfokuskan pada pemerhatian titik n didalam kelompok titik k di mana setiap pemerhatian dipunyai kepada kelompok yang terdekat dengan mencuba mengecilkan jumlah keseluruhan jarak ruang titik Euclidean dengan kelompok sentroid-sentroid.(S. Na, L. Xumin and G. Yong, 2010)

2 PENYATAAN MASALAH

LiDAR *(Light Detection and Ranging)* menghasilkan titik awan LiDAR untuk mengesan hasil kelompok pokok yang terhasil. Dalam kajian lepas, "*Improving Individual Tree Crown Delineation and Attributes Estimation of Tropical Forests Using Airborne LiDAR Data* (Wan Mohd Jaafar, Wan Shafrina. 2018) kaedah imej raster digunakan untuk melihat silara pokok daripada imej raster ataupun titik awan kawasan tinggi. Kaedah imej raster ini hanya mampu untuk mengesan pokok dibahagian atas permukaan sahaja dan tidak mengesan bahagian bawah pokok yang berada dihutan.

Seharusnya pokok dikesan sebagai individu pokok tetapi sekelompok pokok masih berada dalam kelompok yang sama bagi kaedah imej raster ini. Pengesanan sekelompok pokok mengakibatkan penyukaran untuk mendapatkan maklumat spesifik tentang ciri-ciri satu pokok dalam kawasan kajian.

3 OBJEKTIF KAJIAN

Tujuan kajian ini adalah untuk membangunkan model pengelompokkan *DBSCAN* untuk mengenal pasti bilangan kluster yang terhasil. Justeru, dalam pembangunan model ini harus juga objektif kajian mengenalpasti ciri-ciri kluster yang terhasil. Seterusnya, dapat menghasilkan visualisasi kluster yang terhasil daripada proses pengelompokkan kaedah yang diguna.

4 METOD KAJIAN

4.1 Fasa Keperluan

Fasa keperluan ini dimulakan dengan proses mengenalpasti dataset yang akan digunakan dalam kajian ini. Dataset yang akan digunakan dalam penyelidikan ini ialah data koordinat latitud dan longitude (x,y,z) titik awan yang diproses menggunakan LiDAR oleh Dr Wan Shafrina yang

merupakan penyelidik dari ANGKASA UKM. Dataset daripada penyelidik ini dipilih kerana dataset disediakan dan telah disahkan serta diakui dari segi ketepatan. Dataset ini juga telah dibuat setelah membuat rujukan daripada kajian "*Improving Individual Tree Crown Delineation and Attributes Estimation of Tropical Forests Using Airborne LiDAR Data*". Maka, dataset yang dipilih tidak mempunyai sumber yang meragukan dan terjamin daripada segi kesahihan.

Seterusnya, dataset yang diperolehi akan melalui proses pra-pemprosesan data di mana dataset ini akan dipersiapkan dalam bentuk yang boleh dianalisis. Antara perkara yang perlu dititikberatkan ketika proses pra-pemprosesan ini ialah normalisasi data kasar iaitu koordinat x, y, z di fail *Excel*.Hal ini kerana, koordinat dalam bentuk fail *Excel* ini tidak dapat digunakan secara terus dengan pelantar yang akan digunakan untuk pengujian. Objektif kajian ini adalah untuk menentukan hasil kluster daripada dataset yang sudah disediakan.

4.2 Fasa Reka bentuk

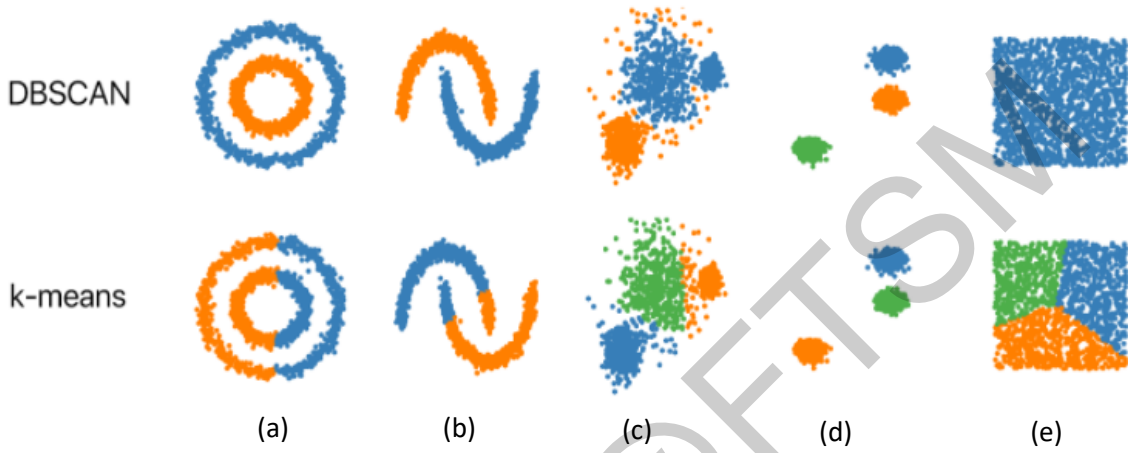
Fasa kedua pula ialah fasa reka bentuk. Fasa ini hanya akan mula dijalankan setelah fasa pertama selesai dengan sepenuhnya, seperti prinsip yang diaplikasikan dalam model air terjun. Terdapat dua jenis kaedah yang diguna pakai dalam kajian ini iaitu kaedah normalisasi dan teknik pembelajaran mesin secara berasaskan ketumpatan atau *DBSCAN*.

4.3 Fasa Pelaksanaan

Fasa ketiga ini merupakan fasa yang penting kerana kaedah yang dan model algoritma yang telah dibina pada fasa sebelumnya akan diaplikasikan pada fasa ini. Fasa pelaksanaan menggunakan dataset yang sudah dibersihkan dengan kaedah normalisasi yang dipilih berbanding hanya sebahagian dataset pada fasa sebelumnya. Hasil analisis yang lengkap akan diperolehi dalam fasa ini. Proses pelaksanaan adalah hasil analisis dan prosedur pengkodan untuk teknik pengolompokan *DBSCAN*. Hasil keseluruhan iaitu dari kaedah pengolompokan akan diperolehi.

Kajian ini menjalankan analisa pengelompokkan menggunakan algoritma *DBSCAN* untuk tujuan pengelompokkan. Ini adalah algoritma pembelajaran mesin yang *'unsupervised'*. Ia sesuai digunakan untuk kelompok berkepadatan tinggi. Ia secara automatik meramalkan penyekat dan

menghapusnya. Ia lebih baik daripada algoritma pengelompokkan hierarki dan *k-means*. Ia membuat kelompok berdasarkan parameter seperti *epsilon*, titik min dan kehingaran. Ia secara berasingan meramalkan titik berpusat, titik sempadan dan penyekat dengan cekap.



Rajah 1 Perbandingan hasil kelompok *DBSCAN* dan k-means

Sumber : *Nagesh Singh Chauhan, 2020*

Rajah 1 menunjukkan perbandingan hasil antara kaedah DBSCAN dan k-means

4.4 Fasa Pengujian

Setelah proses pelaksanaan atau pengekstrakan maklumat selesai dijalankan pada fasa ketiga, kajian akan diteruskan ke fasa terakhir iaitu fasa pengujian. Hasil daripada kaedah-kaedah yang dimplementasi akan disahkan dari segi ketepatan bagi dataset ini. Pengujian ini akan melalui pengesahan ciri-ciri hasil teknik *DBSCAN*. Pengujian adalah penting bagi mengelakkan maklumat yang diekstrak tidak dimanipulasikan dan harus menepati objektif kajian ini. Kemudian, berdasarkan pengujian yang telah diperoleh, prestasi setiap kaedah akan direkodkan. Prestasi ini direkodkan dengan analisis parameter berkait dengan hasil teknik pengelompokkan yang digunakan.

5 HASIL KAJIAN

Bahagian ini membincangkan hasil daripada proses pembangunan kajian silara pokok menggunakan pembelajaran mesin. Penerangan yang mendalam tentang Kajian ini menggunakan kaedah pembelajaran mesin berdasarkan ketumpatan untuk menentukan bilangan kelompok yang terdapat dalam dataset. Untuk menggunakan kaedah ini kelompok awal perlu ditentukan bagi menggunakan kaedah berdasarkan ketumpatan. Kaedah yang digunakan bagi menentukan kelompok awal untuk pengelompokkan merupakan kaedah *DBSCAN.* Dengan menggunakan kaedah ini, pengelompokkan awal dapat ditentukan untuk digunakan dalam model. Seterusnya, model latihan *DBSCAN* harus dibina untuk mengecam kestabilan kaedah pengolompokan dan memahami hasil yang dicipta oleh ciri-ciri *DBSCAN*.

Data yang diguna sudah dinormalisasikan menggunakan kaedah normalisasi yang dihuraikan.Seterusnya, digunapakai untuk model latihan.

```
# Generate sample data
centers = [[1, 1], [-1, -1], [1, -1]]
X, labels_true = make_blobs(n_samples=750, centers=centers, cluster_std=0.4,
                            random_state=0)

X = StandardScaler().fit_transform(X)
```

Rajah 2: Pengkodan dataset ujian untuk model latihan *DBSCAN*

```
# Compute DBSCAN
db = DBSCAN(eps=0.3, min_samples=10).fit(X)
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels = db.labels_

# Number of clusters in labels, ignoring noise if present.
n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
n_noise_ = list(labels).count(-1)

print('Estimated number of clusters: %d' % n_clusters_)
print('Estimated number of noise points: %d' % n_noise_)
print("Homogeneity: %0.3f" % metrics.homogeneity_score(labels_true, labels))
print("Completeness: %0.3f" % metrics.completeness_score(labels_true, labels))
print("V-measure: %0.3f" % metrics.v_measure_score(labels_true, labels))
print("Silhouette Coefficient: %0.3f"
      % metrics.silhouette_score(X, labels))
```

Rajah 3 pengkodan dan parameter yang berkaitan dengan pengujian *DBSCAN*

Rajah 3 menunjukan pengkodan dan ciri-ciri yang berkaitan dengan pengujian *DBSCAN* dengan penghuraian dibawah :

Homogeneity - Metrik homogenitas pelabelan kelompok yang diberi kebenaran asas. Hasil pengelompokkan memuaskan homogenitas jika semua klusternya hanya mengandungi titik data yang merupakan anggota satu kelas. Metrik ini tidak bergantung pada nilai mutlak label: permutasi nilai label kelas atau kluster tidak akan mengubah nilai skor dengan cara apa pun. Metrik ini tidak simetri: menukar label_true dengan label_pred akan mengembalikan skor_kelengkapan yang akan berbeza secara umum.(Andrew Rosenberg and Julia Hirschberg 2007)

Completeness - Metrik kelengkapan pelabelan kelompok yang diberi kebenaran asas. Hasil pengelompokkan memuaskan kelengkapan jika semua titik data yang menjadi anggota kelas tertentu adalah elemen dari kelompok yang sama. Metrik ini tidak bergantung pada nilai mutlak label: permutasi nilai label kelas atau kluster tidak akan mengubah nilai skor dengan cara apa pun. Metrik ini tidak simetrik: menukar label_true dengan label_pred akan mengembalikan skor_ homogen yang akan berbeza secara umum, (Andrew Rosenberg and Julia Hirschberg 2007).

V-measure - Ukuran-V adalah min harmonik antara homogenitas dan kelengkapan. (Andrew Rosenberg and Julia Hirschberg 2007)

Sillouette Coefficient - Dikira menggunakan jarak intra-kluster (a) dan jarak kluster terdekat (b) bagi setiap sampel. Pekali Siluet untuk sampel adalah (b - a) / maks (a, b). Untuk menjelaskan, b adalah jarak antara sampel dan kelompok terdekat yang sampelnya bukan merupakan bahagian. Perhatikan bahawa Pekali Siluet hanya ditentukan jika bilangan label adalah 2 = n_labels = n_sampel - 1. Nilai terbaik adalah 1 dan nilai terburuk adalah -1. Nilai berhampiran 0 menunjukkan kelompok yang bertindih. Nilai negatif pada umumnya menunjukkan bahawa sampel telah ditetapkan ke kluster yang salah, kerana kluster yang berbeza lebih serupa. (Andrew Rosenberg and Julia Hirschberg 2007)

Selepas model latihan diuji dan dibangunkan. Hasil kajian daripada ujian kestabilan algoritma *DBSCAN* adalah seperti di Rajah 4:

```
# Compute DBSCAN
db = DBSCAN(eps=0.3, min_samples=10).fit(X)
core_samples_mask = np.zeros_like(db.labels_, dtype=bool)
core_samples_mask[db.core_sample_indices_] = True
labels = db.labels_

# Number of clusters in labels, ignoring noise if present.
n_clusters_ = len(set(labels)) - (1 if -1 in labels else 0)
n_noise_ = list(labels).count(-1)

print('Estimated number of clusters: %d' % n_clusters_)
print('Estimated number of noise points: %d' % n_noise_)
print("Homogeneity: %0.3f" % metrics.homogeneity_score(labels_true, labels))
print("Completeness: %0.3f" % metrics.completeness_score(labels_true, labels))
print("V-measure: %0.3f" % metrics.v_measure_score(labels_true, labels))
print("Silhouette Coefficient: %0.3f"
      % metrics.silhouette_score(X, labels))
```

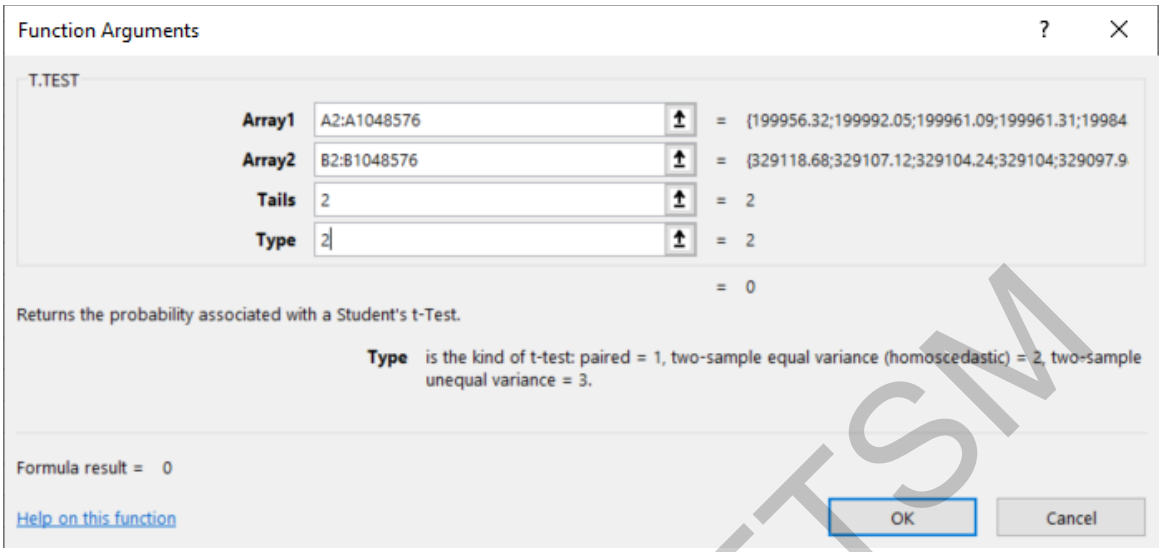
```
Estimated number of clusters: 3
Estimated number of noise points: 18
Homogeneity: 0.953
Completeness: 0.883
V-measure: 0.917
Silhouette Coefficient: 0.626
```

Rajah 4 Kaedah yang dimplementasikan dipapar dalam rajah dan hasil kluster data terbentuk

Merujuk rajah di bawah, hasil kaedah pengolompokan model *DBSCAN* dapat mengecam keseluruhan kelompok dari dataset sebenar. Jumlah kelompok yang dikesan oleh *DBSCAN* ini adalah sebanyak 8327 kelompok dan titik kelompok berstandad -1 hingga 143. Anggaran titik kelompok adalah sebanyak 143 titik.

Rajah 5 Hasil kluster berjaya terhasil dari implementasi model *DBSCAN*

Berikut adalah penilaian signifikan yang dijalan untuk menentukan hasil dapatan kajian adalah tepat. Dataset yang diuji telah dijalankan ujian *t-test* pada fail koordinat *Excel*. Nilai ujian *t-test* yang didapati daripada dataset adalah 0. Ini bermaksud Nilai t=0 menunjukkan bahawa hasil sampel sama dengan hipotesis *null*. Apabila perbezaan antara data sampel dan hipotesis *null* meningkat, nilai mutlak dari nilai-t juga meningkat. Maka nilai hasil kluster adalah jitu dan tepat.



Rajah 6 Hasil ujian t-test

Bukti hasil pengujian *t-test* yang dilakukan menggunakan kesediaan fungsi *Excel* pada dataset yang diuji.

Hasil kajian telah menghuraikan hasil analisa pengelompokkan berdasarkan ketumpatan dengan menggunakan pendekatan kaedah *DBSCAN* dan juga telah memaparkan visualisasi bersama penerangan data bagi dataset yang digunakan. Terdapat 8327 kelompok yang terhasil setelah menjalankan model *DBSCAN*.

6 KESIMPULAN

Keseluruhannya, kajian ini telah berjaya mencapai kesemua objektif kajian seperti yang dinyatakan dalam laporan. Kajian pengelompokkan pokok berdasarkan kaedah ketumpatan mempunyai maklumat dalam pelbagai aspek. Kajian ini merupakan sebuah kajian yang berasaskan pengelompokkan pembelajaran mesin berasaskan tiga pengelompokkan yang lain yang dikaji oleh beberapa pihak yang terlibat. Kelebihan pencapaian ini dapat menggalakkan lebih pengkaji menggunakan kajian ini bagi mendapatkan maklumat yang lebih spesifik dalam meneruskan kajian lanjut.

7 RUJUKAN

Zhang, Jlin, X ,2013 Filtering airborne LiDAR data by embedding smoothness-constrained segmentation in progressive TIN densification. Diakses dari: https://www.sciencedirect.com/science/article/abs/pii/S0924271613001019

S.Saeedi, F. Samadzadegan b , N. El-Sheimy, 2009, Object extraction from LiDAR data using an artificial swarm bee colony clustering algorithm. Diakses dari: http://www.pf.bgu.tum.de/isprs/cmrt09/pub/CMRT09_Saeedi_et_al.pdf.

Denise Laes, Richard Warnick, Wendy Goetz, Paul Maus USDA Forest Service Remote Sensing tips, LiDAR Applications for forestry and geosciences Diakses dari: http://static1.squarespace.com/static/59c944de59cc68469d159b28/t/5a21e5b453450aa90cada3bd/1512170934265/lidar-overview.pdf

Gupta,Sandeep. 2013. S i n gl e t r e e d e l i n e a t i o n Li D A R. Diakses dari: https://www.researchgate.net/publication/298791738_Single_Tree_Delineation_Using_Airborne_LIDAR_Data.

XingboHu, Wei Chen. 2017.Adaptive mean shift-based identification of individual tree using Airborn LiDAR data, identifying individual trees and delineating Diakses dari: https://pdfs.semanticscholar.org/1f49/2950397a7d2d135f5f31e0d088bb3184fbee.pdf

R. Gaulton . 2010.LiDAR maping of canopy gaps in continuous cover forest. Diakses dari: https://www.tandfonline.com/doi/abs/10.1080/01431160903380565

Martin Ester. 2017.Density-based algorithm for discovering clusters. Diakses dari: https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Konstantinos,G.Derpanis. 2005. Mean Shift Clustering. Diakses dari: http://www.cse.yorku.ca/~kosta/CompVis_Notes/mean_shift.pdf

Wilfred van Casteran, 2017. The waterfall model and agile. Diakses dari: https://www.researchgate.net/publication/313768860_The_Waterfall_Model_and_the_

Agile_Methodologies_A_comparison_by_project_characteristics_-_short

Wan Mohd Jaafar, Wan Shafrina. 2018. "*Improving Individual Tree Crown Delineation and Attributes Estimation of Tropical Forests Using Airborne LiDAR Data* ". Earth Observation Centre, Institute of Climate Change (IPI), Universiti Kebangsaan Malaysia. Pages. 1-23.