Jurnal Teknologi dan Riset Terapan (JATRA)
Volume 1, Nomor 1 (Juni 2019)
http://jurnal.polibatam.ac.id/index.php/JATRA

PENERAPAN TRIANGULAR KERNEL NEAREST NEIGHBOR SEBAGAI METODE CLUSTERING DASAR PADA METODE BAGGING

Muhammad Najibulloh Muzaki,

[1] Universitas Nusantara PGRI Kediri
*Corresponding author: m.n.muzaki@gmail.com

Article history

Received:
20 Mei 2019
Accepted:
24 Juni 2019
Published:
29 Juni 2019

Abstrak

Salah satu permasalahan utama dalam *data mining* adalah untuk menemukan metode *clustering* yang powerful. Terdapat suatu pendekatan baru dengan menggabungkan lebih dari satu metode *clustering* dikenal juga sebagai *multi-clustering*. Salah satu metode yang telah diperkenalkan adalah *bagging* (*bootstrap aggregating*). Metode tersebut terdiri dari metode *clustering* dasar dan metode *clustering* hierarki untuk mengkombinasikan partisi yang dihasilkan oleh metode *clustering* dasar. *Triangular kernel nearest neighbor* (TKNN) merupakan salah satu metode *clustering* berbasis densitas yang memiliki kemampuan untuk menghasilkan jumlah *cluster* secara otomatis. Penelitian ini menggunakan TKNN sebagai metode *clustering* dasar dalam metode *bagging*. Analisis komparatif menggunakan nilai F-*Measures* untuk empat metode single clustering (TKNN, ILGC, DBSCAN dan DENCLUE) dengan 9 dataset untuk menganalisa kinerja dari metode yang diusulkan. Berdasarkan hasil percobaan, menunjukkan beberapa hasil *cluster* yang lebih baik.

Kata Kunci: ***Data mining, Multi-clustering, Bagging, Triangular kernel nearest neighbor, Ward's method***

Abstract

One clustering issue is trying to find powerful method. There is a new approach that combines more than one method as known as a multi-clustering method. One method that has been introduced is bagging (bootstrap aggregating), the method consists of basic clustering methods and hierarchical clustering methods to combine partitions produced by basic clustering methods.
Triangular kernel nearest neighbor (TKNN) is a density-based clustering method that has the ability to produce a number of clusters automaticly. This research propose clustering method TKNN as basic clustering methods in the scope of bagging method.
Comparative analysis uses the value of F-Measures for four single clustering methods (TKNN, ILGC, DBSCAN and DENCLUE) with 9 datasets to analyze the performance of the proposed method. Based on the results of the experiment, it shows some better cluster results.

Keywords*: *Data mining, Multi-clustering, Bagging, Triangular kernel nearest neighbor, Ward's method

1.0 PENDAHULUAN

Salah satu isu utama dalam data mining adalah *clustering*. Fokus permasalahan adalah untuk menemukan teknik *clustering* yang powerful dalam menemukan pengetahuan dari data. Teknik *clustering* digunakan pada *data mining* untuk mengelompokkan obyek-obyek yang memiliki kemiripan dalam kelas atau segmen yang sama, sementara obyek-obyek pada kelas yang berbeda menunjukkan karakteristik yang berbeda pula[1].

Secara umum, teknik *clustering* terbagi ke dalam tiga kategori *partitioning method*, *hierarchical method* dan *density-based method*. *Partitioning method* menggunakan sebuah teknik relokasi secara beruntun, bertujuan untuk melakukan pemisahan dengan cara memindahkan obyek-obyek dari satu kelompok ke kelompok yang lain [2]. *Hierarchical method* memiliki karakteristik yang berbeda dari *partitioning method*. *Hierarchical method* tidak bertujuan untuk menemukan suatu segmentasi yang sesuai dengan jumlah *cluster* tertentu, tetapi lebih bertujuan untuk menciptakan solusi dengan jumlah *cluster* sebanyak $K = 1,\ldots,N$ [3]. *Density-based method* menyajikan algoritma pembentukan *cluster* yang berbeda. Algoritma tersebut memiliki kemampuan untuk menentukan jumlah *cluster* secara

otomatis ketika mencapai suatu keadaan yang konvergen, dimana setiap titik telah menjadi anggota pada titik pusat tertentu[4].

Setiap metode *clustering* secara umum diimplementasikan sebagai *single clustering*. Muncul gagasan baru untuk memasangkan lebih dari satu metode *clustering* atau yang lebih dikenal dengan *multi-clustering*. Secara spesifik, hasil-hasil dari beberapa algoritma *clustering* yang dilakukan secara independen, dikombinasikan untuk menghasilkan partisi dari data yang tidak dipengaruhi oleh inisialisasi dan mengatasi ketidakstabilan dari metode *clustering* [5]. Salah satu penelitian yang telah dilakukan adalah *bagging* (*bootstrap aggregating*) untuk *clustering*. Sebuah terobosan teknik pelatihan dengan memanfaatkan sampel *bootstrap* [3]. Penerapan dari metode tersebut adalah dengan menempatkan metode *clustering* berbasis partisi (*K-means*) sebagai metode *clustering* dasar untuk membentuk *cluster* dari setiap sampel *bootstrap*. Seluruh *cluster* dari masing-masing *bootstrap* yang telah terbentuk, selanjutnya akan dikumpulkan menjadi dataset baru sebagai input untuk metode *clustering* akhir menggunakan metode berbasis hirarki (*ward's linkage* untuk menyatukan *cluster* sehingga diperoleh *cluster* yang optimal.

Penelitian ini mengusukan metode *clustering* berbasis densitas untuk tahap *clustering* dasar. Tujuan dari penelitian ini adalah memperbaiki tahap *clustering* dasar untuk setiap sampel *bootstrap* sehingga menghasilkan kualitas *cluster* yang lebih baik untuk proses setelahnya. Algoritma *clustering* berbasis densitas dapat menemukan bentuk cluster yang *non-spherical* dan berguna untuk mengidentifikasi adanya *noise* [6]. *Triangular kernel nearest neighbor* (TKNN) merupakan salah satu metode berbasis densitas, dan metode tersebut yang akan digunakan sebagai metode *clustering* dasar dalam penelitian ini. *Triangular kernel nearest neighbor* (TKNN) merupakan kombinasi *K-nearest neighbor* dengan estimasi densitas menggunakan *triangular kernel* [7]. algoritma tersebut mampu membentuk *cluster* secara otomatis dan mampu mengenali *cluster* dengan *density*, *shape* dan ukuran yang berbeda [8].

Metode *clustering* akhir menggunakan metode berbasis hirarki yaitu *ward's linkage*. Metode *Ward* memiliki perbedaaan dengan metode *agglomerative hierarchical* yang lain. Tujuannya adalah untuk mengurangi sejumlah kelompok dari n ke n-1 menurut suatu cara pembentukan kelompok yang meminimalkan nilai error [9]. Sehingga variasi yang terjadi antar masing-masing obyek dalam satu *cluster* berada dalam kondisi minimal.

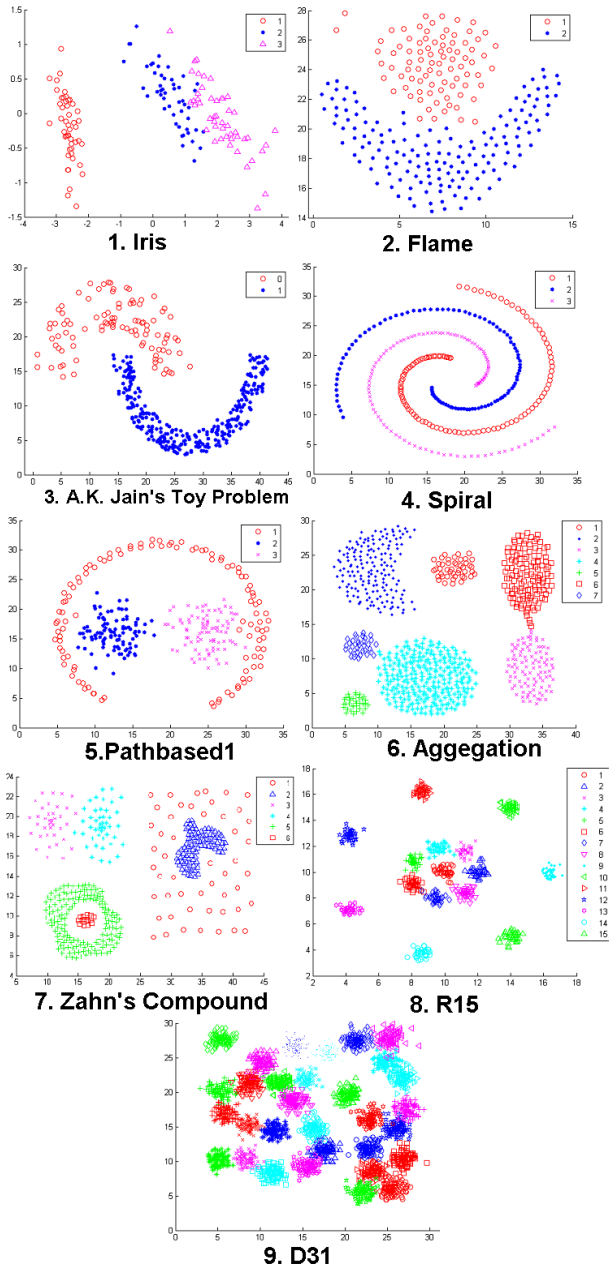
2.0 METODE

2.1. *Dataset*

Penelitian ini menggunakan 9 *dataset* untuk menguji kualitas hasil *cluster* dari metode yang diusulkan yaitu *bagging* (TKNN-AHC *Ward's linkage*) terhadap metode *single clustering*. Informasi mengenai jumlah data, jumlah atribut dan jumlah kelas ditunjukkan pada Tabel 1. Visualisasi sebagai gambaran sebaran dan bentuk data ditunjukkan pada Gambar 1.

Tabel 1: *Dataset* yang digunakan

No.	Dataset	Jumlah Data	Jumlah Atribut	Jumlah Kelas
1.	*Iris*	150	4	3
2.	*Flame*	240	2	2
3.	*A.K. Jain's Toy Problem*	373	2	2
4.	*Spiral*	312	2	3
5.	*Pathbased1*	300	2	3
6.	*Aggregation*	788	2	7
7.	*Zahn's Compound*	399	2	6
8.	R15	600	2	15
9.	D31	3100	2	31



Gambar 1: Visualisasi sebaran data

Sebanyak 8 *dataset* merupakan *artificial dataset* yang memiliki karakteristik *boundary* dan *shape type*

yang berbeda-beda, seperti yang ditampilkan pada gambar 1. *artificial dataset* tersebut adalah *dataset flame*, A.K. Jain's *toy problem*, *spiral*, *pathbased1*, *aggregation*, *Zahn's compound*, R15, dan D31 yang diperoleh dari *Speech and Image Processing Unit*, *School of Computing*, *University of Eastern Finland*. *Dataset* lain yang juga digunakan dalam penelitian ini adalah *dataset iris* dari UCI *machine Learning Repository* [10]

2.2. Usulan Metode

Bagging (*bootstrap aggregating*) merupakan metode yang diperkenalkan dalam [11]. *Bagging* secara umum digunakan untuk meningkatkan performa sebuah algoritma klasifikasi dengan terlebih dahulu menggunakan *bootstrap sampling* pada *dataset* yang telah diberikan untuk melatih sejumlah metode klasifikasi dan kemudian menggunakan mekanisme *majority voting* untuk agregat outputnya [12]. Penerapan *bagging* untuk *clustering* diusulkan dalam [3]. Ide utamanya adalah untuk menstabilkan metode partisi seperti *K-means* atau *competitive learning* dengan secara berulang menjalankan algoritma *cluster* dan mengkombinasikan hasil-hasil yang diperoleh [3].

Penelitian ini menerapkan metode *bagging* dengan *triangular kernel nearest neighbor* (TKNN) digunakan sebagai metode *clustering* dasar dan dipadukan dengan AHC *ward's lingkage* sebagai metode *clustering* akhir. TKNN sendiri termasuk dalam *density-based method*. Algoritma tersebut memiliki kemampuan untuk menentukan jumlah *cluster* secara otomatis ketika mencapai suatu keadaan yang konvergen, dimana setiap titik telah menjadi anggota pada titik pusat tertentu [6]. Secara tipikal metode *density-based method* hanya mempertimbangkan *cluster* yang eksklusif, dan tidak mempertimbangkan *cluster* yang *fuzzy* [2]. Algoritma TKNN akan mengalokasikan satu titik hanya akan memiliki satu *cluster* [13]. Alur dari algoritma *multi-clustering bagging* (TKNN-*Ward's lingkage*) dapat dilihat pada Gambar 2.

Tahapan dari metode *bagging* (TKNN-*Ward's lingkage*) dapat diuraikan kedalam prosedur sebagai berikut :

1. Inisialisasi :
   Dataset yang akan diproses, Inisialisasi parameter, yaitu jumlah *bootstrap* (*B*), kapasitas *bootstrap* (*N*), ketetanggaan (*k*), dan maksimum iterasi.
2. Bentuk sampel *bootstrap* sebanyak *B* dengan kapasitas masing-masing sebanyak *N*.
3. Lakukan tahap proses *clustering* dasar menggunakan metode TKNN untuk setiap *dataset bootstrap*. Tahapan dari metode TKNN adlah sebagai berikut :
   a. Perhitungaan jarak antar titik menggunakan *euclidean distance* menggunakan persamaan (1).

$$dist(x_i, x_j) = \sqrt{\sum_{d=1}^{N} (x_{id} - x_{jd})^2} \qquad (1)$$

   Dimana x_{id} adalah data ke-*i* dan atribut ke-*d*

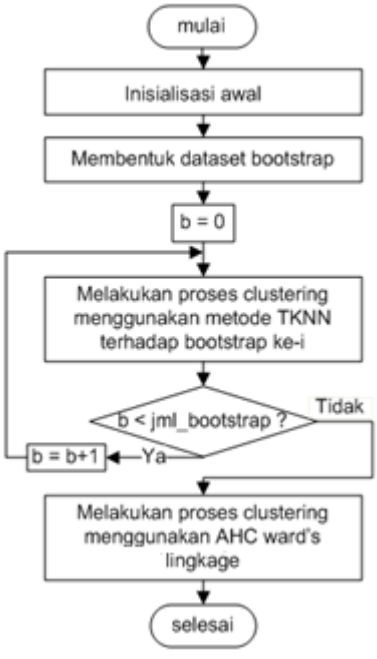
   b. Urutkan hasil perhitungan jarak secara *ascending*.
   c. Cari *k-nearest neighbor* dari setiap titik dengan merujuk pada tabel jarak, kemudian membuat tabel *k-nearest neighbor* yang berukuran (*N* x *k*).
   d. Ulangi :
      - Menghitung fungsi *triangular kernel p(x)* untuk masing-masing titik dengan memanfaatkan nilai yang tersimpan pada tabel jarak dan daftar ketetanggaan yang terdapat pada tabel *k-nearest neighbor* dengan menggunakan persamaan (2).

$$p(x) = \sum_{c=1}^{N_\omega} (1 - \frac{dist(x, x_k)}{Hx}) \qquad (2)$$

        dimana x_k adalah tetangga dari titik *x* dan merupakan anggota dari *cluster* ω, $N\omega$ adalah banyaknya anggota dalam *cluster* ω, $dist(x, x_k)$ adalah jarak antara titik *x* terhadap tetangga ke-*k* menggunakan perhitungan jarak *euclidean distance* pada persamaan (1). *Hx* adalah skala dari titik *x*.
      - Titik *x* ditempatkan pada *cluster* ω yang memiliki nilai fungsi *triangular kernel* tertinggi *max(p(x))* dengan menggunakan persamaan (3).

$$Clust_\omega = \max(p(x)) \qquad (3)$$

      - Selanjutnya adalah melakukan pembaharuan indeks label cluster untuk titik x.
   e. Hingga : Tidak ada titik yang berpindah *cluster*.
   f. Pembentukan data baru sebanyak *cluster* yang terbentuk dengan cara mengambil *centroid* dari setiap *cluster*.
4. Melakukan proses *clustering* terhadap data baru menggunakan metode AHC *Ward's linkage*.

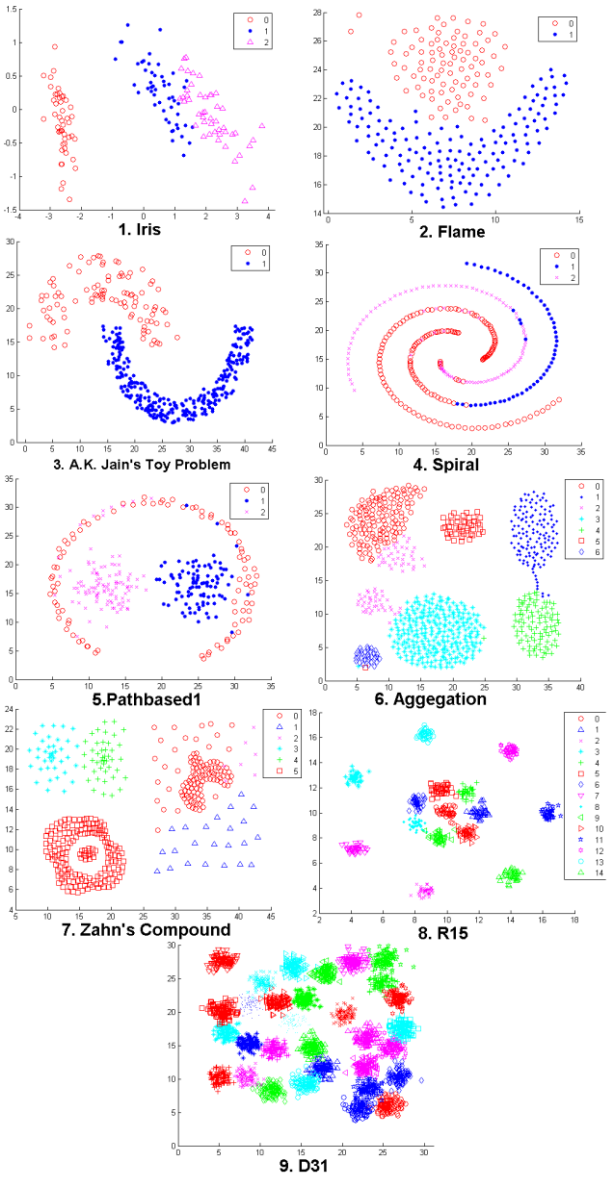


Gambar 2: Algoritma *bagging* (TKNN-*Ward's lingkage*)

Pada penelitian ini dilakukan percobaan dengan menggunakan variasi untuk parameter jumlah *bootstrap* (*B*), kapasitas *bootstrap* (*N*) dan ketetanggaan (*k*).

3.0 HASIL DAN PEMBAHASAN

Alat ukur yang akan digunakan dalam uji perbandingan ini adalah nilai perhitungan *F-measures* yang merupakan perhitungan evaluasi menggunakan nilai *precision* dan *recall*. Kualitas hasil *cluster* terbaik dari *bagging* (TKNN-AHC W*ard's linkage*) ditunjukkan pada Gambar 3. akan dilakukan uji perbandingan dengan hasil dari penerapan metode *single clustering*. Uji perbandingan terhadap metode *single clustering* menggunakan hasil penelitian [8]. Metode-metode *single clustering* tersebut merupakan *density-based method* meliputi TKNN, ILGC, DBSCAN dan DENCLUE.



Gambar 3: Visualisasi cluster terbaik hasil *bagging* (TKNN-AHC *ward's linkage*)

Hasil uji perbandingan ditunjukkan pada tabel 2. Dataset iris menunjukkan bahwa *multi-clustering bagging* (TKNN-AHC *Ward's linkage*) memberikan kualitas hasil cluster yang mengungguli metode single clustering secara keseluruhan dengan nilai F-measures sebesar 99,33%. Hasil uji perbandingan pada dataset flame, *bagging* (TKNN-AHC *Ward's linkage*) juga mengungguli metode single clustering dengan hasil cluster yang sesuai dengan kelas aslinya dengan nilai F-measures sebesar 100%. Performa *bagging* (TKNN-AHC *Ward's linkage*) memiliki hasil yang sama dengan TKNN dan ILGC dalam uji perbandingan untuk dataset A.K. Jain's Toy Problem, ditunjukkan dengan nilai F-measures mencapai 100%.

Pengujian terhadap *dataset spiral* memiliki nilai *F-measures* yang paling rendah sebesar 79,12%. Hasil tertinggi uji perbandingan pada *dataset pathbased1* dicapai oleh *bagging* (TKNN-AHC *Ward's linkage*) dengan nilai *F-measures* sebesar 93.15%. Hasil uji perbandingan tertinggi juga diperoleh *bagging* (TKNN-AHC w*ard's linkage*) untuk *dataset aggregation*, hal ini ditunjukkan dengan nilai *F-measures* sebesar 92,62%. Uji perbandingan untuk *dataset zahn's Compound bagging* (TKNN-AHC w*ard's linkage*) berada satu tingkat lebih rendah dari DBSCAN dengan nilai *F-measures* sebesar 90,46%. Uji perbandingan menggunakan *dataset R15*, *bagging* (TKNN-AHC *Ward's linkage*) memberikan nilai *F-measures* tertinggi sebesar 99,83%. Hasil pengujian untuk *dataset* D31 yang berisi data terbanyak yang digunakan dalam pengujian, *bagging* (TKNN-AHC *Ward's linkage*) sekali lagi menunjukkan kualitas tertinggi dengan nilai *F-measures* mencapai 99,9%.

Tabel 2: Perbandingan persentase *F-measures*

Dataset	Bagging (TKNN-Ward's)	TKNN	ILGC	DBSCAN	DENCLUE
1.	**99,33**	90,02	86,03	80	89,92
2.	**100**	99,17	64,58	64,58	80,83
3.	**100**	**100**	**100**	73,99	81,77
4.	79,12	**100**	**100**	**100**	34,29
5.	**93,15**	87	73,33	80,67	53,67
6.	**92,62**	78,43	78,43	82,23	58,5
7.	90,46	87,22	87,22	**96,99**	66,42
8.	**99,83**	99,67	99,67	53,33	99,5
9.	**99,9**	97,55	97,52	62,54	76,03

4.0 KESIMPULAN

Multi-clustering bagging dengan kombinasi TKNN dan *Ward's linkage* dapat memberikan kualitas hasil *cluster* yang lebih baik dari metode *single clustering* yang dihasilkan, terutama untuk *dataset iris*, *flame*, *pathbased1*, *aggregation*, *R15* dan *D31*. Kualitas yang sama dengan metode *single clustering* TKNN dan ILGC pada *dataset A.K. Jain's toy problem* ditunjukkan dengan persentase *F-measures* sebesar 100%. Hasil *cluster bagging* (TKNN-AHC *Ward's linkage*) memiliki kualitas lebih rendah dari metode TKNN, ILGC dan DBSCAN ketika dijalankan sebagai *single clustering* untuk *dataset spiral* dan kualitas lebih rendah dibawah

DBSCAN terdapat pada pengujian *dataset Zahn's Compound*.

DAFTAR PUSTAKA

[1] J. A.K., M. M.N., and F. P.J., "Data Clustering : A Review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[3] F. Leisch, "Bagged clustering," *Adapt. Inf. Syst. Model. Econ. Manag. Sci.*, vol. 51, no. 51, p. 11, 1999.

[4] T. N. Tran, R. Wehrens, and L. M. C. Buydens, "KNN-kernel density-based clustering for high-dimensional multivariate data," *Comput. Stat. Data Anal.*, vol. 51, no. 2, pp. 513–525, 2006.

[5] D. Frossyniotis, M. Pertselakis, and A. Stafylopatis, "A Multi-clustering Fusion Algorithm," 2007, pp. 225–236.

[6] A. Amini, H. Saboohi, and T. Y. Wah, "A multi density-based clustering algorithm for data stream with noise," in *Proceedings - IEEE 13th International Conference on Data Mining Workshops, ICDMW 2013*, 2013, pp. 1105–1112.

[7] A. Musdholifah and S. Z. Siti, "Triangular kernel nearest neighbor based clustering for pattern extraction in spatio-temporal database," in *Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, ISDA'10*, 2010, pp. 67–73.

[8] A. Musdholifah, S. Zaiton, and M. Hashim, "Cluster Analysis on High-Dimensional Data : A Comparison of Density-based Clustering Algorithms," vol. 7, no. 2, pp. 380–389, 2013.

[9] J. H. Ward, "Hierarchical Grouping to Optimize an Objective Function," *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963.

[10] "Iris Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/iris.

[11] B. Leo, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[12] K. Hsu, "Weight-Adjusted Baggin gof classification Algorithms Sensitive to missing Values," *Int. J. Inf. Educ. Technol. (IJIET)*, vol. 3, no. 5, pp. 560–566, 2013.

[13] A. Musdholifah, S. Z. M. Hashim, and R. Ngah, "Robust Local Triangular Kernel density-based clustering for high-dimensional data," in *2013 5th International Conference on Computer Science and Information Technology, CSIT 2013 - Proceedings*, 2013, pp. 24–32.