PROCEEDINGS OF THE 1st STEEEM 2019
Volume 1, Number 1, 2019, pp. 139-147.
Universitas Ahmad Dahlan, Yogyakarta-Indonesia, December 30, 2019

A Comparison of Outlier Detection Techniques in Data Mining

Endang Wahyuni[1], Suparman[2]

[1,2]*Magister Pendidikan Matematika, Universitas Ahmad Dahlan, Indonesia*

Abstract. Big data (data yang besar) merupakan data yang memiliki volume yang besar, jenis data yang bervariasi, serta kecepatan data yang sangat cepat.Data mining merupakan merupakan teknik analisis data statistik yang bertujuan untuk mencari informasi yang sebelumnya tidak dapat ditentukan atau menemukan informasi dari pola yang tersembunyi. Outlier walaupun muncul dengan nilai yang ekstrem seringkali mengandung informasi yang sangat penting sehingga perlu untuk dikaji dahulu apakah data tersebut tetap digunakan atau dikeluarkan. Deteksi Outlier merupakan topik yang sedang hangat untuk diiteliti. Dengan teknologi yang baru muncul dan berbagai macam aplikasi meningkatkan minat pendeteksian outlier meningkat dengan pesat. Banyak metode outlier yang berhasil diterapkan dalam berbagai bidang, mulai dari pendidikan, ekonomi, bisnis, kesehatan, antariksa, geologi, hingga penipuan kartu kredit. Metode ini bukan merupakan metode yang mudah mengingat deteksi outlier mengidentifikasi perilaku yang langka, unik serta ia dapat mengungkapkan yang jarang akan tetapi penting serta menemukan pola-pola yang menarik atau tidak terduga dari data yang rumit. Pada makalah ini membahas dan memberikan tinjauan singkat mengenai metode deteksi outlier dengan membuat perbandingan eksperimen deteksi outlier dengan metode yang populer yaitu metode KMeans dan K-Nearest Neighbors.

Kata kunci: Big data, data mining, outlier detection, KMeans , K-Nearest Neighbors

1. Pendahuluan

Big data merupakan data yang memiliki volume yang besar, jenis data yang bervariasi, serta kecepatan data yang sangat cepat[1]. Data mining merupakan proses eksplorasi dan analisis yang berjumlah besar dan varietas data yang berbeda untuk menemukan pola dan aturan yang bermakna [2]. Data mining dipelajari dalam bidang penelitian, dimana tugasnya untuk penemuan pengetahuan[3] Salah satu metode dalam data mining untuk menggali data yang bersifat unik adalah outlier detection, metode ini merupakan metode yang paling penting dalam menganalisis suatu data, seperti pengambilan keputusan, pengelompokan, dan klasifikasi pola, karena bahwa ia dapat mengungkapkan yang jarang akan tetapi penting serta menemukan pola-pola yang menarik atau tidak terduga. [1,3-5]. Outlier didefinisikan sebagai pengamatan yang tidak sesuai dengan keseluruhan pola pengelompokan[1,4,6]. Algoritma pendeteksian outlier mencari outlier dengan menerapkan salah satu algoritma pengelompokan dan mengambil set noise, sehingga kinerja algoritma pendeteksian outlier tergantung pada seberapa baik algoritma pengelompokan menangkap konstruksi cluster [7]. Analisis outlier detection ini sangat menarik untuk dibahas karena akan membahas tentang identifikasi pola-pola dari data yang tidak sesuai dengan apa yang diharapkan dan bersifat unik. Dalam outlier detection ini terdapat dua jenis yaitu univariat dan multivariat. Outlier pada univariat dapat ditemukan pada saat melihat distribusi nilai dalam

ruang dimensi tunggal, sedangkan outlier pada multivariat dapat ditemukan dalam ruang n-dimensi. masalah utama dalam penelitian data mining adalah meningkatnya dimensi data memunculkan sejumlah data baru tantangan komputasi. Dalam outlier detection diklasifikasikan menjadi deteksi outlier berbasis statistik (statistical based), berbasis jarak (distance based), berbasis kepadatan (*density based*), berbasis penyimpangan (*deviation based*), berbasis klaster (*clustering based*), dan berbasis subruang (*subspace based*) [1-2].

Deteksi outlier untuk penambangan data biasanya didasarkan pada jarak, pengelompokan dan metode spasial. Penelitian ini berkaitan dengan penempatan outlier dalam set data besar dan multidimensi dengan metode KMeans, k-medoid, dan metode pengelompokan Fuzzy C-Means. Algoritma KMeans clustering mempartisi sebuah dataset ke dalam jumlah cluster dan kemudian efeknya digunakan untuk mengunci outlier dari masing-masing cluster, menggunakan salah satu dari metode deteksi outlier.[7]

Dalam beberapa tahun terakhir, diamati bahwa aktivitas penelitian data mining terutama outlier detection semakin banyak karena deteksi outlier menjadi salah satu masalah yang penting dalam data mining serta memiliki banyak aplikasi di dunia nyata. Misalnya yang dilakukan oleh K.Swapna dan M.S. Prasad Babu yang merancang sistem diagnosis hati secara otomatis untuk mendeteksi secara dini dan akurat untuk mengurangi kematian yang disebabkan oleh penyakit hati serta menganalisis kumpulan data untuk memahami sistem guna membantu mengembangkan sistem diagnosis hati otomatis menggunaakaan metode Cluster-based Bisecting KMeans dan traditional KMeans yang menghasilkan kesimpulan bahwa algoritma Bisecting KMeans Algorithm (IBKM) lebih unggul algoritma KMeans karena menghasilkan cluster yang lebih baik dan memberikan kluster yang efisien tanpa outlier[8]. Melakukan teteksi panggilan penipuan dengan dial-back menggunakan metode pendeteksian outlier dengan klastering[9].

Maka untuk mengatasi permasalahan tersebut, penulis ingin membandingkan beberapa algoritma pendeteksian outlier yaitu metode K-Nearest Neighbors dan KMeans untuk mengetahui algoritma mana yang lebih akurat dalam menyelesaikan permasalahan dari outlier. Penelitian ini disusun sebagai berikut. Pada bagian 2, menjelaskan metode-metode yang digunaklan. Pada bagian 3, menjelaskan pembahasan. Pada bagian 4, Kesimpulan dan ucapan terimakasih.

2. Metode

2.1. Studi Pustaka

Studi pustaka dilakukan diawali dengan mengumpulkan literatur berupa jurnal-jurnal internasional dan buku terbaru yang sesuai dengan tema. Penelitian ini membahas tentang teori big data, outlier detection, KMeans yang dibahas berdasarkan jurnal-jurnal terkait.

2.2. Pengkajian literatur

Penulis membaca, mencatat, dan memahami pokok-pokok dari jurnal dan buku yang sudah dikumpulkan.

2.3. Penyusunan hasil kajian

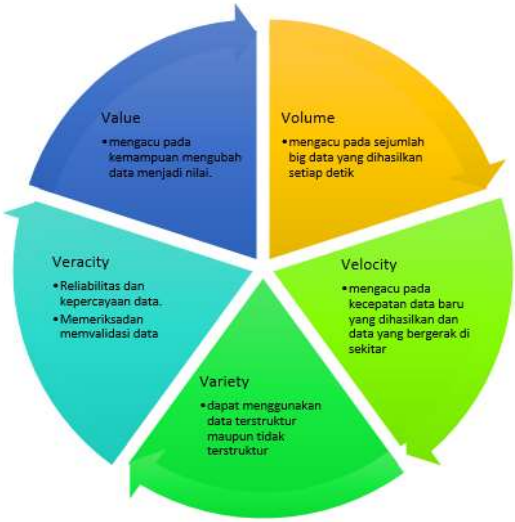
Hasil kajian disusun sebagai berikut

- Pendahuluan
  Meliputi latarbelakang masalah pada big data, outlier detection, KMeans dan K-Nearest Neighbors.
- Metode
  Meliputi langkah-langkah kajian pustaka
- Hasil dan pembahasan
  Menguraikan gagasan utama Big Data, Data Mining, Teori outlier, KMeans, K-Nearest Neighbors, deskripsi, hasil, kelebihan dan kekurangan
- Kesimpulan
  Berisi kesimpulan dan hasil kajian

3. Hasil dan Pembahasan

3.1. Big data

Big data merupakan data yang memiliki volume yang besar, jenis data yang bervariasi, serta kecepatan data yang sangat cepat[1]. Definisi atau kriteria pada big data yang sekarang-mainstream sebagai 5V yang bisa dilihat pada gambar 1.



Gambar 1. Kriteria Big data (5V)

3.2. Data Mining

Data mining merupakan proses eksplorasi dan analisis yang berjumlah besar dan varietas data yang berbeda untuk menemukan pola dan aturan yang bermakna [2]. Data mining berhubungan dengan deteksi informasi nontrivial, tidak terlihat, dan menarik dari beberapa jenis data, karena pertumbuhan teknologi informasi yang terus menerus, ada peningkatan yang besar dalam jumlah database, di samping dimensi dan kesulitannya[10]. Data mining dipelajari dalam bidang penelitian, dimana tugasnya untuk penemuan pengetahuan[3] Salah satu metode dalam data mining untuk menggali data yang bersifat unik adalah outlier detection, metode ini merupakan metode yang paling penting dalam menganalisis suatu data, seperti pengambilan keputusan, pengelompokan, dan klasifikasi pola, karena bahwa ia dapat mengungkapkan yang jarang akan tetapi penting serta menemukan pola-pola yang menarik atau tidak terduga. [1,3-5].

3.3. Outlier Detection

Deteksi outlier adalah subjek penting dalam penambangan data, khususnya telah digunakan secara luas untuk mengidentifikasi dan menghilangkan objek yang tidak relevan atau tidak relevan dari kumpulan data [1-5]. Deteksi pencilan sangat membantu dalam banyak aplikasi seperti deteksi intrusi jaringan, penipuan kartu kredit, pemantauan aktivitas, aplikasi keuangan, analisis penyimpangan pemilihan, prediksi cuaca buruk, dan sistem informasi geografis dll [2]. Namun, tantangan utama dari pendeteksian outlier adalah meningkatnya kompleksitas karena beragamnya dataset dan ukuran dataset [2] Faktor outlier dari cluster menentukan derajat perbedaan dari sebuah cluster dari seluruh dataset [10]. deteksi outlier digunakan untuk menemukan data penipuan. Dalam penelitian ini bertujuan untuk melakukan pengelompokan data dan proses deteksi outlier [4,10]

Outlier (data pencilan) adalah data obserasi yang muncul dengan nilai-nilai ekstrem baik secara univariat maupun multivariate yang dimaksud ekstrem adalah nilai yang jauh atau beda samasekali dengan sebagian besar nilai lain dalam kelompoknya[1].

Alasan kemunculan outlier:

1. Kesalahan prosedur dalam memasukkan data/mengkoding/mrndefinisan

2. Muncul dalam range nilai yang ada, tetapi bila dikimbinasi dengan variabel lain menjadi ekstrem
3. Pengambilan sample

Alasan dilakukannya pendeteksian outlier:

4. Outlier dapat mengubah kesimpulan
5. Pengeluaran data outlier tidak disalahkan, tetapi harus dikaji dahulu apakah data tersebut bagian dari populasi atau bukan.

Jika data outlier tidak dapat dikeluarkan karena masih merupakan subjek penelitian yang sebaiknya tetap digunakan, agar efek outlier dapat direduksi maka data dilakukan transformasi data, misalkan menggunakan logaritma natural atau akar kuadrat

3.4. K-Nearest Neighbors

Algoritma k-Nearest Neighbors (k-NN) merupakan algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (train data sets), yang diambil dari k tetangga terdekatnya (nearest neighbors). Dengan k merupakan banyaknya tetangga terdekat. Tujuan dari algoritma ini adalah mengklasifikasi objek baru berdasakan atribut dan sampel latih. pengklasifikasian tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Diberikan titik uji, akan ditemukan sejumlah k objek (titik training) yang paling dekat dengan titik uji. Klasifikasi menggunakan voting terbanyak di antara klasifikasi dari k objek. Algoritma k-NN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sample uji yang baru. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Eucledian

3.5. KMeans

Algoritma KMeans adalah algoritma pengelompokan dipartisi yang paling dikenal yang metode sederhana untuk memperkirakan rata-rata (vektor) kelompok K set. Kmeans yang paling banyak digunakan di antara semua algoritma clustering karena efisiensi dan kesederhanaannya[3]. KMeans yang pertama kali diusulkan oleh [4], adalah pengelompokan yang terkenal dan banyak digunakan algoritma. KMeans adalah salah satu algoritma pengelompokan paling sederhana dalam pembelajaran mesin yang bisa digunakan untuk secara otomatis mengenali kelompok objek serupa dalam pelatihan data.

Algoritma KMeans adalah sebagai berikut:

1. memilih titik data k sebagai klaster pusat
2. ulangi untuk setiap titik data $x \in D$
3. hitung jarak dari x ke masing-masing pusat;
4. menetapkan x ke pusat terdekat yang mewakili sebuah cluster
5. end for
6. menghitung ulang terpusat menggunakan keanggotaan cluster saat ini
7. sampai kriteria berhenti terpenuhi

KMeans clustering sensitif terhadap outlier dan dapat dianggap sebagai titik data yang tidak sesuai dengan titik normal yang menjadi ciri set data [6]

Dari hasil studi literatur didapatkan komparasi metode KNN dan KMeans terdapat pada Table 1.

Table 1. Komparasi.

Penulis/ Judu/ Tahun Publikasi	Deskripsi dan hasil	Algoritma
K.Swapna, Prof. M.S. Prasad Babu / *A Framework for Outlier*	merancang sistem diagnosis hati secara otomatis untuk mendeteksi secara dini dan akurat untuk mengurangi kematian yang disebabkan oleh penyakit hati serta menganalisis kumpulan data untuk memahami sistem guna	Cluster-based Bisecting kMeans,

Detection Using Improved Bisecting kMeans Clustering Algorithm (2017)	membantu mengembangkan sistem diagnosis hati otomatis. algoritma Bisecting KMeans Algorithm (IBKM) lebih unggul algoritma KMeans karena menghasilkan cluster yang lebih baik dan memberikan kluster yang efisien tanpa outlier. [11]	cluster validation
Sarunya Kanjanawattana / *A Novel Outlier Detection Applied to an Adaptive KMeans* (2019)	metode baru pemilihan pusat awal berdasarkan data kepadatan dan pendekatan baru deteksi outlier berdasarkan data jarak. Untuk metode baru pemilihan pusat awal, membandingkan jumlah iterasi dan skor Silhouette dari metode ini dan KMeans tradisional. Untuk sistem deteksi outlier, mengukur kinerja sistem dengan menggunakan matriks Fuzzy. Sebagai hasilnya, KMeans tradisional unggul karena kecepatan yang lebih tinggi dan hebat akurasi diperoleh[12].	KMeans dan pendekatan baru deteksi outlier berdasarkan data jarak.
Parmeet Kaur, Kanwarpreet Kaur / *A Review on Outlier Detection for Data Cleaning in Data Mining* (2016)	pembersihan data berkurang kesalahan dan meningkatkan kualitas data. Dengan bantuan deteksi outlier, dapat mendeteksi outlier dan meningkatkan kualitasnya data. Deteksi outlier digunakan untuk mendeteksi dan menghapus outlier dari data. meninjau berbagai pendekatan pengelompokan deteksi outlier yang digunakan untuk pembersihan data. algoritma kmean lebih baik daripada algoritma yang lain karena sederhana dan efisien. Dengan bantuan kmeans kita dapat mendeteksi outlier dan kemudian menghapus outlier itu untuk membersihkan data [13].	KMeans
C.Sumithiradevi, Dr.M.Punithavalli / *Enhanced KMeans with Greedy Algorithm for Outlier Detection* (2012)	teknik standar KMeans ditingkatkan menggunakan algoritma Greedy untuk deteksi dan penghapusan outlier yang efektif (EKMOD). Eksperimen pada set data iris mengungkapkan bahwa EKMOD secara otomatis mendeteksi dan menghapus pencilan. hasil keakurasian standard KMeans 89.80%, SKOD 93,25%, dan EKMOD 97,23% [14].	standard KMeans,EKMOD, dan SKOD
Yadigar Erdem, Caner Ozcan / *Fast Data Clustering and Outlier Detection Using KMeans Clustering On Apache Spark* (2017)	Hasil eksperimen membuktikan bahwa deteksi outlier berhasil dicapai setelah menggunakan algoritma KMeans di implementasikan dalam Spark MLlib. deteksi outlier diterapkan secara efisien pada dataset konsumsi daya listrik rumah tangga individu untuk menemukan data penipuan. data clustering dan deteksi outlier proses direalisasikan lebih cepat dengan menggunakan teknologi Spark pada big data[15].	KMeans (KNIME), KMeans (Spark), GMM (Spark)
Vishal Bhatt, Mradul Dhakar, Brijesh Kumar Chaurasia / *Filtered Clustering Based on Local Outlier Factor in Data Mining* (2016)	komparatif dari lima metodologi yang berbeda menggunakan KMeans sebagai algoritma dasar bersama dengan berbagai metode jarak yang digunakan dalam menemukan perbedaan antara objek maka untuk menganalisis efek outlier pada analisis cluster dataset dalam data mining. KMeans Sederhana dengan parameter yang difilter dan jarak Chebyshev sangat cocok untuk pengelompokan dalam penambangan data[16]	KMeans dengan filter parameter LOF dan jarak Chebyshev

Priyanga Dilini Talagala, Rob J. Hyndman, Catherine Leigh, Kerrie Mengersen and Kate Smith-Miles / *A feature-based framework for detecting technical outliers in water-quality data from in situ sensors* (2019)	Mengusulkan kerangka kerja otomatis yang menyediakan lebih awal deteksi outlier dalam data kualitas air dari sensor in situ yang disebabkan oleh masalah teknis Kerangka kerja yang pertama digunakan untuk mengidentifikasi fitur data yang membedakan. Teknik penilaian outlier tanpa pengawasan adalah kemudian diterapkan pada ruang data yang diubah dan pendekatan yang didasarkan pada teori nilai ekstrem digunakan untuk menghitung ambang batas untuk setiap outlier potensial[17].	KNN-AGG and KNN-SUM algorithms
Reema Aswani, S. P. Ghrera and Satish Chandra / *A Novel Approach to Outlier Detection using Modified Grey Wolf Optimization and k-Nearest Neighbors Algorithm* (2016)	model hybrid menggunakan meta-heuristik untuk mendeteksi anomali dataset secara efisien. Metode / Analisis Statistik: Algoritma optimasi serigala abu-abu yang dimodifikasi berbasis jarak dirancang yang menggunakan algoritma k- Nearest Neighbor untuk hasil yang lebih baik. Pendekatan yang diusulkan bekerja dengan baik dengan regresi serta dataset klasifikasi dalam skenario yang diawasi[18].	kNN dan kNN hybrid
Hegazy Zaher, Abd El-Fattah Kandil and Rehab Shehata / *An Alternative Artificial Intelligence Technique for Detecting* Outliers (2014)	mengusulkan algoritma hybrid termasuk K Nearest Neighbor dan Support Vector Machine (KSVM) yang mendeteksi outlier dengan mengambil keuntungan dari dua teknik cerdas, Support Vector Machine (SVM) dan K Nearest Neighbor (KNN). Skor efisiensi global tertinggi, (0,9148), untuk mendeteksi pencilan puas dengan dua metode SVM (RBF) dan KSVM. Meskipun metode yang diusulkan KSVM memiliki efisiensi yang sama dengan SVM (RBF) untuk data yang diberikan. Mungkin lebih cocok untuk data lain karena KSVM memiliki kelebihan dari dua metode SVM (RBF) dan KNN [19].	KSVM, SVM, dan KNN
R. Selvi1 and S. Saravan Kumar and A. Suresh / *An Intelligent Weighted Outlier Detection Method For Intrusion Detection Using Mst And K-NN* (2015)	Sistem deteksi intrusi yang efektif diperlukan untuk menyediakan komunikasi yang efektif di dunia masa lalu. Deteksi outlier adalah proses yang efektif untuk meningkatkan kinerja klasifikasi. Di masa lalu, banyak metode deteksi outlier dengan kombinasi metode clustering yang berbeda telah diusulkan. Ini semua memiliki keterbatasan dalam hal akurasi dan kecepatan. mengusulkan model pendeteksi outlier baru yang disebut Intelligent Weighted MST dan k-NN Based Outlier Detection (IWMKOD) untuk mendeteksi penyusup di semua jenis lingkungan jaringan. Hasil percobaan menunjukkan bahwa algoritma yang diusulkan meningkatkan akurasi deteksi dan mengurangi tingkat alarm palsu [20].	k-NN Based Outlier Detection (IWMKOD)

Heta Naik / *Credit Card Fraud Detection for Online Banking Transactions* (2018)	meningkatnya transaksi online berbanding lurus dengan meningkatnya jumlah penipuan. Dalam makalah ini berbagai algoritma seperti K- Nearest Neighbor, Random Tree, AdaBoost dan Logistic Regression beberapa tantangan yang termasuk membedakan antara transaksi normal dan penipuan yang tampaknya sangat mirip satu sama lain. Parameter untuk mendeteksi transaksi tersebut adalah Waktu, Jumlah dan Frekuensi Transaksi. Dalam tulisan ini, Berbeda empat algoritma KNN, AdaBoost, Random tree dan Regresi logistik dibandingkan untuk mekanisme deteksi penipuan. Regresi logistik lebih baik dibandingkan dengan algoritma lainnya. Model ini digunakan untuk data penipuan kartu kredit yang tidak seimbang. Semua algoritme ini tidak berlaku untuk deteksi penipuan pada saat transaksi [21].	KNN, AdaBoost, Random tree dan Regresi logistik
Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang and Xiangnan He / *Generative Adversarial Active Learning for Unsupervised Outlier Detection* (2019)	mendekati deteksi outlier sebagai masalah klasifikasi-biner dengan mengambil sampel pencilan potensial dari distribusi referensi yang seragam. Namun, karena jarangnya data dalam ruang dimensi tinggi, sejumlah outlier potensial mungkin gagal menyediakan informasi untuk membantu pengklasifikasi dalam menggambarkan batas yang dapat memisahkan pencilan dari data normal secara efektif. Untuk mengatasi ini, mengusulkan metode Pembelajaran Aktif Sasaran Aktif (SO-GAAL) Generatif Objektif Tunggal untuk deteksi outlier, yang dapat secara langsung menghasilkan pencilan potensial informative. mengusulkan algoritma deteksi outlier baru SOGAAL, yang secara langsung dapat menghasilkan outlier potensial informatif, untuk mengatasi kurangnya informasi yang disebabkan oleh kutukan dimensi. memperluas struktur GAAL dari generator tunggal (SO-GAAL) ke beberapa generator dengan tujuan yang berbeda (MO-GAAL) untuk mencegah generator dari jatuh ke mode masalah runtuh. dibandingkan dengan beberapa metode pendeteksi outlier yang canggih, MO-GAAL mencapai peringkat rata-rata terbaik pada dataset dunia nyata, dan menunjukkan kekokohan yang kuat untuk berbagai parameter. Selain itu, MO-GAAL dapat dengan mudah menangani berbagai jenis kluster dan rasio variabel tidak relevan yang tinggi, yang dapat diilustrasikan dengan hasil eksperimen pada set data sintetis. Meskipun runtime tidak memiliki keuntungan untuk dataset kecil, itu bukan cacat fatal dengan peningkatan daya komputasi [22].	algoritma deteksi outlier baru SOGAAL dan KNN
K.T.Divya, N.Senthil Kumaran / *Improved Outlier Detection Using Classic Knn Algorithm* (2016)	Pendekatan deteksi outlier didasarkan pada pembelajaran jarak jauh untuk atribut kategori (DILCA), kerangka pembelajaran jarak jauh diperkenalkan. Intuisi kunci dari DILCA adalah bahwa jarak antara dua nilai atribut kategoris dapat ditentukan dengan cara, di mana mereka terjadi bersamaan dengan nilai atribut lainnya dalam kumpulan data. Klasik KNN menghasilkan utilitas data yang unggul, tetapi menimbulkan overhead komputasi yang lebih tinggi. Selain itu teknik reduksi dimensi digunakan dalam dataset kesehatan kerja ini digunakan [23].	KNN

N. Nagarathinam, K. Karpagam / *Reverse Nearest Neighbours in Unsupervised Distance-Based Outlier Detection using FCM* (2016)	metode berbasis jarak dapat menghasilkan skor outlier yang lebih kontras dalam pengaturan dimensi tinggi. dimensi tinggi dapat memiliki dampak yang berbeda, dengan menguji kembali gagasan membalikkan tetangga terdekat dalam konteks deteksi outlier yang tidak diawasi. metode hybrid yang efisien untuk pemilihan fitur set kasar berdasarkan KNN dengan FCM clustering dan deteksi outlier berbasis jarak[24].	KNN dan FCM clustering

Berdasarkan pada table 1. KMeans tradisional unggul karena kecepatan yang lebih tinggi, hebat akurasi, sederhana, fleksibel, efisien, dan umum digunakan. Tetapi metode ini juga memiliki kelemahan dalam hal overlapping, sulit mencapai global optimum, sensitive terhadap titik awal. sedanglan metode KNN memiliki keunggulan berupa umum digunakan, mudah unntuk dipahami dan diimplementasikan, dan sangat non linier. Tetapi metode ini memiliki keterbatasan dalam hal akurasi dan kecepatan, perlu menentukan parameter K yang merupakan jumlah tetangga terdekat, memerlukan biaya komputasi yang lebih tinggi karena perlu menghitung jarak dari setiap sample , serta menimbulkan overhead komputasi yang lebih tinggi tetapi menghasilkan utilitas data yang unggul

4. Kesimpulan

Berdasarkan uraian diatas bahwa untuk mendeteksi outlier bisa menggunakan KNN maupun KMeans kedua metode tersebut sama-sama memiliki kelebihan dan kekurangan masing-masing. Walaupun kedua metode ini umum untuk digunakan dalam pendeteksian outlier. Secara umum metode pendeteksian outlier menggunakan KMeans lebih unggul dikarenakan fleksibel, efisien, kecepatan tinggi, serta sederhana. Tetapi untuk menentukan metode mana yang terbaik tergantung pada jenis data yang akan digunakan.

Ucapan Terimakasih

Terimakasih diucapkan kepada Ibu dan Bapak yang telah, sedang dan selalu mendoakan serta mendukung penulis untuk berprestasi. Terimakasih juga diucapkan kepada pimpinan Universitas Ahmad Dahlan dan Fakultas Pascasarjana UAD yang selalu memberikan dukungan kepada kami untuk selalu mengembangkan diri. Kepada ketua prodi Magister Pendidikan Matematika yang telah membimbing kami untuk menjadi mahasiswa S2 yang berkualitas dan berkaliber internasional dengan menghasilkan paper yang berkualitas. Kepada bapak ibu dosen prodi Magister Pendidikan Matematika yang selalu membagi ilmu kepada mahasiswa dengan tulus ikhlas.

References

[1] Marr, B 2016 big data in practice: *how 45 succesful companies used big data analytics to deliver extraordinary result*. John wiley &sons

[2] Bhatt V, Dhakar M, and Chaurasia B K 2016 . International Journal of Database Theory and Application *Filtered Clustering Based on Local Outlier Factor in Data Mining* **9** 275282

[3] Christopher T, and Divya T 2015 Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications *A Study of Clustering Based Algorithm for Outlier Detection in Data streams* 195197

[4] Macqueen J.B.,1967. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.

[5] Erdem Y, Ozcan C 2017 International Journal of Advanced Computational Engineering and Networking *Fast Data Clustering and Outlier Detection Using KMeans Clustering On Apache Spark* **5** 8690
[6] Kim S 2015 Communications for Statistical Applications and Methods *Variable Selection and Outlier Detection for Automated KMeans Clustering* **22** 5567
[7] Kannan K S and Manoj K 2016 International Journal of Computer Application *Clustering Algorithm Based Outlier Detection in Data Mining* **6** 0919
[8] Swapna K, and Babu M S P 2017 International Journal of Electrical & Computer Sciences IJECS-IJENS *A Framework for Outlier Detection Using Improved Bisecting KMeans Clustering Algorithm*. **17** 0812
[9] Wu B, Liao F, and Zhang D 2012 international conference on computer and information application ICCIA
[10] Sumithiradevi C, and Punithavalli M 2012 International Journal of Advanced Research in Computer Science *Enhanced KMeans with Greedy Algorithm for Outlier Detection* **3** 294297
[11] Swapna K, Babu M S P 2017 International Journal of Electrical & Computer Sciences IJECS-IJENS *A* Framework *for Outlier Detection Using Improved Bisecting kMeans Clustering Algorithm* **17 0**812
[12] Kanjanawattana S 2019 International Journal of Machine Learning and Computing *A Novel Outlier Detection Applied to an Adaptive Kmeans* **9** 569574
[13] Kaur P, Kaur K 2016 International Journal of Innovative Research in Computer and Communication Engineering *A Review on Outlier Detection for Data Cleaning in Data Mining.* **4** 1437314376
[14] Sumithiradevi C, Punithavalli M 2012 . International Journal of Advanced Research in Computer Science Enhanced *KMeans with Greedy Algorithm for Outlier Detection* **3** 294297
[15] Erdem Y, Ozcan C 2017 International Journal of Advanced Computational Engineering and Networking *Fast Data* Clustering *and Outlier Detection Using KMeans Clustering On Apache Spark* **5** 8690
[16] Bhatt V, dkk 2016 International Journal of Database Theory and Application *Filtered Clustering Based on Local Outlier Factor in Data Mining* **9** 275282
[17] Talagala P D, dkk 2019 Department of Econometrics and Business Statistics Monash University *A feature-based framework for detecting technical outliers in water-quality data from in situ sensors* .
[18] Aswani R, Ghrera S P, and Chandra S 2016 Indian Journal of Science and Technology *A Novel Approach to Outlier Detection using* Modified *Grey Wolf Optimization and k-Nearest Neighbors Algorithm* ***9***
[19] Zaher H, Kandil A E and Shehata R 2014 British Journal of Mathematics & Computer Science *An Alternative Artificial Intelligence Technique for Detecting Outliers* 27992810
[20] Selvil R dkk 2015 ARPN Journal of Engineering and Applied Sciences *An Intelligent Weighted Outlier Detection Method For Intrusion Detection Using Mst And K-NN* **10** 79527958
[21] Naik H 2018 International Journal for Research in Applied Science & Engineering Technology (IJRASET) *Credit Card Fraud Detection for Online Banking Transactions* ***6*** *453457*
[22] Liu Y dkk 2019 IEEE Transactions On Knowledge And Data Engineering *Generative Adversarial Active Learning for Unsupervised Outlier Detection*.
[23] Divya K T , Kumaran N S 2016 International Research Journal of Engineering and Technology (IRJET) *Improved Outlier Detection Using Classic Knn Algorithm* ***3*** 892898
[24] Nagarathinam N, Karpagam K 2016 International Journal of Advance Research in Computer Science and Management Studies *Reverse Nearest Neighbours in Unsupervised Distance-Based Outlier Detection using FCM* **4** 235241