ANALISIS GEROMBOL PADA PEUBAH CAMPURAN DENGAN DATA HILANG DAN PENCILAN TUNGGAL

NAFISA BERLIANA INDAH PRATIWI

PROGRAM STUDI STATISTIKA DAN SAINS DATA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul "Analisis Gerombol pada Peubah Campuran dengan Data Hilang dan Pencilan Tunggal" adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.
Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2024

Nafisa Berliana Indah Pratiwi
G1501211044

RINGKASAN

NAFISA BERLIANA INDAH PRATIWI. Analisis Gerombol pada Peubah Campuran dengan Data Hilang dan Pencilan Tunggal. Dibimbing oleh INDAHWATI dan ANWAR FITRIANTO.

Analisis gerombol merupakan teknik *unsupervised learning* yang telah banyak digunakan untuk mengelompokkan objek-objek yang mirip. Analisis gerombol menarik perhatian di berbagai bidang ilmiah seperti *machine learning*, *data mining*, dan *information retrieval*. Perhatian ini telah mendorong pengembangan pada berbagai pendekatan algoritme penggerombolan. Teknik penggerombolan yang umum digunakan biasanya berbasis jarak (*distance-based*) dan bervariasi tergantung pada jenis data yang digunakan (*data-based*), termasuk data numerik, data kategorik, dan data campuran yang terdiri atas peubah numerik dan kategorik. Salah satu tantangan yang dapat mengurangi kinerja algoritme penggerombolan adalah adanya data hilang dan pencilan tunggal. Dalam studi analisis gerombol dengan kasus data hilang, metode imputasi yang efektif menjadi kunci untuk meningkatkan akurasi hasil penggerombolan. Pengembangan terkini dalam algoritme penggerombolan melibatkan integrasi proses imputasi dengan proses penggerombolan. Oleh karena itu, penelitian ini berfokus pada kajian simulasi dan penerapan algoritme penggerombolan pada peubah campuran dengan data hilang dan pencilan tunggal.

Penelitian ini membandingkan tiga algoritme penggerombolan untuk data peubah campuran, algoritme *k-prototype*, yang merupakan algoritme penggerombolan pertama yang diciptakan untuk data peubah campuran; *simple k-medoids*, yang merupakan algoritme penggerombolan untuk data campuran yang dikembangkan berbasis *medoids*, umumnya kekar terhadap pencilan; dan *clustering mixed numerical and categorical data with missing values* (*k*-CMM), yang merupakan algoritme penggerombolan pertama untuk data peubah campuran yang mengintegrasikan proses imputasi dengan proses penggerombolan. Proses imputasi data hilang berbasis teknik *machine learning* dengan pendekatan berbasis pohon (*tree-based*), baik yang terpisah maupun terintegrasi. Selanjutnya, hasil penggerombolan dari ketiga algoritme tersebut dievaluasi menggunakan indeks validitas berdasarkan kriteria internal (*Silhouette*) dan kriteria eksternal (*Purity*, NMI, *Homogeneity*, *Completeness*, *dan V-measure*).

Penelitian ini terdiri atas kajian simulasi dan kajian empiris. Kajian simulasi bertujuan untuk memperoleh informasi mengenai performa ketiga algoritme tersebut dengan mempertimbangkan beberapa kondisi yang dapat memengaruhi gerombol hasil. Kondisi yang diamati dalam penelitian ini meliputi jumlah gerombol (k=3), jumlah observasi (N=100, 250, 500), banyak peubah campuran (p=32), proporsi pencilan (*Out* = 0.00, 0.10, 0.20), proporsi data hilang (*Emp* = 0.0004, 0.0008, 0.0012), dan proporsi tumpang tindih gerombol (Δ = 0.001, 0.10, 0.20). Setiap data yang dibangkitkan direplikasi sebanyak 30 kali, dan dari seluruh kombinasi kondisi tersebut diperoleh 81 skenario keseluruhan yang diamati.

Kajian empiris dilakukan dengan menerapkan ketiga algoritme tersebut pada data potensi desa Kabupaten Bogor tahun 2021, dengan acuan lima dimensi yang mewakili Indeks Pembangunan Desa (IPD). Lima dimensi tersebut meliputi (i) pelayanan dasar, (ii) kondisi infrastruktur, (iii) aksesibilitas / transportasi, (iv) pelayanan umum, dan (v) penyelenggaraan pemerintahan. Dalam penelitian ini,

digunakan 32 peubah campuran yang mewakili seluruh dimensi. Hasil penggerombolan kemudian dievaluasi menggunakan kriteria internal, sedangkan untuk evaluasi berdasarkan kriteria eksternal dibentuk dua skenario pembentukan *ground truth*: (i) menggunakan publikasi status Indeks Desa Membangun (IDM) untuk desa di Kabupaten Bogor tahun 2021, dan (ii) membangun *ground truth* berdasarkan algoritme penggerombolan *agglomerative nesting* (AGNES) yang dievaluasi menggunakan koefisien aglomeratif.

Hasil kajian simulasi menunjukkan bahwa algoritme *k-prototype* dengan proses imputasi terpisah memiliki performa paling unggul dibandingkan dengan kedua algoritme lainnya, termasuk *k*-CMM yang merupakan algoritme penggerombolan dengan proses imputasi terintegrasi. Algoritme *k-prototype* memiliki nilai indeks validitas tertinggi berdasarkan kriteria internal dan seluruh kriteria eksternal dibandingkan dengan *simple k-medoids* dan *k*-CMM. Algoritme *k-prototype* memberikan hasil terbaik terutama pada data yang tidak mengandung pencilan, sementara *simple k-medoids* menunjukkan kemampuan yang lebih baik dalam menggerombolkan data pada skenario dengan beberapa proporsi pencilan. Sebaliknya, *k*-CMM cenderung stabil untuk semua kondisi dengan nilai indeks validitas kedua kriteria yang relatif rendah. Proporsi data hilang yang dicakup oleh ketiga algoritme bervariasi dan mencakup seluruh jenis proporsi data hilang yang diuji.

Untuk memperoleh informasi terkait signifikansi pengaruh masing-masing skenario terhadap indeks validitas berdasarkan kriteria internal dan eksternal, dilakukan pemodelan regresi untuk ketiga algoritme. Hasilnya menunjukkan bahwa pada *k-prototype* dan *simple k-medoids*, kondisi seperti jumlah observasi, pencilan, dan tingkat tumpang tindih gerombol berpengaruh signifikan terhadap seluruh indeks validitas, sementara proporsi data hilang tidak memiliki pengaruh signifikan. Berbeda dengan *k*-CMM, untuk seluruh kondisi skenario, terutama kondisi data hilang, berpengaruh signifikan terhadap performa algoritme *k*-CMM yang diukur berdasarkan kriteria eksternal dan internal. Hal ini menunjukkan kesesuaian dengan tujuan utama dari pembentukan algoritme *k*-CMM, yaitu mengintegrasikan proses imputasi dengan proses penggerombolan untuk meningkatkan performa penggerombolan. Penelitian ini juga mengamati interaksi antar skenario untuk ketiga algoritme penggerombolan.

Hasil kajian empiris juga sejalan dengan hasil kajian simulasi, yakni algoritme *k-prototype* dengan proses imputasi terpisah tetap unggul dibandingkan kedua algoritme lainnya dalam menggerombolkan desa di Kabupaten Bogor, berdasarkan dimensi Indeks Pembangunan Desa (IPD) 2021. *k-prototype* konsisten memberikan nilai indeks validitas tertinggi berdasarkan kriteria eksternal pada kedua skenario *ground truth*. *Algoritme* k-prototype berhasil mengidentifikasi tiga gerombol desa yang memiliki karakteristik mirip dengan gerombol yang dihasilkan status IDM 2021 (Mandiri, Maju, Berkembang). Gerombol pertama (50 desa) yang dihasilkan oleh *k-prototype* setara dengan desa dalam kelompok "Mandiri", sementara gerombol kedua (264 desa) dan ketiga (119 desa) selaras dengan kelompok desa dengan status "Maju" dan "Berkembang". Integrasi data potensi desa yang diterapkan pada *k-prototype* dan status desa berdasarkan IDM memberikan wawasan penting untuk peningkatan kebijakan dan alokasi sumber daya yang lebih terarah bagi pembangunan desa di Kabupaten Bogor.

Kata kunci: Analisis gerombol, data hilang, pencilan tunggal, peubah campuran

SUMMARY

Nafisa Berliana Indah Pratiwi. Cluster Analysis in Mixed Variables with Missing Values and Univariate Outliers. Supervised by Indahwati and Anwar Fitrianto.

Cluster analysis is an unsupervised learning technique which has been widely employed to group similar objects. This technique gather attention across various scientific fields, including machine learning, data mining, and information retrieval. Such attention has led to the development of numerous clustering algorithm approaches. Commonly used clustering techniques are typically distance-based and also vary depends on the data type, such as numerical data, categorical data, and mixed data comprising both numerical and categorical variables. One of the key challenges that may obstruct clustering algorithms' performance is missing data and outliers' presence. In studies of cluster analysis involving missing data, effective imputation methods are crucial for enhance clustering results' accuracy. Recent advancements in clustering algorithms have involved the integration of imputation processes directly into the clustering process. Therefore, this research focuses on simulation study and the application of clustering algorithms on mixed data with missing values and univariate outliers.

This study compares three clustering algorithms for mixed data: k-prototype algorithm, which is the first algorithm developed for mixed data; simple k-medoids, that is medoid-based algorithm and typically robust to outliers; and clustering mixed numerical and categorical data with missing values (k-CMM), the first clustering algorithm for mixed data that integrates imputation process with clustering process. The both separated and integrated missing data imputation are based on machine learning techniques with tree-based approach. Subsequently, three algorithms' clustering results are evaluated by validity indices based on internal criteria (Silhouette) and external criteria (Purity, NMI, Homogeneity, Completeness, and V-measure).

This research comprises both simulation and empirical studies. The simulation study aims to gather further information on previously mentioned algorithms' performances under various conditions that may influence clustering results. The observed conditions include number of clusters (k = 3), number of observations (N = 100, 250, 500), number of mixed variables (p = 32), proportion of outliers (Out = 0.00, 0.10, 0.20), proportion of missing data (Emp = 0.0004, 0.0008, 0.0012), and degree of cluster separation / overlap (Δ = 0.001, 0.10, 0.20). Each dataset was replicated 30 times, and total of 81 scenarios were observed across all combinations of these conditions.

The empirical study was conducted by applying all three algorithms to the 2021 village potential data (PODES) from Bogor Regency, based on five dimensions representing the *indeks pembangunan desa* (IPD). These five dimensions include (i) basic services, (ii) infrastructure conditions, (iii) accessibility/transportation, (iv) public services, and (v) governance. In this study, 32 mixed variables reflect all five dimensions. The clustering results were then evaluated using internal criteria, while for evaluation based on external criteria, two ground truth scenarios were established: (i) utilize the published status of the *indeks desa membangun* (IDM) for villages in Bogor Regency in 2021, and (ii) form ground truth based on the agglomerative nesting (AGNES) clustering algorithm, which is evaluated using the agglomerative coefficient.

The simulation study results indicate the k-prototype algorithm with separate imputation processes outperformed the other two algorithms, also include k-CMM, which integrates the imputation process into the clustering process. The k-prototype algorithm achieved the highest validity index scores based on internal criteria and overall external criteria compared to simple k-medoids and k-CMM. k-prototype algorithm produced the best results, particularly for data without outliers, while simple k-medoids demonstrated better performance in clustering data with certain outlier scenarios. Conversely, k-CMM tended to be stable across all conditions, with relatively lower validity index values for both internal and external criteria. The range of missing data proportions handled by the three algorithms varied and encompassed all missing data proportions.

In order to determine the significance of each scenario's influence on validity indices based on internal and external criteria, regression model was performed for the three algorithms. The results revealed that k-prototype and simple k-medoids, conditions such as the number of observations, outliers, and the degree of cluster separation/overlap significantly influenced all validity indices on two different criteria, while the proportion of missing data did not have a significant impact. In contrast, for only k-CMM, all scenario conditions, particularly the presence of missing data, significantly influenced the performance of the k-CMM algorithm as measured by both criteria. This finding aligns with the primary goal of k-CMM algorithm, which is to integrate the imputation process with clustering to improve clustering performance. The study also examined the interaction between scenarios for all three clustering algorithms.

The empirical study results also aligned with the simulation study findings, reveals the k-prototype algorithm with separate imputation processes remained better to the other two algorithms in clustering villages in Bogor Regency based on the 2021 IPD dimensions. The k-prototype algorithm consistently provided the highest validity index values based on external criteria in both ground truth scenarios. The k-prototype algorithm successfully identified three village clusters with characteristics similar to those generated by the 2021 IDM status namely "Mandiri" (Self-reliant), "Maju" (Advanced), and "Berkembang" (Developing). The first cluster (50 villages) identified by the k-prototype algorithm corresponds to villages in the "Self-reliant" cluster, while the second (264 villages) and third clusters (119 villages) align with villages in the "Advanced" and "Developing" status clusters, respectively. The integration of village potential data applied to the k-prototype algorithm and village status based on IDM provides valuable insights for enhancing policies and more targeted resource allocation for village development in Bogor Regency.

Keywords: cluster analysis, missing data, mixed data, univariate outlier

ANALISIS GEROMBOL PADA PEUBAH CAMPURAN DENGAN DATA HILANG DAN PENCILAN TUNGGAL

NAFISA BERLIANA INDAH PRATIWI

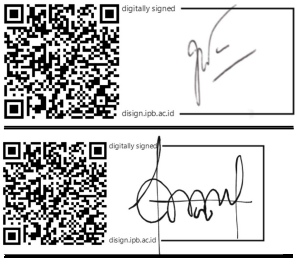
Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Statistika dan Sains Data

PROGRAM STUDI STATISTIKA DAN SAINS DATA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024

Tim Penguji pada Ujian Tesis:
Prof. Dr. Ir. Muhammad Nur Aidi, M.S.

Judul Tesis : Analisis Gerombol pada Peubah Campuran dengan Data Hilang dan Pencilan Tunggal
Nama : Nafisa Berliana Indah Pratiwi
NIM : G1501211044

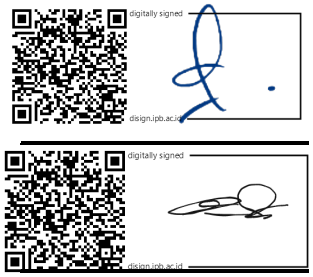
Disetujui oleh



Pembimbing 1:
Dr. Ir. Indahwati, S.Si., M.Si.

Pembimbing 2:
Dr. Anwar Fitrianto, S.Si., M.Sc.

Diketahui oleh



Ketua Program Studi:
Dr. Agus Mohamad Soleh, S.Si., M.T.
NIP 197503151999031004

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam:
Dr. Berry Juliandi, S.Si., M.Si.
NIP 19780723 2007011001

Tanggal Ujian: 15 Agustus 2024 Tanggal Lulus:

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Januari 2023 sampai bulan Juli 2024 ini ialah "Analisis Gerombol pada Peubah Campuran dengan Data Hilang dan Pencilan Tunggal".

Terima kasih penulis ucapkan kepada para pembimbing, Dr. Indahwati, M.Si. dan Dr.Anwar Fitrianto, S.Si., M.Sc. yang telah membimbing dan banyak memberi saran, ucapan terima kasih juga disampaikan kepada penguji luar komisi pembimbing Prof. Dr. Ir. Muhammad Nur Aidi, M.S, serta kepada pimpingan sidang Sachnaz Desta Oktarina S.Stat., M.Agr.Sc., Ph.D. Ungkapan terimakasih juga disampaikan kepada Tai Dinh, Ph.D. (dosen The Kyoto College of Graduate Studies for Informatics) sebagai pengembang algoritme *k*-CMM yang telah ikut mengembangkan penelitian ini, dan Dr.nat.techn Weksi Budiaji, M.Sc. (dosen Universitas Sultan Ageng Tirtayasa) sebagai pengembang algoritme *simple k-medoids* atas waktu yang telah diluangkan untuk berdiskusi selama penelitian berlangsung. Ungkapan terima kasih juga disampaikan kepada bapak Suranto, ibu Enny Pratiwi, kakak Annisa Nur Pratiwi & adik Fadhila Diah Ayu Pratiwi, dan seluruh rekan yang telah memberikan dukungan, doa, dan kasih sayangnya sehingga penulis dapat menyelesaikan studi.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Agustus 2024

Nafisa Berliana Indah Pratiwi

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN