ARS

by Endi Sukemi

Submission date: 01-Oct-2021 04:41PM (UTC+0700)
Submission ID: 1662414227
File name: paper.pdf (1.29M)
Word count: 1798
Character count: 11861

ARS 2019 The 5th Annual Research Seminar

Prosiding Annual Research Seminar 2019

Computer Science and ICT

ISBN : 978-979-587-846-9

Vol.5 No.1

Agglomerative Hierarchical Clustering Dengan Berbagai Pengukuran Jarak Dalam Mengklaster Daerah Berdasarkan Tingkat Kemiskinan

Endy Suherman
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
endysuherman11@google.com

Ermatita
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
ermatitaz@yahoo.com

Sukemi
Fakultas Ilmu Komputer
Universitas Sriwijaya
Palembang, Indonesia
sukemi@ilkom.unsri.ac.id

***Abstract*—Information that is efficient, effective, right on target and can be trusted, is a powerful instrument to be the basis for policy making. This has become an important aspect in supporting problem-solving strategies, one of which is poverty. Problems that become global and national issues. The purpose of this research is to develop a framework that applies the CRISP-DM methodology and the Agglomerative Hierarchical Clustering technique using several variations of distance measurements, in producing regional cluster analysis based on poverty levels in the East Kalimantan region.**

***Abstrak*—Informasi yang efisien, efektif, tepat sasaran dan dapat dipercaya, merupakan instrument tangguh untuk menjadi dasar dalam pengambilan kebijakan. Hal ini menjadi aspek penting dalam mendukung strategi penanggulangan masalah, salah satunya masalah kemiskinan. Masalah yang menjadi isu global maupun nasional. Tujuan dari penelitian ini adalah membangun kerangka kerja yang menerapkan metodologi CRISP-DM dan teknik *Agglomerative Hierarchical Clustering* menggunakan beberapa variasi pengukuran jarak, dalam menghasilkan analisis klaster wilayah berdasarkan tingkat kemiskinan di daerah Kalimantan Timur.**

***Keywords*—Agglomerative Hierarchical Clustering, CRISP-DM, Distance Measure**

I. PENDAHULUAN

Di Indonesia, masalah kemiskinan menjadi prioritas pembangunan. Masalah ini merupakan masalah yang kompleks dan bersifat multidimensional sehingga program-program pembangunan terus digerakan untuk meningkatkan kesejahteraan masyarakat.

Aspek penting dalam mendukung strategi penanggulangan kemiskinan adalah tersedianya data kemiskinan yang akurat. Badan pusat Statistik (BPS) terus melakukan perhitungan jumlah dan persentase penduduk miskin di Indonesia. Pengukuran kemiskinan yang dapat dipercaya mampu menjadi instrumen tangguh bagi pengambilan kebijakan sehingga upaya penanggulangan kemiskinan dapat berjalan lebih efisien, efektif, dan tepat sasaran [1].

Clustering adalah teknik dari *data mining* dimana data yang dinilai serupa ditempatkan ke dalam kelompok yang terkait atau homogen tanpa harus memiliki pengetahuan tentang definisi dari kelompok tersebut [14]. Dengan mengelompokan objek yang memiliki kemiripan maksimum atau kemiripan minimum dengan grup objek lain, pendekatan ini dapat berguna untuk mengekplorasi data dalam dataset yang tidak memiliki laber secara objektif dan menjadi ringkasan untuk tugas-tugas *data mining* pada permasalahan yang kompleks [15].Untuk mendapatkan informasi yang lebih efisien, efektif dan tepat sasaran untuk mendukung strategi penanggulangan kemiskinan ini, penulis mencoba membuat kerangka kerja dengan menerapkan metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM)[2-3] dan metode *Agglomerative Hierarchical Clustering* (AHC). Dalam pengukuran klaster ini, juga diterapkan beberapa variasi pengukuran, untuk mendapatkan klaster yang paling tepat untuk data yang dipakai.

Penelitian terdahulu [12], menerapkan *Complate Linkage* dan metode *Hierarchical Clustering Multiscale Bootstrap*, untuk menentukan jumlah klaster menggunakan data kemiskinan di Kalimantan Timur Tahun 2016. Penelitian lain [17], melakukan perbandiangan hasil klaster menggunakan *single linkage* dan metode *C-Means*. Perbedaan penelitian yang dilakukan adalah dari segi penggunaan metode dan metodologi yang dipakai.

II. TINJAUAN PUSTAKA

A. *Kemiskinan*

Kemiskinan adalah kondisi kehidupan yang serba kekurangan yang dirasakan oleh seorang atau rumah tangga sehingga tidak mampu memenuhi kebutuhan primer atau yang layak bagi kebutuhan. Aspek-aspek yang menyangkut kemiskinan adalah aspek ekonomi, politik, dan sosial-psikologi[5].

Kemiskinan berarti kondisi dimana orang atau kelompok orang memiliki peluang lebih besar terhadap resiko dan tekanan penyakit dan kenaikan harga-harga bahan makanan dan uang sekolah secara tiba-tiba [16].

B. *Agglomerative Hierarchical Clustering*

Pengelompokan klaster menggunakan model hirarki, dapat dilakukan dengan dua arah, menggabungkan titik-titik individual ke klaster pada tingkat paling tinggi atau membaginya dari klaster paling atas ke objek yang paling kecil. *Agglomerative Hierarchical Clustering* (AHC) adalah pendekatan yang melakukan proses dari objek paling bawah hingga membentuk klaster teratas [6]. AHC akan menyajikan hasil dalam bentuk dendrogram yang memrepresentasikan pengempokan pola dan tingkat kesamaan pengelompokan.

Teknik algoritma untuk mengukur jarak klaster yang biasanya digunakan pada AHC adalah *Single Linkage, Complate Linkage,* dan *Average Linkage*. *Single Linkage* melihat jarak antara 2 anggota klaster terdekat, *Complate*

ISBN : 978-979-587-846-9
Vol.5 No.1

Linkage melihat jarak antara 2 anggota klaster terjauh, dan *Average Linkage* melihat jarak rata antara masing-masing anggota.

C. Pengukuran Jarak

Fungsi yang digunakan untuk mengidentifikasi kesamaan antara dua data menjadi kunci penting dari proses pengelompokan. Data dapat bervariasi dimana dapat dibentuk sebagai nilai mentah dengan panjang yang sama atau tidak, atau dibentuk sebagai vektor [11].

Misalkan x_i dan v_i dalam sebuah dimensi vektor P, maka untuk *Euclidean Distance* (ED) [6-8], *Manhattan Distance* (MD) [8], *Minkowski Distance* (MKD) [9-10] masing-masing akan diukur seperti :

$$d_{ij} = \sqrt{\sum_{i=1}^{p} (x_{ik} - v_{jk})^2} \quad (1)$$

$$d_{ij} = \sum_{i=1}^{p} |x_{ik} - v_{jk}| \quad (2)$$

$$d_{ij} = \left(\sum_{i=1}^{p} (x_{ik} - v_{jk})^q \right)^{1/q} \quad (3)$$

Dalam persamaan di atas, q adalah bilangan bulat positif. Versi yang dinormalisasi dapat didefinisikan jika nilai yang diukur dinormalisasi melalui pembagian dengan nilai maksimum dalam urutan.

III. Metodologi Penelitian

Penelitian ini menggunakan metodologi *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [2-3]. Alur kerja dari CRISP-DM meliputi *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment*.

A. Business Understanding

Fase awal ini difokuskan untuk memahami tujuan dan persyaratan dan kemudian mengubah pengetahuan menjadi definisi masalah pada *datamining*, serta merancang rencana untuk mencapai tujuan.

B. Data Understanding

Tahapan dimana memulai untuk memahami data dan melakukan pengumpulan data awal, untuk mengidentifikasi masalah dari kualitas data, menemukan wawasan awal, dan mendeteksi hipotesis dari informasi yang tersembunyi.

C. Data Preparation

Fase ini mencakup semua kegiatan untuk menyusun dataset akhir dari data mentah. Data yang dihasilkan dapat konstruksi atribut baru, dan transformasi data untuk permodelan nantinya.

D. Modeling

Pada fase ini, berbagai teknik permodelan akan dipilih dan diterapkan. Beberapa teknik dapat digunakan untuk jenis masalah data yang sama. Namun beberapa teknik memerlukan bentuk format data yang tertentu.

E. Evaluation

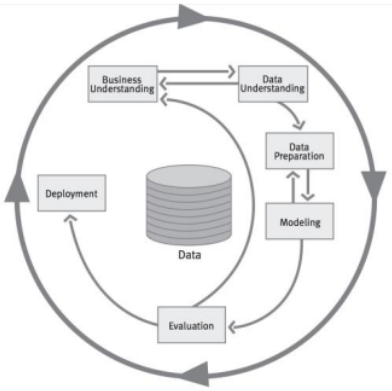
Pada tahap ini, model yang terbangun memiliki kualitas yang baik dari prespektif analisis data. Tujuan utama pada tahap ini adalah untuk menentukan apakah ada beberapa masalah bisnis yang belum dipertimbangkan sebelumnya. Tahapan ini juga harus menunjukan hasil *data mining* yang diingikan.

F. Deployment

Pengetahuan yang diperoleh masih perlu untuk diatur dan disajikan sedemikian rupa agar dapat digunakan. Penting untuk memahami dari awal, kegiatan yang akan dilakukan agar dapat benar-benar memanfaatkan model yang dibuat.

IV. Hasil Dan Pembahasan

Dengan menerapkan CRISP-DM, alur model dapat digambarkan seperti Gambar 1. Namun pada penelitian ini, pekerjaan dilakukan mulai dari *Business Understanding* hingga *Modeling*.



Gambar. 1. Keseluruhan Alur Model CRISP-DM

Data yang digunakan pada penelitian ini adalah data yang dipublis oleh Badan Pusat Statistik Kalimatan Timur yang ditebitkan 2017 [4]. Untuk lebih spesifik, data yang digunakan adalah data tentang kemiskinan di daerah Kalimatan Timur. Mengikuti penelitian sebelumnya [12], objek pengamatan yang digunakan pada penelitian ini, sebagai berikut :

- Kota Balikpapan (Objek 1).
- Kabupaten Kutai Kartanegara (Objek 2).
- Kota Samarinda (Objek 3).
- Kabupaten Penajam Paser Utara (Objek 4).

- Kabupaten Paser (Objek 5).
- Kabupaten Kutai Barat (Objek 6).
- Kabupaten Kutai Timur (Objek 7).
- Kabupaten Berau (Objek 8).
- Kota Bontang (Objek 9).
- Kabupaten Mahakam Ulu (Objek 10).

Sedangkan fitur yang digunakan adalah :

- Persentase jumlah penduduk miskin usia 15 tahun ke atas yang tidak bekerja (satuan dalam %) (Fitur A).
- Persentase jumlah rumah tangga yang pernah membeli beras raskin(satuan dalam %) (Fitur B).
- Persentase jumlah pengeluaran perkapita untuk non makanan (satuan dalam %) (Fitur C).
- Persentase jumlah penduduk miskin usia 15 tahun ke atas yang tidak tamat SD (satuan dalam %) (Fitur D).
- Persentase angka Melek huruf penduduk miskin usia 15-55 tahun (satuan dalam %) (Fitur E).
- Persentase jumlah pengguna alat KB di rumah tangga miskin (satuan dalam %) (Fitur F)
- Persentase jumlah rumah tangga miskin dengan luas lantai perkapita ⩽8 m2 (satuan dalam %) (Fitur G).
- Persentase jumlah rumah tangga miskin yang menggunakan air bersih.(satuan dalam %) (Fitur H).
- Persentase jumlah rumah tangga miskin yang mendapatkan pelayanan jaminan kesehatan (satuan dalam %) (Fitur I).
- Persentase rumah tangga miskin yang menggunakan jamban sendiri/bersama (satuan dalam %) (Fitur J).

Untuk pengelompokan klaster, penelitian ini menggunakan model hirarki, *Agglomerative Hierarchical Clustering* (AHC). Pada model AHC akan diterapkan *Complate Linkage* dengan tiga pengukuran jarak termasuk *Euclidean Distance* (ED) [6-8], *Manhattan Distance* (MD) [8] dan *Minkowski Distance* (MKD) [9-10]. Hasil dari tiga pengukuran jarak yang berbeda ini, digunakan sebagai pembanding dalam menganalisis klaster.

V. KESIMPULAN

Dengan menerapkan metodologi CRISP-DM, tahapan-tahapan yang akan dijalani untuk menganalisis kluster wilayah berdasarkan tingkat kemiskinan ini, menjadi lebih jelas dan terstruktur. Menggunakan sepuluh objek dan sepuluh fitur sebagai data pengukuran kemiskinan di Kalimantan Timur. Penerapan AHC dan berbagai pengukuran jarak, memberikan gambaran model untuk menganalisis klaster wilayah berdasarkan tingkat kemiskinan di Kalimantan Timur. Tahapan *evaluation* dan *deployment* menjadi target pada penelitian selanjutnya.

REFERENCES

[1] Badan Pusat Statistik. (2018). *Penghitungan dan Analisis Kemiskinan Makro dan Indonesia Tahun 2018*. Jakarta: BPS.

[2] Wirth, R., and Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000.

[3] The CRISP-DM process model (1999), http://www.crisp-dm.org/

[4] Badan Pusat Statistik Kalimantan Timur, (2017). *Kalimantan Timur Dalam Angka 2017*. Samarinda.

[5] Suharto, E. (2005). *Membangun Masyarakat Memberdayakan Rakyat, Kajian Strategis Pembangunan Kesejahteraan Sosial Dan Pekerjaan Sosial*. Bandung: PT Rafika Aditama.

[6] Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. Expert Systems with Applications, 42(5), 2785–2797.

[7] Danielsson, P.-E. (1980). Euclidean distance mapping. Computer Graphics and Image Processing, 14(3), 227–248.

[8] Thakare, Y. S., & Bagal, S. B. (2015). Performance evaluation of K-means clustering algorithm with various distance metrics. *International Journal of Computer Applications*, *110*(11), 12-16.

[9] Sammour, M., & Othman, Z. (2016). An agglomerative hierarchical clustering with various distance measurements for ground level ozone clustering in Putrajaya, Malaysia. *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 1127-1133.

[10] Chouikhi, H., Saad, M. F., & Alimi, A. M. (2017). Improved fuzzy possibilistic C-means (IFPCM) algorithms using Minkowski distance. 2017 International Conference on Control, Automation and Diagnosis (ICCAD).

[11] T. Warren Liao. Clustering of time series data-a survey. Pattern Recognition, 38: 1857– 1874, 2005.

[12] Ramadhani, L., Purnamasari, I., & Amijaya, F. D. T. (2018). Penerapan Metode Complete Linkage dan Metode Hierarchical Clustering Multiscale Bootstrap. *JURNAL EKSPONENSIAL*, *9*(1), 1- 10.

[13] Mirkin, B. (2016). *Clustering: a data recovery approach*. Chapman and Hall/CRC.

[14] P. Rai, S. Singh, A survey of clustering techniques, Int. J. Comput. Appl. 7 (12) (2010) 1–5.

[15] Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering – A decade review. Information Systems, 53, 16–38.

[16] Indra, P. (2001). An Analysis Towards Urban Poverty Alleviation Program in Indonesia. Philosophy Doctor Dissertation. Faculty of the School Policy, Planning, and Development. University of Southern California, California.

[17] Goreti, M., Nasution, Y. N., & Wahyuningsih, S. (2017). Perbandingan Hasil Analisis Cluster dengan Menggunakan Metode Single Linkage dan Metode C-Means. *JURNAL EKSPONENSIAL*, *7*(1), 9-16.

ARS

ORIGINALITY REPORT

16%	%	16%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Nugroho Irawan Febianto, Nicodias Palasara. "Analisa Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018", Jurnal Sisfokom (Sistem Informasi dan Komputer), 2019 Publication	3%
2	Shantika Martha, Yundari Yundari, Setyo Wira Rizki, Ray Tamtama. "PENERAPAN METODE GEOGRAPHICALLY WEIGHTED PANEL REGRESSION (GWPR) PADA KASUS KEMISKINAN DI INDONESIA", BAREKENG: Jurnal Ilmu Matematika dan Terapan, 2021 Publication	2%
3	Muhammad Ibnu Sa'ad, Kusrini, M. Syukri Mustafa. "Student Prediction of Drop Out Using Extreme Learning Machine (ELM) Algorithm", 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020 Publication	2%

4	Rifqi Hammad, Kurniadin Abd Latif, Kartarina Kartarina, Pahrul Irfan et al. "Sosialisasi Computational Thingking Pada Guru MTs Yayasan NW Darul Abror Gunung Rajak Lombok Barat", Jurnal Pengabdi, 2021 Publication	2%
5	Ummul Hairah. "Pengembangan Sistem Manajemen Database dan Pengambilan Keputusan kriteria Penduduk Miskin Kabupaten Kutai Kartanegara Kalimantan timur", ILKOM Jurnal Ilmiah, 2016 Publication	1%
6	Johannes Riedl, Daniel Puckmayr, Dominik Brunner. "Traktor-Assistenzsysteme mittels Machine Learning", ATZheavy duty, 2020 Publication	1%
7	Johannes Petrus, Ermatita, Sukemi. "Soft and Hard Clustering for Abstract Scientific Paper in Indonesian", 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2019 Publication	1%
8	Marban, O.. "Toward data mining engineering: A software engineering approach", Information Systems, 200903 Publication	1%

9	Mohammad Iwan Wahyuddin, Rima Tamara Aldisa, Fauziah Fauziah, Ira Diana Sholihati. "Sistem Informasi Administrasi Kemahasiswaan dan Alumni (Smart Adma) dengan Metode Extreme Programming (XP)", Jurnal JTIK (Jurnal Teknologi Informasi dan Komunikasi), 2021 Publication	1 %
10	Muhammad Ilham Alhari, Widia Febriyani, Wader Trisepa Jonson, Asti Amalia Nur Fajrillah. "Perancangan Smart Village Platform Aplikasi Edukatif untuk Pengentasan Stunting serta Monitoring Kesehatan Ibu Hamil", Jurnal Ilmiah Teknologi Informasi Asia, 2021 Publication	1 %
11	Rudi Nurdiansyah. "Optimasi Penjadwalan Flow Shop Menggunakan Algoritma Hybrid Differential Evolution", Rekayasa Energi Manufaktur, 2017 Publication	1 %
12	"Emerging Trends in Intelligent Computing and Informatics", Springer Science and Business Media LLC, 2020 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off

Exclude bibliography On