JURNAL ILMIAH SEMESTA TEKNIKA
Vol. 18, No. 1, 76-82, Mei 2015

Penerapan Metode K-Means Untuk *Clustering* Mahasiswa Berdasarkan Nilai Akademik Dengan Weka *Interface* Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang

(Implementation Method for K-Means Clustering Based Student Value with Weka Interface a Case Study of Department of Information UMM Magelang)

ASRONI, RONALD ADRIAN

ABSTRACT

The selection process among outstanding students in a department has a big problem. This process is not fair because only involve one criteria and ignore the other criteria. We need the best student to participate in a competition held by the Indonesia Security Incident Response Team on Internet Infrastructure (ID SIRTII) of the Ministry of Communication and Information. This process uses Weka software to calculate the best student. It provides the various method to explore the data. One of them is clustering method. There are many algorithms in clustering method. In this research, we will investigate widely about one of that algorithms. Its name is K-Means. This algorithm (K-Means) will give the recommendations about the best student based on the cluster. It will represent the many clusters of a student group. The best cluster can be calculated more to get the names of the best students group. They are eligible to enter the competition. K-means involve the GPA (Grade Point Average) and related course to support the academic skill in order to get the best student. This research helps the teacher select the best student to enter the competition. Many similar cases can use this algorithm in order to get the best student.

Keywords: Clustering, K-Means, Algorithm

PENDAHULUAN

Universitas Muhammadiyah Magelang sebagai salah satu PTS di Indonesia memiliki 6 Fakultas dan telah memiliki beberapa prestasi di bidang Teknologi Informasi salah satunya adalah Juara 1 Lomba *Cyber Forensic* Tingkat Nasional oleh Tim Teknik UMM Magelang pada tahun 2014. Dengan prestasi yang telah diraih, maka diperlukan pola rekruitmen yang baik untuk memperoleh mahasiswa yang akan mewakili Lomba tersebut. Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Magelang mengalami kesulitan untuk mencari 5 orang mahasiswa yang akan dikirimkan untuk mengikuti lomba pada kompetisi *event* Cyberjawara yang diselenggarakan oleh Indonesia *Security Incident Response Team on Internet Infrastructure* (ID SIRTII) Kementerian Komunikasi dan Informatika RI. Pada kompetisi tersebut akan dibutuhkan kemampuan untuk melakukan analisis untuk memecahkan permasalahan terkait logika pemrograman untuk bisa meraih juara yang diharapkan. Dari permasalahan tersebut dan target yang ingin dicapai, maka diperlukan sebuah proses seleksi yang dilakukan dengan menyeleksi mahasiswa yang memiliki kriteria yang baik sesuai variabel-variabel yang dijadikan acuan penyeleksian. Variabel-variabel yang akan dijadikan acuan adalah: nilai matakuliah Algoritma dan Pemrograman, Fisika Dasar, Kalukulus dan Indek Prestasi Komuliatif (IPK). Dari permasalahan diperlukan sebuah pengelompokan terhadap kriteria-kriteria yang ada. Kriteria tersebut diproses dengan menggunakan metode pengelompokan K-Means. Metode K-Means diperlukan karena mampu menentukan pengelompokan mahasiswa dengan kriteria yang bisa jadi acuan untuk menentukan keputusan terhadap mahasiswa yang akan dikirimkan sebagai peserta Lomba.

Penelitian ini menguji data yang telah ada di *data warehouse* Universitas Muhammadiyah Magelang untuk memudahkan untuk mencari 5 orang mahasiswa pada jurusan Teknik Informatika dalam melakukan penyeleksian untuk mengikuti lomba. Lomba yang akan diikuti adalah kompetisi event Cyberjawara yang diselenggarakan oleh Indonesia *Security Incident Response Team on Internet Infrastructure* (ID SIRTII) Kementerian Komunikasi dan Informatika RI. Pengujian data

ini pada fase awal memilih siapa saja mahasiswa yang berpeluang untuk mengikuti event Cyberjawara. Fase awal sistem melakukan proses pengelompokan terhadap kriteria-kriteria yang ada. Kriteria tersebut diproses dengan menggunakan metode pengelompokan K-Means.

Diharapkan dengan adanya pengujian data ini dapat memberikan rekomendasi mahasiswa yang layak maju sebagai peserta event Cyber jawara.

Dewasa ini pengolahan *data warhouse* telah menjadi kebutuhan yang sangat utama. Perkembangan pesat dalam teknologi informasi yang menjadikan semua informasi dapat disimpan dalam jaringan komputer telah membuat munculnya sistem basis data yang sangat besar yaitu *data warehouse*. Dalam hitungan detik, data-data dalam berbagai basis data akan senatiasa terbarukan, baik dikarenakan adanya *update* maupun penambahan data baru. Permasalahan yang kemudian muncul adalah bagaimana mengetahui informasi yang terdapat dalam *data warehouse* yang sangat besar.

Knowledge discovery in Database (KDD) didefinisikan sebagai ekstraksi informasi potensial, implisit dan tidak dikenal dari sekumpulan data. Proses *knowledge discovery* melibatkan hasil dari proses *data mining* (proses mengekstrak kecenderungan pola suatu data), kemudian mengubah hasilnya secara akurat menjadi informasi yang mudah dipahami.

Ada beberapa macam pendekatan berbeda yang diklasifikasikan sebagai teknik pencarian informasi/pengetahuan dalam KDD. Ada pendekatan kuantitif, seperti pendekatan *probabilistic* and statistik. Beberapa pendekatan memanfaatkan teknik visualisasi, pendekatan klasifikasi seperti logika induktif, pencarian pola, dan analisis pohon keputusan. Pendekatan yang lain meliputi deviasi, analisis kecenderungan, algoritma genetik, jaringan syaraf tiruan dan pendekatan campuran dua atau lebih dari beberapa pendekatan yang ada.

Wrigt (1998) melakukan pembagian enam elemen yang paling esensial dalam teknik pencarian informasi/ pengetahuan dalam KDD, yaitu:

1. mengerjakan sejumlah besar data,
2. diperlukan efisiensi berkaitan dengan volume data,
3. mengutamakan ketepatan/keakuratan,
4. membutuhkan pemakaian bahasa tingkat tinggi,
5. menggunakan beberapa bentuk dari pembelajaran otomatis, dan
6. menghasilkan hasil yang menarik

Clustering

Berkhin (2006) menyatakan salah satu metode yang diterapkan dalam KDD adalah *clustering*. *Clustering* adalah membagi data ke dalam grup-grup yang mempunyai obyek yang karakteristiknya sama. Garcia (2002) menyatakan *clustering* adalah mengelompokkan item data ke dalam sejumlah kecil grup sedemikian sehingga masing-masing grup mempunyai sesuatu persamaan yang esensial.

Andayani (2007) menyatakan *clustering* memegang peranan penting dalam aplikasi data mining, misalnya eksplorasi data ilmu pengetahuan, pengaksesan informasi dan text mining, aplikasi basis data spasial, dan analisis web. *Clustering* diterapkan dalam mesin pencari di Internet. Web mesin pencari akan mencari ratusan dokumen yang cocok dengan kata kunci yang dimasukkan. Dokumen-dokumen tersebut dikelompokkan dalam *cluster-cluster* sesuai dengan kata-kata yang digunakan.

Du (2010) menjelaskan bahwa klasterisasi adalah proses membagi data yang tidak berlabel menjadi kelompok-kelompok data yang memiliki kemiripan. Misalkan *K* adalah jumlah klaster, *C* merupakan label klaster, dan P merupakan dataset. Klasterisasi harus memenuhi kriteria sebagai berikut:

$$C_i \neq \Phi, \forall i \in \{1,2,...,K\} \quad (1)$$

$$C_i \cap C_j = \Phi, \forall i \neq jandi, j \in \{1,2,...,K\} \quad (2)$$

$$\bigcup_{i=1}^{K} Ci = P \quad (3)$$

Kategori clustering

Tan (2006) membagi *clustering* dalam dua kelompok, yaitu *hierarchical and partitional clustering*. *Partitional Clustering* disebutkan sebagai pembagian obyek-obyek data ke dalam kelompok yang tidak saling *overlap* sehingga

setiap data berada tepat di satu *cluster*. *Hierarchical clustering* adalah sekelopok cluster yang bersarang seperti sebuah pohon berjenjang (hirarki).

METODE PENELITIAN

Penelitian ini menggunakan algoritma K-Means untuk menentukan cluster yang terbaik. Cluster terbaik ini dipergunakan untuk pemilihan mahasiswa-mahasiswa terbaik yang dapat diikutkan lomba. Sehingga peluang untuk mendapatkan juara dalam lomba bisa semakin besar.

K-Means

Algoritma K-Means merupakan algoritma klasterisasi yang mengelompokkan data berdasarkan titik pusat klaster (*centroid*) terdekat dengan data. Tujuan dari K-Means adalah pengelompokkan data dengan memaksimalkan kemiripan data dalam satu klaster dan meminimalkan kemiripan data antar klaster. Ukuran kemiripan yang digunakan dalam klaster adalah fungsi jarak. Sehingga pemaksimalan kemiripan data didapatkan berdasarkan jarak terpendek antara data terhadap titik centroid.

Tahapan awal yang dilakukan pada proses klasterisasi data dengan menggunakan algoritma K-Means adalah pembentukan titik awal *centroid* cj Pada umumnya pembentukan titik awal *centroid* dibangkitkan secara acak. Jumlah *centroid* cj yang dibangkitkan sesuai dengan jumlah klaster yang ditentukan di awal. Setelah k *centroid* terbentuk kemudian dihitung jarak tiap data xi dengan *centroid* ke-j sampai k, dinotasikan dengan d (xi,cj). Terdapat beberapa ukuran jarak yang digunakan sebagai ukuran kemiripan suatu *instance* data, salah satunya adalah jarak *Euclid*. Perhitungan jarak *Euclidean* seperti pada Persamaan 4.

$$d(Xi, Cj) = \sqrt{\sum_{i=1}^{N} (Xi - Cj)^2} \quad (4)$$

Duran dan Odell (1974) menyatakan jika $d(Xi, Cj)$ semakin kecil, kesamaan antara dua unit pengamatan semakin dekat. Syarat menggunakan jarak *Euclid* adalah jika semua fitur dalam dataset tidak saling berkorelasi. Jika terdapat fitur yang berkorelasi maka menggunakan konsep jarak Mahalanobis.

Agusta (2007) menyatakan kelanjutn dari jarak tersebut dicari yang terdekat sehingga data akan mengelompok berdasarkan *centroid* yang paling dekat. Tahap berikutnya adalah update titik *centroid* dengan menghitung rata-rata jarak seluruh data terhadap *centroi*d. Selanjutnya akan kembali lagi ke proses awal. Iterasi ini akan diulangi terus sampai didapatkan *centroid* yang konstan artinya titik *centroid* sudah tidak berubah lagi. Atau iterasi dihentikan berdasarkan jumlah iterasi maksimal yang ditentukan.

Software yang digunakan dalam penelitian ini adalah Weka. Tujuan dari penggunaan *software* ini adalah membandingkan hasil dengan perhitungan secara teoritis dengan hasil yang didapatkan dengan proses di Weka *Interface* ini. Alat penelitian Weka *Interface,* seperti tampak pada Gambar 1 adalah aplikasi data mining *open source* berbasis Java. Aplikasi ini dikembangkan pertama kali oleh Universitas Waikato di Selandia Baru. Weka memiliki banyak algoritma *machine learnin*g yang dapat digunakan untuk melakukan generalisasi atau formulasi dari sekumpulan data sampling. Salah satunya adalah *clustering* dengan menggunakan algoritma K-Means.

Sharma (2012) menyatakan teknik *clustering* memiliki penggunaan yang luas dan saat ini memiliki kecenderungan yang semakin meningkat seiring dengan jumlah data yang terus berkembang. K-means adalah teknik sederhana untuk analisis klastering. Tujuannya adalah untuk menemukan divisi terbaik entitas n ke dalam kelompok k (disebut cluster), sehingga total jarak antara anggota kelompok dan entroid sesuai, terlepas dari kelompok diminimalkan. Setiap entitas milik *cluster* dengan *mean* terdekat. Ini hasil ke partisi ruang data ke Voronoi Sel.

GAMBAR 1. Weka *Interface*

Lokasi Penelitian

Penelitian dilaksanakan di Prodi Teknik Informatika, Universitas Muhammadiyah Magelang. Universitas Muhammadiyah Magelang sebelumnya sudah membangun sebuah Sistem *Data Warehouse.*

HASIL DAN PEMBAHASAN

Data Pengujian

Data pengujian yang digunakan adalah berupa tabel yang memiliki komponen penyusun sebagai berikut:

1. Memiliki 5 attribute yaitu nim mahasiswa, nilai mata kuliah algoritma dan pemrogaman 1, nilai mata kuliah fisika dasar, nilai kalkulus 1, dan IPK.
2. Jumlah *instance* adalah 124

Algoritma K-Means

Oyelade (2010) menyatakan algoritma ini disusun atas dasar ide yang sederhana. Sebaran obyek dan elemen pertama dalam *cluster* dapat dipilih untuk dijadikan sebagai titik tengah (*centroid point*) *cluster*. Algoritma metode K-Means selanjutnya akan melakukan pengulangan langkah-langkah berikut sampai terjadi kestabilan (tidak ada obyek yang dapat dipindahkan):

1. Menentukan koordinat titik tengah setiap *cluste.* Penentuan *cluster* dibuat 4 buah obyek dengan 3 atribut. Metode *Clustering* dengan algoritma K-Means akan menghasilkan 4 *cluster* berdasarkan *class* Indek Prestasi Komulatif (IPK). Pengesetan nilai awal tengah dengan menentukan titik tengah (*centroid*) dari *cluster* seperti pada table 1.
2. Penentuan nilai dari *cluster-cluster* tersebut untuk dijadikan acuan untuk melakukan perhitungan pada setiap baris tabel data pengujian. Contoh penentuan jarak obyek ke *centroid*, pengujian dilakukan pada NIM = 12.0504.0009 yang memiliki IPK = 0.25, dengan mengacu pada rumus Euclid yang telah disederhanakan (*cluster* x-IPK):
   Jarak 0 = (0.5 - 0.25) = 0.25
   Jarak 1 = (3.4 - 0.25) = 3.15
   Jarak 2 = (2.3 - 0.25) = 2.05
   Jarak 3 = (2.9 - 0.25) = 2.65
   Dari hasil perhitungan maka didapatkan hasil seperti pada Tabel 2

   Dari hasil diperoleh jarak 0 = 0.25 akan mendekati nilai pada *cluster* 0, maka proses yang sama akan dilakukan untuk semua data pengujian.
3. Pengelompokan obyek-obyek tersebut berdasarkan pada jarak minimumnya dilakukan dengan menggunakan hasil proses pada langkah 2. Hasil pada perhitungan jarak akan digunakan untuk penentuan *clustering,* seperti pada Gambar 2.

Pengujian dengan Software Weka

Pengujian data dengan *Software* Weka menghasilkan data berupa:

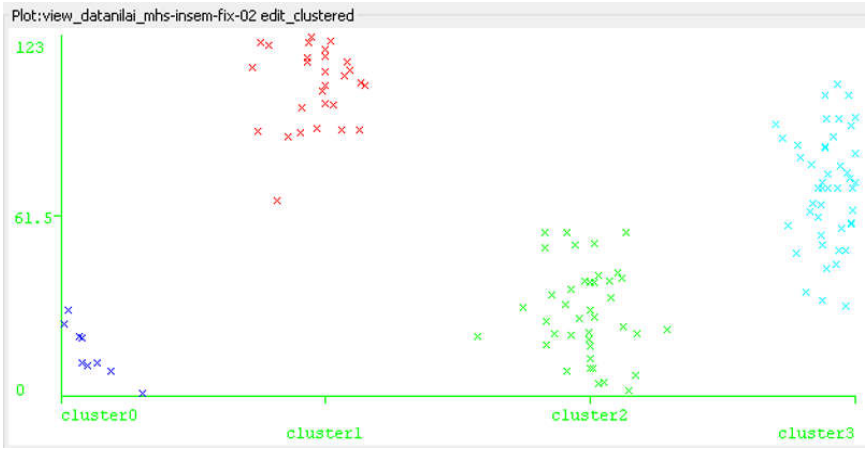
1. Nilai *cluster centroids* dan *cluster instances* seperti pada Gambar 3.
2. Grafik *clustering* posisi mahasiswa pada setiap *cluster* masing-masing seperti pada Gambar 4.

TABEL 1. *Cluster centroids*

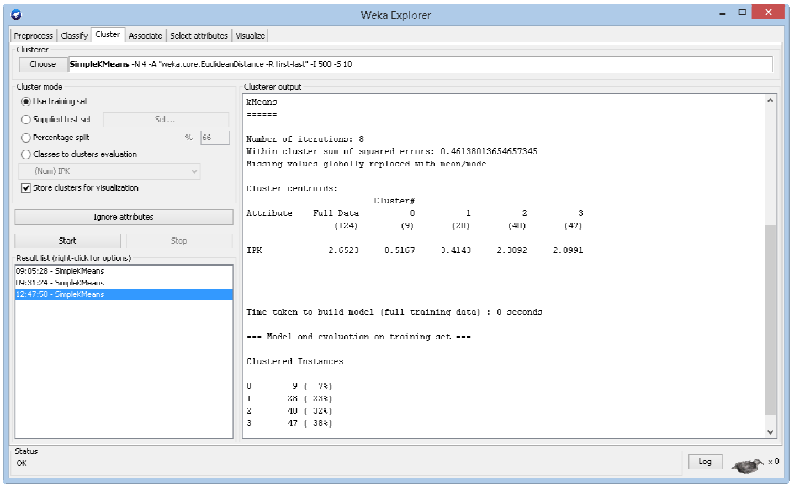
cluster 0	*cluster* 1	*cluster* 2	*cluster* 3
0.5	3.4	2.3	2,9

TABEL 2. Penentuan Jarak Obyek

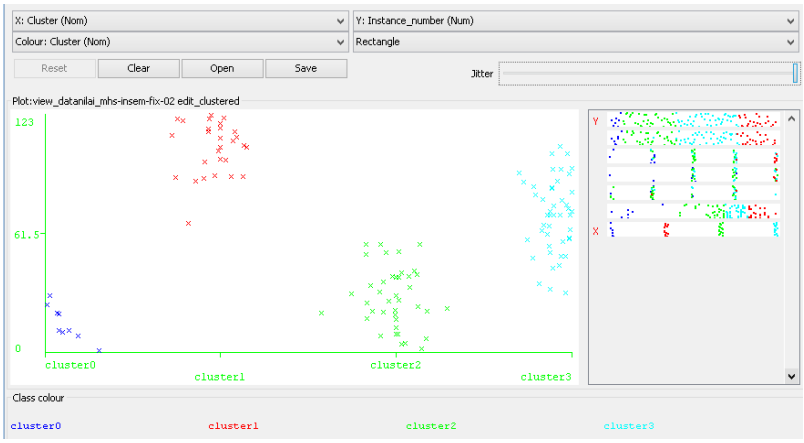
NIM	**IPK**	**Jarak 0**	Jarak 1	Jarak 2	Jarak 3
12.0504.0009	**0.25**	**0.25**	3.15	2.05	2.65



GAMBAR 2. Plot grafik *clustering*



GAMBAR 3. Hasil *cluster centroids* dan *clustered Instances* dengan Weka



GAMBAR 4. Hasil Grafik *clustering* dengan Weka

Hasil Pengujian dengan Software Weka

Dari data yang dilatih, didapatkan 4 kelompok dengan hasil sebagai berikut:

1. Mahasiswa dengan IPK = 0.5167 untuk *cluster* 0, sebanyak 9 Mahasiswa dari 124 Mahasiswa (7%)
2. Mahasiswa dengan IPK = 3.4143 untuk *cluster* 1, sebanyak 28 Mahasiswa dari 124 Mahasiswa (23%)
3. Mahasiswa dengan IPK = 3.3092 untuk *cluster* 2, sebanyak 40 Mahasiswa dari 124 Mahasiswa (32%)
4. Mahasiswa dengan IPK = 3.8991 untuk *cluster* 3, sebanyak 47 Mahasiswa dari 124 Mahasiswa (38%)

Maka *cluster* 1 dengan IPK tertinggi bisa digunakan untuk memilih 5 Mahasiswa untuk bisa mewakili lomba.

KESIMPULAN

Berdasarkan penelitian yang dilakukan, dapat disimpulkan bahwa algoritma K-Means bisa digunakan untuk mengelompokkan mahasiswa berdasarkan IPK dan beberapa atribut mata kuliah.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada pihak-pihak yang telah membantu dalam penelitian yang dipublikasikan dalam jurnal ini. Pihak-pihak yang terkait adalah Jurusan Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Magelang.

DAFTAR PUSTAKA

Agusta, Yudi. "K-means–penerapan, permasalahan dan metode terkait." *Jurnal Sistem dan Informatika* 3.47-60 (2007).

Andayani, Sri. Pembentukan cluster dalam Knowledge Discovery in Database dengan Algoritma K-Means. *SEMNAS Matematika dan Pendidikan Matematika 2007 dengan tema "Trend Penelitian Matematika dan Pendidikan Matematika di Era Global*, 2007.

Berkhin, Pavel. A survey of clustering data mining techniques. In: *Grouping multidimensional data*. Springer Berlin Heidelberg, 2006. p. 25-71.

DU, K.-L. Clustering: A neural network approach. *Neural Networks*, 2010, 23.1: 89-107.

Duran, Benjamin S.; ODELL, Patrick L. Cluster analysis. 1974.

Garcia, H. M.; ULLMAN, J.; WIDOM, Jennifer. Database systems: The complete book. 2002.

Oyelade, O. J., O. O. Oladipupo, and I. C. Obagbuwa. "Application of k Means Clustering algorithm for prediction of Students Academic Performance." arXiv preprint arXiv:1002.2425 (2010).

Sharma, Ritu; Alam, M. Afshar; Rani, Anita. K-Means clustering in spatial data mining using weka interface. In: *International Conference on Advances in Communication and Computing Technologies (ICACACT Proceedings published by International Journal of Computer Applications®(IJCA), pp. 26-30*. 2012.

Sunardi, Tommy Anandra; Rochimah, Siti; AKBAR, Rizky Januar. Rancang Bangun Aplikasi Rekomendasi Mawapres. *Rancang Bangun Aplikasi Rekomendasi Mawapres*, 2015.

Tan, Pang-Ning, et al. *Introduction to data mining*. Boston: Pearson Addison Wesley, 2006.

Wright, Peggy. Knowledge discovery in databases: Tools and Techniques.*Crossroads*, 1998, 5.2: 23-26.

PENULIS:

Asroni

Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Muhammadiyah Yogyakarta, Jalan Lingkar Selatan, Bantul 55183, Yogyakarta.

Email: asroni@umy.ac.id

Ronald Adrian

Program Studi Teknologi Informasi, Fakultas Teknik, Universitas Muhammadiyah Yogyakarta, Jalan Lingkar Selatan, Bantul 55183, Yogyakarta..

Email: ronald@ft.umy.ac.id