Supervised Clustering of Label Ranking Data by Grbovic

Dalam makalah ini kami mempelajari pengelompokan terawasi dalam konteks data peringkat label. Segmentasi data kompleks tersebut memiliki banyak potensi aplikasi dunia nyata. Misalnya, dalam pemasaran target, tujuannya adalah untuk mengelompokkan pelanggan di ruang fitur dengan mempertimbangkan preferensi produk yang ditetapkan dan berpotensi tidak lengkap, sehingga preferensi instans dalam suatu kluster lebih mirip daripada preferensi pelanggan di kluster lain. . Kami menetapkan beberapa dasar heuristik untuk aplikasi ini yang menggunakan algoritma terkenal seperti K-means, dan mengusulkan algoritma berprinsip yang dirancang khusus untuk jenis pengelompokan ini. Hal ini didasarkan pada model peringkat probabilistik Plackett-Luce (PL). Setiap cluster direpresentasikan sebagai gabungan sel Voronoi yang didefinisikan oleh satu set prototipe dan diberi satu set skor label PL yang menentukan peringkat label spesifik cluster. Parameter PL cluster yang tidak diketahui dan posisi prototipe ditentukan menggunakan teknik pembelajaran terawasi. Keanggotaan cluster dan peringkat untuk instance baru ditentukan oleh keanggotaan prototipe terdekatnya. Algoritma yang diusulkan dievaluasi secara empiris pada data peringkat label sintetis dan kehidupan nyata. Metode berbasis PL lebih unggul daripada pendekatan pengelompokan terawasi berbasis heuristik. Algoritma berbasis PL yang diusulkan juga dievaluasi pada tugas prediksi peringkat label. Hasilnya menunjukkan bahwa ini sangat kompetitif dengan algoritme peringkat label yang canggih, dan sangat akurat pada data dengan peringkat parsial.

Aplikasi tambahan meliputi: meta-learning [24], di mana, diberikan set data baru, tugasnya adalah untuk menginduksi peringkat total algoritma yang tersedia sesuai dengan kesesuaian mereka berdasarkan properti set data; memprediksi preferensi makanan untuk pelanggan baru berdasarkan hasil survei, demografi, dan karakteristik responden lainnya [17]; menentukan urutan pertanyaan dalam survei untuk pengguna tertentu berdasarkan atribut responden. Lihat [8] untuk ikhtisar aplikasi peringkat label di bidang ekonomi, penelitian operasi, dan database. Pengelompokan data peringkat label yang diawasi adalah masalah terbuka dan tidak sepele, yang belum dibahas dalam literatur penambangan data. Secara tradisional, teknik pengelompokan tidak diawasi, dan dengan demikian tidak mempertimbangkan keanggotaan kelas (klasifikasi) atau nilai target (regresi). Pengelompokan yang diawasi [9, 10, 11], di sisi lain, tidak. Ini bertujuan untuk menghasilkan pengelompokan yang diinginkan dengan informasi tambahan (misalnya, label kelas). Contoh aplikasi mencakup pengelompokan artikel berita berdasarkan apakah artikel tersebut merujuk pada topik yang sama. Tergantung pada aplikasi yang sebenarnya, biasanya ada ukuran kinerja khusus atau penghargaan untuk mengevaluasi solusi pengelompokan potensial.

Seseorang dapat membayangkan banyak aplikasi potensial dari pengelompokan terawasi untuk data peringkat label. Misalnya, dalam pemasaran sasaran, sebuah perusahaan dengan beberapa produk ingin mengelompokkan pelanggannya untuk tujuan merancang materi promosi khusus klaster. Untuk setiap klaster, perusahaan dapat membuat katalog yang berbeda, dengan mempromosikan subset produk yang dipilih dan merancang katalog dengan cara yang paling menolak selera pelanggan sasarannya. Penerapan lainnya adalah bagaimana mengurutkan pertanyaan dalam survei untuk kelompok sasaran tertentu untuk memaksimalkan keberhasilannya. Misalnya, mengingat data berlabel pengguna yang menyelesaikan survei, tujuannya adalah mengelompokkan pengguna ini ke dalam jumlah grup yang telah ditentukan sebelumnya dan mempelajari urutan pertanyaan terbaik untuk setiap grup. Secara khusus, dalam pengelompokan terawasi, tujuannya adalah untuk mengelompokkan data berdasarkan fitur instans dan peringkat label yang ditetapkan, yang berpotensi tidak lengkap, sehingga peringkat label instans dalam sebuah kluster lebih mirip satu sama lain daripada sebelumnya. label peringkat instance di cluster lain. Jika kita menganggap peringkat label sebagai klasifikasi, di mana setiap permutasi diperlakukan sebagai kelas yang berbeda, kita akan memiliki L! kelas, dan bahkan lebih ketika berhadapan dengan peringkat yang tidak lengkap. Ini akan membuat sangat sulit untuk menerapkan algoritma pengelompokan terawasi standar. Karena pengelompokan data peringkat label yang diawasi, sejauh pengetahuan kami, belum pernah dipelajari sebelumnya, kami mengusulkan beberapa algoritma dasar heuristik dan juga mengusulkan model probabilistik berprinsip. Pendekatan baseline pertama menggunakan algoritma K-means yang terkenal, baik dengan memperlakukan peringkat label sebagai salah satu fitur, atau dengan pengelompokan pertama berdasarkan fitur instan saja dan kemudian menetapkan peringkat label ke cluster yang diperoleh menggunakan Model Mallow umum [ 18]. Pendekatan dasar kedua pertama mengelompokkan peringkat label [19], tanpa memperhitungkan fitur instans, untuk mendapatkan jumlah kelas yang telah ditentukan sebelumnya. Ini kemudian melatih pengklasifikasi multi-kelas menggunakan data klasifikasi yang baru dibentuk.

Model Campuran Plackett-Luce (MMPL) yang diusulkan menyajikan kerangka kerja umum untuk peringkat label yang dapat digunakan baik untuk pengelompokan yang diawasi maupun untuk prediksi. Ini didasarkan pada representasi klaster multi-prototipe, di mana preferensi klaster yang mendasari dimodelkan menggunakan parameter skor Plackett-Luce khusus klaster. Modelnya cepat, dengan pelatihan linier dan waktu prediksi, penskalaan memori konstan dengan jumlah prototipe dan mampu bekerja secara efisien dengan peringkat yang tidak lengkap.

K-berarti! Algoritma Mallows pertama-tama melakukan K-means clustering berdasarkan fitur instance saja, tanpa memperhitungkan peringkat labelnya. Ini kemudian menurunkan peringkat pusat k untuk setiap cluster dari peringkat label yang dimiliki cluster ini menggunakan Model Mallows. Pendekatan ini diharapkan bekerja dengan baik ketika

cluster di ruang fitur sesuai dengan cluster di ruang label. Na ve K-means Dalam pendekatan ini, peringkat label diperlakukan sebagai salah satu fitur. Hal ini dilakukan dengan menambahkan L atribut tambahan ke vektor fitur x, yang menghasilkan vektor baru dengan panjang d+L. Nilai dari (d+j)- atribut ke-j diatur ke posisi label ke-j di peringkat , atau L=2 jika label tertentu tidak tersedia di . Setelah tahap preprocessing ini, data yang baru terbentuk di cluster menggunakan algoritma benchmark K-means. Peringkat pusat k untuk setiap cluster diturunkan dari centroid cluster yang diperoleh dengan mengurutkan atribut L terakhir dalam urutan menaik. Pemeringkatan ini digunakan untuk memprediksi peringkat label dari instance baru ketika keanggotaan clusternya ditentukan dengan menemukan centroid cluster terdekat di ruang fitur asli.

Peringkat Label Tanpa Pengawasan! SVM pertama-tama mengelompokkan peringkat label menggunakan pendekatan peringkat label tanpa pengawasan untuk mendapatkan kelas K (peringkat pusat cluster). Kemudian melatih kelas SVM multi-kelas

dengan kernel RBF menggunakan data yang baru diberi label. Ketika sebuah instance baru diklasifikasikan, itu diberi peringkat yang sesuai dengan kelas yang diprediksi.

Segmentation Model of Customer Lifetime Value in Small and Medium Enterprise (SMEs) using K-Means Clustering and LRFM Model by Marisa Fitri

Model CLV adalah ukuran keuntungan pelanggan bagi perusahaan yang dapat digunakan untuk mengevaluasi nilai masa depan pelanggan. Model CLV adalah ukuran keuntungan pelanggan bagi perusahaan yang dapat digunakan untuk mengevaluasi nilai masa depan pelanggan. Penelitian ini bertujuan untuk mendapatkan Customer Lifetime Value (CLV) pada setiap segmen pelanggan. Pengelompokan menggunakan metode K-Means Clustering berdasarkan model LRFM (Length, Recency, Frequency, Monetary). Proses pembentukan cluster menggunakan Metode Elbow dan SSE dengan jumlah cluster terbaik = 2 cluster. Nilai CLV dihasilkan dari perkalian hasil normalisasi LRFM dan nilai bobot LFRM kemudian dijumlahkan, dan dilakukan pada setiap cluster yang telah terbentuk. Peringkat tertinggi diantara 2 klaster tersebut berada pada klaster kedua dengan nilai CLV paling tinggi dibandingkan rata-rata klaster lainnya sebesar 0,362. Berdasarkan matriks LRFM, cluster ini memiliki nilai loyalitas yang tinggi dengan simbol LRFM L R F M yang merupakan pelanggan setia (segmen terbaik yang memiliki nilai loyalitas pelanggan tinggi). Berdasarkan simbol LRFM, perusahaan dapat membuat strategi untuk mempertahankan pelanggan dan memperoleh pelanggan untuk menjadi pelanggan setia dengan profitabilitas tinggi.

Pelanggan adalah aset bagi perusahaan, dan mereka memiliki preferensi yang berbeda.[3]. Saat ini, perusahaan berlomba-lomba untuk menarik perhatian pelanggan, dan penelitian ini [4] adalah salah satu bidang yang mempelajarinya. Pada kenyataannya, setiap pelanggan memiliki kebutuhan, harapan, dan perilaku yang berbeda, sehingga perusahaan diharapkan tidak memperlakukan mereka dengan perlakuan yang sama. Pelanggan tidak hanya memiliki preferensi, harapan, dan kebutuhan yang berbeda tetapi juga pendapatan dan profil biaya yang berbeda, oleh karena itu harus dikelola dalam perlakuan yang berbeda. [5],[6],[7]. Perlakuan yang sama terhadap semua pelanggan akan menyebabkan pelanggan yang bernilai kecil bagi perusahaan akan menurunkan nilai pelanggan bagi perusahaan. Oleh karena itu, diperlukan suatu metode untuk mengetahui nilai pelanggan bagi perusahaan. Metode CLV adalah metode yang menghubungkan nilai yang diberikan perusahaan kepada pelanggan dan nilai yang diberikan pelanggan kepada perusahaan selama periode hubungan antara keduanya. Menurut [8] Nilai pelanggan dapat dilihat dengan mengetahui nilai yang dihasilkan saat ini dari semua keuntungan untuk masa depan perusahaan. Sedangkan menurut [6] hasil yang diperoleh dari model CLV dapat digunakan untuk mengevaluasi nilai masa depan pelanggan kepada perusahaan dengan teknik data mining dalam mendeteksi pola dan hubungan dalam histori data pelanggan. RFM (Recency, Frequency and Monetary) merupakan salah satu model yang dapat mengukur nilai CLV, yang mampu memperkirakan nilai kehidupan pelanggan [9],[10].

Selain itu, kelebihannya adalah model ini juga dapat mengukur nilai profitabilitas pelanggan. [11]. Model RFM pertama kali dikembangkan oleh Hughes (1994), merupakan salah satu metode segmentasi yang paling umum yang dapat mengidentifikasi nilai pelanggan di perusahaan dengan 3 variabel: kebaruan, komunikasi dan moneter [12]. Perhitungan recency menunjukkan waktu sejak pelanggan terakhir melakukan transaksi pembelian, frekuensi menunjukkan jumlah transaksi pembelian, sedangkan moneter menunjukkan nilai pembelian [13]. Kemudian model RFM diperluas dengan satu variabel lain: Length oleh Chang dan Tasy (2004) [5] yang disebut model LRFM (Length, Recency, Frequency and Monetary) [14]. Panjang variabel ini adalah variabel yang mengukur lamanya transaksi pertama dan terakhir pelanggan pada suatu periode waktu. Hal ini penting untuk mengukur dalam memperkuat nilai resensi yang dihasilkan nasabah berdasarkan lamanya transaksi terakhir dengan jangka waktu tertentu. [5]. Model LRFM memiliki keuntungan karena lebih tepat untuk menganalisis waktu hidup pelanggan. Kerugian dari model RFM adalah tidak mempertimbangkan loyalitas pelanggan, dan ini telah direvisi dalam model LRFM. [15]. Langkah utama dalam mengukur nilai profitabilitas pelanggan adalah mengidentifikasi pelanggan dengan menggunakan pendekatan segmentasi pelanggan [9]. Dalam melakukan proses segmentasi, pelanggan yang sudah ada dibagi menjadi beberapa kelompok menurut standar tertentu. [11]. Data mining merupakan ilmu yang dapat digunakan untuk melakukan proses segmentasi. [16] “Data mining adalah serangkaian proses untuk menemukan hubungan antar item untuk membentuk pola pengetahuan baru dari data berukuran besar dan diproses dengan beberapa teknik yang melibatkan ilmu pengetahuan termasuk statistik, dan matematika.”.

CLV adalah bagian dari aktivitas Model Hubungan Pelanggan (CRM). Data mining dapat mengekstrak informasi tentang pelanggan yang penting untuk membangun strategi yang efektif (CRM). Kegiatan CRM yang dapat didukung melalui penggunaan data mining antara lain: (1) segmentasi pelanggan akan menghasilkan segmen pelanggan berdasarkan karakteristik perilaku pelanggan, (2) kampanye pemasaran langsung untuk berkomunikasi langsung dengan pelanggan dalam meningkatkan jumlah pelanggan baru dan pembelian produk. , dan (3) analisis keranjang pasar untuk mengidentifikasi produk terkait yang biasanya dibeli bersama. [17].

Metode yang digunakan dalam segmentasi pelanggan adalah metode clustering. Clustering juga disebut sebagai segmentasi data dalam beberapa aplikasi karena pengelompokan data yang besar ke dalam kelompok memiliki kesamaan dengan kelompok tersebut.[18]. Clustering adalah proses pengelompokan dataset fisik dan abstrak dari objek ke dalam kelompok objek yang memiliki kemiripan. [19], [20]. Clustering merupakan metode yang banyak digunakan di berbagai bidang antara lain segmentasi pelanggan, perilaku pelanggan, profitabilitas pelanggan, pendataan kebakaran hutan dan lain sebagainya. Berbagai algoritma digunakan dalam clustering: K-Means, DBSCAN (Density Based Spatial Clustering of Application with Noise), SOM (Self Organizing Map) dan lain-lain. Metode yang paling umum digunakan adalah algoritma K-Means. [5].

Namun, clustering K-Means memiliki kelemahan pada akurasi penentuan jumlah cluster. [21], [5], [13], [12]. Banyak peneliti telah meneliti metode validasi dalam menentukan jumlah cluster terbaik. Salah satu metode yang baik untuk menentukan jumlah cluster adalah Metode Elbow. [22]. Untuk menentukan jumlah cluster menggunakan metode Elbow. Metode ini digunakan dalam analisis klaster untuk interpretasi dan pengujian kinerja tingkat konsistensi jumlah klaster yang tepat dengan melihat nilai SSE. [23]. Pada titik tertentu grafik akan menurun drastis dengan kurva yang disebut kriteria siku. Nilai tersebut kemudian menjadi nilai terbaik dari k atau jumlah cluster. [23].

Oleh karena itu penelitian ini bertujuan untuk menghasilkan Customer Lifetime Value (CLV) dengan algoritma clustering K-Means berdasarkan model LRFM (Length, Recency, Frequency dan Monetary). Dataset yang digunakan sebagai eksperimen adalah data UKM penjualan pulsa semua operator selama periode 1 Januari 2018 sampai dengan 30 Juni 2018. Sebelum proses clustering dilakukan dilakukan normalisasi dan transformasi untuk menghasilkan dataset yang siap untuk proses clustering. Penentuan jumlah cluster terbaik dihitung dengan metode elbow. Kemudian hasil CLV dari pembobotan data LFRM dan proses analisis dan pemeringkatan.

Analysis on E-commerce Order Cancellations Using Market Segmentation Approach by JIngi Ye

Penelitian ini menginvestigasi penerapan segmentasi pasar pada e-commerce cancelled order. Ini menggunakan dataset transnasional yang berisi transaksi toko ritel online selama setahun. Proses analisis meliputi 1) analisis data eksplorasi pada pesanan yang dibatalkan yang membentuk sejumlah besar dataset untuk menunjukkan karakteristiknya. 2) segmentasi produksi yang memanfaatkan k-means clustering untuk menghasilkan 5 klaster produk. 3) segmentasi pelanggan dengan k-means clustering menggunakan segmen produksi dan fitur pelanggan yang menghasilkan 7 segmen. Dalam prosesnya, studi membandingkan skor siluet dan menerapkan analisis komponen utama untuk mengoptimalkan jumlah cluster. Kesimpulannya menunjukkan bahwa segmentasi pasar berfungsi sebagai alat yang efektif untuk membedakan produk dan konsumen dengan karakteristik yang berbeda dan membantu memberikan saran kepada bisnis. Selain itu, memasukkan fitur sikap ke dalam proses analisis akan menghasilkan profil pelanggan yang lebih baik.

Clustering and Classification Based on Distributed Automatic Feature Engineering for Customer Segmentation by Lee Jung Zne

Untuk mengalahkan persaingan dan mendapatkan informasi berharga, pengambil keputusan harus melakukan pembelajaran mesin atau penambangan data yang mendalam untuk analitik data. Secara tradisional, pengelompokan dan klasifikasi adalah dua metode umum yang digunakan dalam penambangan mesin. Untuk clustering, data dibagi menjadi berbagai kelompok sesuai dengan kesamaan atau kesamaan fitur. Di sisi lain, klasifikasi mengacu pada membangun model dengan data pelatihan yang diberikan, di mana kelas atau label target diprediksi untuk data uji. Dalam beberapa tahun terakhir, banyak peneliti fokus pada hybrid clustering dan klasifikasi. Teknik-teknik ini memiliki prestasi yang mengagumkan, tetapi masih ada ruang untuk memperbaiki kinerja, seperti proses terdistribusi. Oleh karena itu, kami mengusulkan pengelompokan dan klasifikasi berdasarkan rekayasa fitur otomatis terdistribusi (AFE) untuk segmentasi pelanggan dalam makalah ini. Dalam algoritme yang diusulkan, AFE menggunakan koloni lebah buatan (ABC) untuk memilih fitur berharga dari data masukan, dan kemudian RFM menyediakan analisis data dasar. Dalam AFE, pertama-tama menginisialisasi jumlah cluster k. Selain itu, metode pengelompokan k-means, metode Wald, dan fuzzy c-means (FCM) diproses untuk mengelompokkan contoh dalam kelompok varian. Akhirnya, metode klasifikasi dari pohon keputusan fuzzy yang ditingkatkan mengklasifikasikan data target dan menghasilkan aturan keputusan untuk menjelaskan situasi detail. AFE juga menentukan nilai bilangan split pada pohon keputusan fuzzy yang ditingkatkan untuk meningkatkan akurasi klasifikasi. Pengelompokan dan klasifikasi yang diusulkan berdasarkan rekayasa fitur otomatis didistribusikan, dilakukan di platform Apache Spark. Topik makalah ini adalah tentang pemecahan masalah pengelompokan dan klasifikasi untuk pembelajaran mesin. Dari hasil, akurasi klasifikasi yang sesuai mengungguli pendekatan lain. Selain itu, kami juga menyediakan strategi dan aturan keputusan yang berguna dari analitik data untuk pengambil keputusan.

Dalam pekerjaan terkait, RFM dan clustering digunakan untuk segmentasi pelanggan. RFM berarti keterkinian, frekuensi, dan nilai moneter untuk analisis pemasaran [6-8]. Ini dapat menganalisis informasi penting untuk jumlah pembelian, volume penjualan, dan perilaku pelanggan, dan ini adalah metode praproses yang berguna untuk analisis data. Untuk metode clustering adalah proses menggabungkan dataset ke dalam segmentasi yang berbeda sesuai dengan kesamaannya. Ini adalah algoritma yang banyak digunakan dalam analisis pemasaran [9-12]. Baru-baru ini, integrasi RFM dan clustering memberikan informasi yang lebih berguna untuk segmentasi pelanggan [13]. Menggabungkan clustering dan RFM, dapat mengadopsi strategi yang berbeda untuk berbagai pelanggan sesuai dengan hasil clustering [14-18]. Hibrida clustering dan RFM memang mampu membagi pelanggan dengan baik, namun kurang memiliki sarana interpretasi dan verifikasi [19]. Metode clustering juga perlu menebak jumlah awal cluster. Meskipun beberapa metode untuk menebak jumlah awal cluster telah dikembangkan, mereka tidak dapat secara jelas mencerminkan perbedaan antara pelanggan varian dan sulit untuk digunakan dalam aplikasi nyata [20]. Metode klasifikasi, seperti pohon keputusan, dapat menemukan aturan untuk pengambilan keputusan [21]. Pohon klasifikasi dan regresi (CART) menggunakan indeks Gini untuk membuat aturan seperti pohon bagi pengambil keputusan, dan dapat memberikan informasi yang berguna. Namun, CART dapat memperbaiki akurasi klasifikasi dengan meningkatkan indeks Gini [22,23].

Untuk mengatasi kelemahan di atas, AFE adalah cara baru untuk mengembangkan strategi untuk memilih fitur subset yang berguna. Itu dapat menemukan fitur-fitur berguna ini dalam kumpulan data untuk menghasilkan serangkaian fitur baru untuk pengelompokan, dan dapat meningkatkan kinerja dengan menghapus fitur yang tidak relevan dan fitur yang berlebihan [24]. Selain itu, dapat menginisialisasi jumlah cluster untuk clustering dan menyesuaikan nilai parameter untuk klasifikasi. Dalam makalah ini, kami mengusulkan masalah pengelompokan dan klasifikasi terbaru berdasarkan rekayasa fitur otomatis terdistribusi untuk segmentasi pelanggan. Dalam algoritma yang diusulkan, AFE menggunakan ABC untuk memilih fitur yang berharga dan menginisialisasi jumlah cluster. RFM menyediakan analisis data dasar. Metode clustering k-means, metode Wald, dan FCM diproses untuk segmentasi pelanggan. Metode klasifikasi dari pohon keputusan fuzzy yang ditingkatkan menghasilkan akurasi klasifikasi dan aturan keputusan. AFE juga menentukan nilai bilangan split pada pohon keputusan fuzzy yang ditingkatkan untuk meningkatkan akurasi klasifikasi. Pengelompokan dan klasifikasi yang diusulkan berdasarkan rekayasa fitur otomatis untuk segmentasi pelanggan didistribusikan dilakukan di platform Apache Spark. Ini didedikasikan untuk meningkatkan kinerja pembelajaran mesin dalam ilmu komputer dan teknik dan simetri. Kontribusi dari makalah ini adalah sebagai berikut:

Menggunakan teknologi clustering dari metode Wald, k-means, dan FCM, yang menyediakan

pengelompokan yang lebih baik untuk segmentasi pelanggan.

Algoritma klasifikasi yang diusulkan dari pohon keputusan fuzzy yang ditingkatkan dapat memperbaiki

akurasi klasifikasi dan menyediakan aturan keputusan untuk pengambil keputusan.

Algoritma yang diusulkan mengintegrasikan manfaat simetri dari pengelompokan dan klasifikasi

untuk pembelajaran mesin.

AFE digunakan untuk memilih fitur yang berguna dan menyesuaikan jumlah cluster dan

nilai parameter dari pohon keputusan fuzzy yang ditingkatkan.

Apache Spark menyediakan lingkungan terdistribusi untuk metode di atas untuk

meningkatkan kinerja.

Kami menyediakan beberapa strategi dan aturan keputusan yang berguna bagi para pembuat keputusan.

Kmeans

k-means adalah salah satu teknik pembelajaran tanpa pengawasan yang terkenal untuk analisis klaster. Analisis klaster digunakan untuk mengagregasi atau membagi kumpulan data menjadi beberapa klaster menurut nilai kesamaan. Untuk k-means perlu ditentukan terlebih dahulu jumlah cluster (k). Ini dimulai dengan centroid yang dihasilkan secara acak dan secara iteratif menghitung centroid baru untuk konvergen ke cluster terakhir. Ada empat langkah dalam k-means [26].

Langkah 1: Posisi k centroid dibangkitkan secara acak.

Langkah 2: Setiap titik data dalam kumpulan data akan ditetapkan ke centroid terdekat, dan kemudian centroid baru dihasilkan.

Langkah 3: Untuk menghitung ulang cluster baru dengan menetapkan semua titik data ke yang terdekat centroid, dan kemudian cluster baru dibuat.

Langkah 4: Proses akan diulang antara langkah 2 dan langkah 3 sampai kriteria berhenti telah terpenuhi.

Metode Ward menggunakan varians dalam-cluster minimum untuk menghasilkan cluster. Metode ini menggunakan error sum of squares (ESS) untuk menghitung jarak dari titik data ke kluster ke-m.

Clustering optimization in RFM analysis based on k-means by Rendra Gustriansyah

RFM adalah singkatan dari Recency, Frequency, dan Monetary. RFM adalah metode sederhana namun efektif yang dapat diterapkan pada segmentasi pasar. Analisis RFM digunakan untuk menganalisis perilaku pelanggan yang terdiri dari seberapa baru pelanggan melakukan pembelian (recency), seberapa sering pelanggan melakukan pembelian (frekuensi), dan berapa banyak uang yang dibelanjakan pelanggan (moneter). Dalam penelitian ini, analisis RFM telah digunakan untuk segmentasi produk yang akan disusun dalam hal penjualan terakhir (R), penjualan yang sering (F), dan total uang yang dikeluarkan (M) dengan menggunakan metode data mining. Penelitian ini mengusulkan prosedur baru untuk analisis RFM (dalam segmentasi produk) menggunakan metode k-Means dan delapan indeks validitas untuk menentukan jumlah cluster yang optimal yaitu Metode Elbow, Indeks Silhouette, Indeks Calinski-Harabasz, Indeks Davies-Bouldin, Ratkowski Index, Hubert Index, Ball-Hall Index, dan Krzanowski-Lai Index, yang dapat meningkatkan objektivitas dan kesamaan data dalam segmentasi produk sehingga dapat meningkatkan akurasi proses manajemen stok. Hasil evaluasi menunjukkan bahwa jumlah cluster yang optimal untuk metode k-Means yang diterapkan pada analisis RFM terdiri dari tiga cluster (segmentasi) dengan nilai varians sebesar 0,19113.

Clustering atau segmentasi data adalah suatu proses pengelompokan (partitioning) kumpulan data yang besar menjadi kelompok-kelompok (partisi) sesuai dengan kesamaannya. Ketika jumlah transaksi menjadi lebih besar, proses pengelolaan database produk untuk manajemen stok bukanlah tugas yang mudah. Masalah ini dapat diatasi dengan pendekatan yang lebih baik dengan menggunakan metode data mining yang diperlukan untuk mengelompokkan semua produk ke dalam jumlah cluster yang tepat sesuai dengan beberapa kesamaannya. Nilai dari berbagai kelompok kemudian dapat diperkirakan dan dievaluasi untuk memberikan keputusan yang berguna bagi manajemen dalam memanfaatkan sumber daya secara rasional.

Salah satu model sederhana namun efektif yang dapat diterapkan pada segmentasi produk berdasarkan kesamaan atribut data dengan memeriksa kapan (recency), seberapa sering (frekuensi), dan uang yang dikeluarkan (moneter) pada barang atau jasa tertentu adalah recency (R), frekuensi ( F), dan moneter (M) model [1, 2]. Penelitian ini akan mengelompokkan data produk menggunakan salah satu metode data mining yaitu metode k-Means [3, 4] yang akan diusulkan untuk analisis RFM. Penggunaan metode k-Means untuk pengelompokan produk berdasarkan nilai RFM diharapkan memiliki akurasi yang lebih baik dibandingkan dengan pengelompokan produk secara manual [5-7].

Sedangkan penentuan jumlah k cluster yang optimal pada metode k-Means akan dievaluasi menggunakan delapan indeks validitas yaitu Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball- Hall Index, dan Krzanowski Index -Lai, yang diharapkan dapat meningkatkan objektivitas dan akurasi dalam segmentasi produk dibandingkan dengan hanya menggunakan satu metode [1, 2, 8], serta dapat mempermudah proses manajemen stok [9].

K-Means [14] merupakan salah satu metode non-hierarchical clustering data yang mempartisi data n ke dalam cluster k, sehingga intra-cluster similarity yang dihasilkan tinggi (minimal within-clusters sum of squares), sedangkan inter-cluster similarity adalah rendah (jumlah kuadrat maksimum antar-cluster). K-Means merupakan salah satu metode clustering yang paling populer, karena kesederhanaan algoritma dan kecepatan pemilihan pusat cluster (centroid).

Metode k-Means sering menerapkan rumus jarak Euclidean untuk menentukan kesamaan data dalam suatu cluster secara iteratif.

Langkah-langkah clustering data menggunakan metode k-Means dapat dilakukan dengan cara:

a) Tentukan jumlah cluster k;

b) Inisialisasi nilai k sebagai pusat cluster (centroid) secara acak;

c) Mengelompokkan setiap data ke dalam cluster terdekat. Kedekatan dua data dihitung menggunakan jarak Euclidean;

d) Hitung ulang setiap centroid dengan menghitung rata-rata semua data centroid dengan anggota cluster saat ini;

e) Pengelompokan ulang setiap data (kembali ke langkah 3) menggunakan semua centroid baru sampai semua centroid tidak berubah lagi;

f) Jika centroid tidak berubah lagi, proses clustering selesai.

Salah satu masalah utama dari metode k-Means adalah bagaimana menentukan jumlah cluster k yang optimal. Penelitian oleh Subbalakshmi et al. [15] telah menunjukkan bahwa akurasi metode k-Means bisa lebih tinggi, jika sesuai dalam memilih nilai awal dan jumlah cluster [2, 13].

Ada berbagai cara yang dapat digunakan untuk mengestimasi jumlah cluster yang optimal k. Dalam penelitian ini, jumlah cluster k yang optimal akan diukur menggunakan Metode Elbow, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall Index, dan Krzanowski-Lai Index.

Validity Index for Determining the Optimal Number of Clusters in the k-Means Method

Metode Elbow: Metode Elbow (EM) [16] merupakan metode yang digunakan untuk menentukan jumlah cluster yang optimal, dengan melihat persentase perbandingan antara jumlah cluster yang akan membentuk sudut pada kurva. Jika nilai cluster pertama dengan nilai cluster kedua membentuk sudut (siku) pada kurva atau nilai penurunan terbesar, nilai cluster adalah nilai cluster terbaik. Jumlah cluster 'k' terbaik akan dipilih pada titik tersebut (titik balik). Metode ini merupakan metode visual yang melihat total variasi intra-cluster atau total Within-Clusters Sum of Squares (WSS) sebagai fungsi dari jumlah cluster. Semakin besar jumlah cluster k maka nilai AMPL akan semakin kecil atau sebaliknya.

Silhouette Index: Nilai Silhouette Index (SI) digunakan untuk mengukur seberapa baik cluster berada pada suatu titik tertentu [17]. Rousseeuw [18] mengusulkan pendekatan yang menghitung nilai indeks maksimum. Siluet mengacu pada metode interpretasi dan validasi konsistensi dalam klaster data.

Calinski-Harabasz Index: Calinski-Harabasz Index (CHI) [19] mengevaluasi validitas cluster berdasarkan perhitungan Between-Clusters Sum of Square (BSS) dan WSS. CHI mengukur rasio pemisahan berdasarkan jarak maksimum antara centroid dan mengukur kekompakan berdasarkan jumlah jarak antara setiap data dengan centroid. Konfigurasi cluster yang kompak dan terpisah dengan baik diharapkan memiliki varians antar-cluster yang tinggi dan varians intra-cluster yang relatif rendah [20, 21].

Davies-Bouldin Index: Davies-Bouldin Index (DBI) [22] adalah salah satu metode yang digunakan untuk mengukur validitas cluster dalam metode pengelompokan, kohesi didefinisikan sebagai jumlah kedekatan data dengan titik pusat cluster dari cluster yang diikuti. Sedangkan pemisahan didasarkan pada jarak antara titik pusat cluster dengan cluster.

Pengukuran menggunakan DBI akan memaksimalkan jarak antar cluster antara cluster ci dan cj dan pada saat yang sama akan meminimalkan jarak antar data dalam cluster. Jika jarak antar klaster maksimal berarti kesamaan karakteristik antar klaster kecil sehingga perbedaan antar klaster dapat terlihat lebih jelas. Jika jarak intra-cluster minimal, berarti setiap objek dalam cluster memiliki tingkat kemiripan karakteristik yang tinggi.

Ratkowsky-Lance Index: Indeks Ratkowsky-Lance [23] didasarkan pada rata-rata antara jumlah kuadrat antara cluster kemudian untuk setiap data (BGSS) dan jumlah total kuadrat setiap data dalam cluster (TSS).

Cluster Kualitas cluster yang dihasilkan akan diuji dengan mengevaluasi nilai varians (R). R adalah nilai perbandingan antara jarak rata-rata data pada cluster yang sama (intra-cluster distance) dan jarak rata-rata data pada cluster yang lain (inter-cluster distance) [26]. Nilai R mendekati 0 menunjukkan bahwa data dalam cluster yang sama lebih mirip.Quality Testing

CONSUMER SEGMENTATION IN FOOD RETAILING IN CROATIA: A LATENT CLASS ANALYSIS by Pavlic

Tujuan – Untuk mengantisipasi kepuasan dengan retail toko dan penawarannya, sangat penting untuk mempertimbangkan apakah toko memenuhi harapan fragmen konsumen yang berbeda

berkaitan dengan atribut toko. Tujuan utama dari makalah ini adalah untuk mengidentifikasi segmen konsumen yang berbeda sesuai dengan tingkat kepuasannya terhadap berbagai elemen toko ritel dan untuk menentukan profil masing-masing segmen.

Desain/Metodologi/Pendekatan – Sebuah studi empiris dilakukan pada 500 konsumen Kroasia, menerapkan sampel kenyamanan melalui kuesioner yang sangat terstruktur. Menggunakan analisis faktor, pernyataan dibagi menjadi lima dimensi yang bermakna (bermacam-macam produk, harga, atmosfir toko, kualitas toko retail dan retail kualitas layanan). Faktor-faktor ini digunakan sebagai indicator untuk analisis kelas laten.

Temuan dan implikasi – Temuan telah mengindikasikan solusi optimal dari lima segmen konsumen dengan tingkat kepuasan yang berbeda. Masing-masing teridentifikasi segmen (berorientasi pada kenyamanan, tidak tertarik, bermacam- berorientasi, harga-sensitif dan konsumen abu-abu) dijelaskan dalam hal sosio-demografis dan perilaku variabel. Penelitian ini memberikan wawasan tentang faktor-faktor yang mempengaruhi kepuasan pelanggan. Ini juga membantu ritel manajer mengembangkan strategi yang tepat untuk didekati segmen konsumen yang berbeda dan meningkatkan kepuasan mereka.

Keterbatasan – Penelitian ini dilakukan dalam satu Kabupaten Kroasia, yang dapat membatasi generalisasi temuan. Batasan lain mengacu pada penggunaan factor hasil alih-alih daftar asli item dalam analisis dari segmen.

Orisinalitas – Studi ini berkontribusi pada literatur yang ada dengan memberikan wawasan yang berguna ke konsumen yang berbeda segmen berdasarkan tingkat kepuasan mereka dengan ritel konteks toko dalam ritel makanan Kroasia. Petunjuk praktis untuk menargetkan segmen konsumen tertentu secara lebih efisien diusulkan. Sejauh yang kami tahu, tidak ada penelitian lain telah menggabungkan segmentasi konsumen dan laten analisis kelas dalam konteks pasar ritel Kroasia.

Customer Behavior Mining Framework (CBMF) using clustering and classification techniques by Abdi Farshid

Studi ini mengusulkan Kerangka Kerja Penambangan Perilaku Pelanggan berdasarkan teknik penambangan data di perusahaan telekomunikasi. Kerangka kerja ini memperhitungkan pola perilaku pelanggan dan memprediksi cara mereka bertindak di masa depan. Pertama, teknik clustering digunakan untuk mengimplementasikan analisis portofolio dan pelanggan sebelumnya dibagi berdasarkan fitur sosio-demografis menggunakan algoritma k-means. Kemudian, analisis klaster dilakukan berdasarkan dua kriteria, yaitu jumlah jam layanan telekomunikasi yang digunakan dan jumlah layanan yang dipilih oleh pelanggan masing-masing grup. Enam kelompok pelanggan diidentifikasi dalam tiga tingkat daya tarik menurut hasil analisis portofolio pelanggan. Fase kedua telah dikhususkan untuk menambang perilaku pelanggan di masa depan. Memprediksi tingkat daya tarik pelanggan pendatang baru dan juga perilaku churn pelanggan ini dilakukan pada tahap kedua. Kerangka kerja ini secara efektif membantu manajer telekomunikasi menambang perilaku pelanggan mereka. Hal ini dapat menyebabkan pengembangan taktik yang tepat sesuai dengan daya tarik pelanggan dan perilaku terburu-buru mereka. Peningkatan kemampuan manajer dalam manajemen hubungan pelanggan adalah salah satu hasil penelitian yang diperoleh.

Manajemen hubungan pelanggan (CRM) telah berubah menjadi bidang yang signifikan dalam bisnis telekomunikasi. CRM diasumsikan sebagai aset tidak berwujud yang memberikan potensi persaingan bagi organisasi (Ryals 2002). Jadi, perusahaan mencoba mengidentifikasi dan menganalisis perilaku pelanggan mereka. Peningkatan teknologi telah memungkinkan perusahaan telekomunikasi untuk menyimpan catatan pelanggan. Menganalisis data historis membantu perusahaan untuk menemukan pola perilaku pelanggan yang ada yang dapat menghasilkan dampak signifikan dalam memprediksi perilaku pelanggan di masa depan.

Segmentasi pelanggan adalah penggunaan data masa lalu untuk membagi pelanggan ke dalam kelompok serupa berdasarkan berbagai fitur (Hsu et al. 2012). Dengan menggunakan proses segmentasi pelanggan, perusahaan akan dapat mengidentifikasi pelanggan yang secara strategis penting dan menguntungkan. Pelanggan ini dapat dikategorikan ke dalam dua kelas utama termasuk pelanggan nilai masa depan yang tinggi atau pelanggan volume tinggi (Buttle dan Maklan 2015).

Segmentasi pelanggan membagi pelanggan ke dalam kelompok-kelompok dengan karakteristik, persyaratan, dan perilaku yang serupa. Segmentasi pelanggan dapat dilakukan berdasarkan kelas fitur yang berbeda seperti atribut pengguna dan atribut penggunaan. Atribut pengguna termasuk atribut demografis (yaitu, usia, jenis kelamin, status pekerjaan, dan status perkawinan), atribut geografis (yaitu, negara, wilayah) dan atribut psikografis (yaitu, gaya hidup). Atribut penggunaan berisi informasi tentang frekuensi dan tingkat pembelian dan perilaku pembeli (Mohammadi dkk. 2013; Buttle, dan Maklan 2015). Pengelompokan pelanggan digunakan untuk membangun profil pelanggan yang merupakan inti dari sistem informasi yang berpusat pada pelanggan (Bose dan Chen 2015). Beberapa penulis menggunakan berbagai kriteria segmentasi dan berbagai teknik pengelompokan untuk mengelompokkan pelanggan. Ringkasan studi sebelumnya tentang segmentasi pelanggan disorot pada Tabel 2. Dapat disimpulkan dari Tabel 2 bahwa berbagai metode pengelompokan digunakan untuk mengelompokkan pelanggan dalam berbagai bidang termasuk perbankan, telekomunikasi, restoran, manufaktur tekstil, rumah sakit dan pariwisata. Pada penelitian kali ini, segmentasi pelanggan pertama-tama dilakukan berdasarkan fitur sosio demografis pelanggan dan kemudian diimplementasikan analisis klaster berdasarkan fitur perilaku pelanggan.

A clustering approach to domestic electricity load profile characterisation using smart metering data by Fintan McLoughlin.

Tersedianya peningkatan jumlah data ke utilitas listrik melalui implementasi kampanye pengukuran pintar domestik berarti bahwa cara tradisional menganalisis informasi pembacaan meter seperti statistik deskriptif menjadi semakin sulit. Informasi karakteristik utama untuk ata sering hilang, terutama ketika proses rata-rata atau agregasi diterapkan. Oleh karena itu, lainnya metode analisis data perlu digunakan agar informasi ini tidak hilang. Salah satu metode yang cocok untuk menganalisis sejumlah besar informasi adalah data mining. Ini memungkinkan data tersegmentasi sebelum proses agregasi tersebut diterapkan. Selain itu, segmentasi memungkinkan untuk dimensi reduksi sehingga memudahkan manipulasi data. Metode clustering telah digunakan dalam industri kelistrikan selama beberapa waktu. Namun, penggunaannya di tingkat domestik agak terbatas sampai saat ini. Makalah ini menyelidiki tiga yang paling banyak digunakan metode pengelompokan tanpa pengawasan: k-means, k-medoid dan Self Organizing Maps (SOM). Performa terbaik Teknik ini kemudian dievaluasi untuk mengelompokkan rumah tangga individu ke dalam klaster berdasarkan pola penggunaan listrik mereka sepanjang hari. Proses ini diulang untuk setiap hari selama periode enam bulan untuk mengkarakterisasi variasi harian, intra-harian dan musiman dari permintaan listrik domestik. Berdasarkan hasil ini, serangkaian Kelas Profil (PC) disajikan yang mewakili pola umum penggunaan listrik di dalam rumah. Akhirnya, setiap PC dikaitkan dengan karakteristik rumah tangga dengan menerapkan sebuah regresi logistik multi-nominal ke data. Akibatnya, rumah tangga dan cara mereka penggunaan listrik di rumah dapat dicirikan berdasarkan atribut pelanggan individu.

Di seluruh Uni Eropa, telah terjadi gerakan menuju jaringan listrik yang lebih cerdas, di mana peningkatan visibilitas atas pembangkitan dan konsumsi listrik telah dicapai dengan pemasangan Infrastruktur Pengukuran Tingkat Lanjut (AMI). Pengukuran cerdas adalah bagian dari ini dan dipandang sebagai komponen yang diperlukan untuk mencapai sasaran kebijakan energi Uni Eropa 20-20-20 pada tahun 2020: untuk mengurangi emisi gas rumah kaca sebesar 20%, untuk meningkatkan efisiensi energi sebesar 20% dan untuk 20% dari Permintaan energi UE berasal dari energi terbarukan

sumber energi [1].

Dalam beberapa tahun terakhir, instalasi smart meter telah meningkat di seluruh dunia dalam upaya untuk memodernisasi jaringan listrik yang sudah tua [2]. Selanjutnya, perbaikan dalam lingkungan peraturan, khususnya di sektor perumahan di Eropa telah menghasilkan sejumlah program percontohan smart metering [3]. Akibatnya, ada banyak data baru untuk utilitas, memberikan konsumsi listrik terperinci dengan granularitas yang meningkat untuk sejumlah besar pelanggan di sektor perumahan [4]. Ketersediaan sumber data ini berpotensi dapat digunakan oleh utilitas untuk membuat Kelas Profil (PC) beban listrik yang disesuaikan dan dapat membantu di bidang-bidang seperti: perencanaan dan prakiraan beban yang lebih baik; desain tarif Time of Use (ToU); penyelesaian listrik; dan strategi Demand Side Management (DSM) [5].

Makalah ini menyajikan metodologi baru untuk profil beban listrik karakterisasi. Dengan demikian, serangkaian PC listrik rumah tangga dibangun yang mencerminkan berbagai cara penggunaan listrik di dalam rumah. Saat ini, PC diturunkan berdasarkan menggabungkan banyak pola penggunaan listrik yang berbeda secara bersama-sama [6]. Penerapan pendekatan jenis ini, di mana rumah tangga individu yang mungkin menggunakan listrik dengan cara yang sangat berbeda disatukan, menghasilkan pembentukan profil beban yang sangat agregat. Namun, pada kenyataannya ini bukanlah cerminan yang sebenarnya tentang bagaimana listrik sebenarnya dikonsumsi dan yang dapat berubah secara signifikan di antara pelanggan yang berbeda [7]. Makalah ini mengusulkan metode alternatif yang menggunakan pengelompokan untuk mengidentifikasi pola penggunaan listrik yang serupa sebelum proses agregasi diterapkan. Dengan cara ini, informasi yang berkaitan dengan bentuk profil beban listrik tidak hilang. Selain itu, makalah ini juga menyajikan metode menghubungkan PC ke pelanggan individu sehingga rumah tangga dan cara mereka menggunakan listrik di dalam rumah dapat dicirikan berdasarkan atribut pelanggan masing-masing.

Domestic electricity load profile characterisation

Berdasarkan literatur, metode yang ada digunakan untuk mengkarakterisasi

penggunaan listrik rumah tangga secara umum dapat dibagi menjadi empat kategori: statistik; rekayasa; deret waktu dan pengelompokan. Metode statistik telah banyak digunakan di pasar listrik yang tidak diatur untuk membentuk PC beban standar [6].

PC beban standar digunakan untuk tujuan penyelesaian dan memberikan perkiraan jumlah dan Waktu Penggunaan (ToU) listrik yang digunakan. Serangkaian PC diproduksi untuk segmen pasar yang berbeda (misalnya perumahan, komersial, industri) dan diturunkan berdasarkan rata-rata untuk semua pelanggan yang terdapat dalam satu kelas pelanggan [8]. Pasar listrik Inggris memiliki dua PC domestik; Tidak Terbatas dan Ekonomi 7.

Di Irlandia, ada empat PC untuk sektor domestik; 24 jam dan Night Saver yang dipisahkan oleh batas kota dan desa [9]. Meskipun PC cocok untuk keperluan pemukiman, pada kenyataannya mereka tidak mencerminkan bagaimana listrik sebenarnya dikonsumsi di dalam rumah setiap hari dan hanya mewakili rata-rata untuk semua pelanggan yang berada dalam kelas yang sama.

Teknik statistik lainnya terdiri dari penggunaan statistik deskriptif dan probabilitas [10–16] serta regresi [17–22] untuk menggambarkan penggunaan listrik di dalam rumah. Serupa dengan yang disebutkan di atas, metode ini menghasilkan bentuk profil beban yang sangat beragam, hasil dari menggabungkan banyak pola penggunaan listrik yang berbeda secara bersamaan [10].

Pendekatan teknik untuk karakterisasi profil beban domestik bervariasi tetapi umumnya mencirikan penggunaan listrik sebagai fungsi parameter seperti hunian atau kepemilikan alat [23-28]. Metode-metode ini dianggap sebagai pendekatan dari bawah ke atas di mana banyak profil dibangun untuk rumah tangga yang berbeda dan oleh karena itu tidak mengalami masalah yang sama yang disoroti di atas untuk pendekatan statistik. Namun, metode rekayasa sulit untuk digeneralisasi dan membutuhkan pengetahuan rinci tentang penghuni rumah tangga dan Penggunaan Waktu (TU) alat [29].

Sebaliknya, pendekatan deret waktu terbatas dalam penerapannya pada rumah tangga domestik, tetapi hal ini kemungkinan besar disebabkan oleh kurangnya data historis yang tersedia untuk sektor tersebut [7]. Metode telah digunakan secara luas untuk menggambarkan penggunaan listrik pada tingkat Operator Sistem Transmisi (TSO) [30-34]. Namun, pendekatan ini mengalami masalah yang serupa dengan yang disorot di atas untuk teknik statistik ketika banyak profil yang berbeda digabungkan bersama sehingga menghasilkan bentuk profil beban listrik yang beragam [35].

Akhirnya teknik penambangan data seperti analisis klaster telah digunakan untuk mengelompokkan pelanggan yang menunjukkan perilaku listrik serupa melalui data meteran pintar ToU, tetapi sebagian besar telah diterapkan pada tingkat agregat [36-38]. Selanjutnya, pelanggan juga telah dikelompokkan berdasarkan nilai parameter agregat seperti penggunaan listrik tahunan atau fitur yang berkaitan dengan bentuk profil beban listrik (misalnya faktor beban) [39,40]. Demikian pula, profil beban telah dibangun untuk pelanggan perumahan komersial, industri dan sebagian besar agregat berdasarkan metode pengelompokan: Self Organizing Maps (SOM), k-means dan Follow the Leader [41-43]. Secara khusus, satu penelitian besar terhadap sekitar 3000 pelanggan perumahan dipantau selama satu tahun dan menggunakan metode: SOM; k-berarti; dan hierarkis untuk mengelompokkan dan membangun profil beban [44].

Namun, analisis dibatasi hanya sebagian kecil dari rangkaian waktu (5%) karena tuntutan komputasi. Metode pengelompokan tidak mengalami banyak masalah yang disorot di atas terutama ketika diterapkan sebelum melakukan analisis statistik apa pun. Lebih jauh lagi dengan perbaikan dalam tugas perangkat keras komputer seperti pengelompokan, yang dapat menjadi komputasi intensif menjadi lebih mudah untuk diimplementasikan.

Makalah ini mengisi celah dalam literatur dengan mengelompokkan berdasarkan ToU untuk sampel besar pelanggan perumahan selama periode enam bulan. Hal ini memungkinkan beban PC diturunkan berdasarkan pola penggunaan listrik individu di dalam rumah selama periode ini dan tidak mengalami beberapa masalah agregasi yang sama yang disorot di atas. Selanjutnya, karena seluruh kumpulan data dikelompokkan, pola diurnal, intra-harian dan musiman untuk penggunaan listrik dapat dicirikan, serta di antara variasi pelanggan. Selain itu, karena karakteristik hunian, penghuni, dan peralatan berkorelasi dengan masing-masing PC, ia juga menyediakan metode untuk menetapkan pola penggunaan listrik kepada pelanggan individu. Akhirnya, karena ukuran sampel relatif besar, PC dapat dianggap mewakili populasi yang lebih luas di Irlandia. Metode serupa juga dapat digunakan di pasar listrik lain di luar Irlandia.

K-LRFMD: Method of Customer Value Segmentation in Shared Transportation Filed Based on Improved K-means Algorithm by Hong Li , Xiaosheng Yang

Munculnya era informasi telah mengubah fokus pemasaran perusahaan dari produk-sentris ke customer-centric, dan manajemen hubungan pelanggan menjadi masalah inti perusahaan. Hasil klasifikasi nilai pelanggan yang akurat merupakan dasar penting bagi perusahaan untuk mengoptimalkan alokasi sumber daya pemasaran, dan klasifikasi nilai pelanggan menjadi salah satu masalah utama yang perlu segera dipecahkan dalam manajemen hubungan pelanggan. Dalam menghadapi persaingan pasar yang ketat dari industri berbagi kendaraan, setiap perusahaan transportasi bersama telah memperkenalkan lebih banyak metode pemasaran preferensial untuk menarik lebih banyak pelanggan. Dalam makalah ini, dengan bantuan platform berbagi kendaraan di kampus universitas domestik, kami membuat model evaluasi nilai pelanggan yang masuk akal yang disebut K-LRFMD. K-LRFMD melakukan beberapa analisis pengelompokan dengan pelanggan berdasarkan rekayasa fitur spesifik dan algoritma K-means yang ditingkatkan. Dalam makalah ini, kami membandingkan nilai pelanggan yang berbeda yang berasal dari K-LRFMD. Analisis dapat merumuskan strategi pemasaran yang sesuai untuk menyediakan layanan pelanggan yang dipersonalisasi untuk pelanggan yang berbeda.

The Analyze of Relationship between Revenue and Customer Payment Methods in Small Medium Enterprise Based on Clustering K-Means

Modal usaha dan pendapatan tidak hanya menentukan kesehatan UKM tetapi juga harus seimbang. Secara umum, pelanggan mendapatkan keuntungan dari metode pembayaran yang fleksibel sementara di sisi lain UKM harus mendapatkan keuntungan mereka juga. Sehingga perlu dikaji apakah UKM perlu mendapatkan keuntungan sesuai dengan keadaan tersebut. Salah satu metode yang cocok diterapkan adalah dengan menerapkan pengelompokan pelanggan berdasarkan pendapatan dan pembayaran yaitu metode clustering K-means karena dapat memunculkan beberapa grup yang belum diketahui sebelumnya. Informasi ini berguna bagi UKM untuk dimanfaatkan sesuai dengan kebutuhannya. Data dalam penelitian ini dikumpulkan dari atribut pelanggan, jumlah transaksi, dan metode pembayaran. Jumlah centroid adalah 3. Hasil pengelompokan dihentikan pada iterasi ke-5. Hasil penelitian menunjukkan bahwa nilai rasio iterasi ke-4 dan iterasi ke-5 memiliki nilai rasio yang sama yaitu 0,07393. Dari hasil iterasi dapat ditemukan; pertama, berdasarkan jumlah pelanggan, kelompok dapat diklasifikasikan menjadi tiga C1 (18%), C2 (45%), C3 (36%). Kedua, berdasarkan rata-rata jumlah transaksi, pembayaran pascabayar menempati urutan pertama (12,7/minggu). Dari hasil tersebut dapat dianalisa bahwa keadaan ini memberatkan UKM karena semakin banyak jumlah transaksi maka semakin banyak investasi yang harus disiapkan untuk piutang.

UKM merupakan aset bagi Indonesia, dan terbukti mampu bertahan dari krisis ekonomi. [1]. Oleh karena itu, penting kiranya bagaimana membantu UKM untuk mengembangkan diri dan menyelesaikan permasalahannya.[2]. Pelanggan merupakan aset bagi Usaha Kecil Menengah (UKM) [3],[4], dimana sebagai upaya perusahaan memberikan kemudahan bagi pelanggan. Fleksibilitas pembayaran merupakan salah satu komponen yang dibutuhkan pelanggan, yang dapat menjadi daya tarik pelanggan terhadap perusahaan. Namun, fleksibilitas pembayaran bagi pelanggan dapat memiliki konsekuensi bagi perusahaan dengan menyediakan lebih banyak modal. Jika akibat diimbangi dengan besarnya laba maka akibat itu seimbang dengan besarnya laba yang diterima perusahaan, maka hal ini tidak menjadi masalah atau sebaliknya. Tentunya perlu dilakukan kajian tentang bagaimana hubungan jumlah transaksi nasabah dengan sistem pembayaran nasabah, sehingga perusahaan dapat mengambil keputusan yang lebih tepat untuk kelangsungan dan perkembangan usahanya. UKM dapat menggunakan database transaksinya untuk menambang dan menganalisis kebutuhan tersebut.[5]. Sebagian besar perusahaan telah menyadari bahwa database pelanggan adalah sumber daya yang sangat penting yang dapat digunakan untuk menganalisis karakteristik pelanggan untuk membentuk strategi pemasaran yang tepat dan untuk mengaturnya.[6]. Segmentasi pelanggan merupakan salah satu kegiatan yang biasa digunakan dengan memanfaatkan data tersebut. Secara umum segmentasi nasabah diartikan sebagai proses pengelompokan nasabah dengan cara dibagi menjadi beberapa kelompok berdasarkan karakteristik dan perilaku transaksinya. [7]. Banyak penelitian telah dilakukan tentang segmentasi pelanggan. setiap penelitian memiliki tujuan yang berbeda. Salah satu penelitian yang menggunakan K-Means menemukan peringkat saham yang penting. [8], dan penelitian lain untuk menemukan segmentasi pelanggan pengguna CRM dan K-Means. [9]. Kajian ini dianalisis berdasarkan RFM (Recency, Frequency, Monetary) untuk mencari calon pelanggan. [10]. Salah satu pendekatannya adalah mengekstrak data menggunakan metode clustering dalam data mining. Data mining berkontribusi besar untuk mengekstraksi pengetahuan dan informasi tersembunyi yang terkandung dalam data yang digunakan oleh peneliti.[11]. Proses data mining pada data UKM dilakukan dengan cara clustering menggunakan algoritma K-means untuk segmentasi pelanggan. [12]. Dalam penelitian [13] telah menggabungkan metode clustering dengan algoritma yang dirakit untuk mensegmentasi pelanggan.

Penemuan pengetahuan dalam Database (KDD) didefinisikan sebagai ekstraksi informasi yang potensial, implisit dan tidak diketahui dari sekumpulan data [14], [15]. Datamining dalam kerangka KDD yang merupakan bagian dari komponen. Algoritma K-means merupakan algoritma untuk pengelompokan data yang hanya berlaku untuk data numerik, sehingga jika data tersebut bukan data numerik harus dilakukan normalisasi.[16]. K-Means efektif untuk mengurangi jarak euclidean dari setiap objek. [17]. Beberapa penelitian telah dilakukan antara lain: Penelitian ini menggunakan metode clustering untuk mengelompokkan segmentasi pelanggan dengan model RFM.[18]. Penelitian lain mencoba menggabungkan metode clustering dengan pohon keputusan untuk segmentasi pelanggan.[19][20]. Pengembangan clustering dengan beberapa atribut juga telah dilakukan pada segmentasi pemasaran berdasarkan greedy heuristic. [21]. Segmentasi pelanggan terhadap analisis pemasaran menggunakan K-Means telah dilakukan dalam penelitian [22]. Penelitian lainnya, mengembangkan pengelompokan segmentasi pelanggan untuk menganalisis niat pelanggan. [23][24].

Kesimpulan

Pendekatan clustering telah mampu menggali data UKM untuk menganalisis kemungkinan-kemungkinan yang dapat dijadikan pertimbangan untuk mengambil keputusan dalam pengembangan dan evaluasi perusahaan. Dari\ hasil iterasi dapat ditemukan; pertama, berdasarkan jumlah pelanggan, kelompok dapat diklasifikasikan menjadi tiga

C1(18%) adalah pembayaran transfer otomatis, C2 (45%) adalah pembayaran setelah tanggal, C3 (36%) adalah pembayaran non-transfer otomatis dan kombinasi. Tiga (3) cluster yang terbentuk menghasilkan temuan transaksi yang paling banyak

kontributornya adalah pelanggan dengan sistem pembayaran post-date. Berdasarkan rata-rata jumlah transaksi, pembayaran lewat tanggal berada di peringkat pertama (12,7/minggu). Temuan ini dapat memberikan saran bagi UKM bahwa banyak transaksi yang dilakukan nasabah juga harus diimbangi dengan ketersediaan modal. Jumlah transaksi terkecil adalah pelanggan dengan sistem auto-transfer artinya pelanggan memiliki kecenderungan kurang tertarik dengan sistem pembayaran di awal, di sisi lain mereka lebih mengandalkan

modal dari UKM. Hal ini dapat menjadi perhatian bagi pengelola UKM untuk menyikapinya. Penentuan jumlah cluster pada penelitian ini masih dilakukan secara manual. Dalam pekerjaan di masa depan, perlu untuk memikirkan

validitas jumlah cluster yang tepat sehingga hasil pengelompokan akan lebih akurat.

Segmentation of Tourist Interest on Tourism Object Categories by Comparing PSO K-Means and DBSCAN Method by Meri Chindyana\*, Lili Ayu Wulandhari

Marketing di perusahaan travel biasanya akan menawarkan promo atau rekomendasi mengenai berbagai kategori objek wisata acak kepada pelanggannya. Promo atau rekomendasi tersebut berisi kategori objek wisata yang sering dikunjungi dan mendapat penilaian yang baik dari banyak pelanggan. Namun karena perusahaan tidak begitu mengetahui dan memahami karakteristik atau minat setiap pelanggan, terkadang beberapa promo tidak sesuai dengan minatnya sehingga tidak tertarik untuk mengambil promo yang ditawarkan. Sudah ada beberapa makalah yang membahas rekomendasi pariwisata, namun hanya fokus pada 1 kategori tempat wisata atau objek wisata. Berdasarkan permasalahan tersebut, skripsi ini dibuat untuk membahas segmentasi minat wisatawan pada kategori objek wisata dengan membandingkan metode PSO K-Means dan metode DBSCAN yaitu tentang rekomendasi paket wisata yang lebih spesifik sesuai rating. Karakteristik atau kesamaan minat antara 1 wisatawan dengan wisatawan lainnya akan dikelompokkan menjadi 1 cluster. Dari setiap cluster yang terbentuk dapat memudahkan perusahaan untuk mengetahui kategori objek wisata apa saja yang diminati atau disukai oleh setiap pelanggan dan dapat menawarkan promo atau rekomendasi paket wisata sesuai minat wisatawan.

Menurut sebuah artikel [1], bepergian ke suatu tempat adalah salah satu hobi yang paling populer bagi banyak orang. Berwisata ke suatu tempat sangat ditunggu-tunggu saat liburan tiba karena bisa menghabiskan waktu luang bersama keluarga atau teman. Berwisata ke suatu tempat bisa dilakukan di luar negeri atau di dalam negeri tergantung dari rencana liburan yang sudah kita buat. Menurut sebuah artikel [2], ada 10 manfaat yang bisa kita dapatkan saat berwisata ke suatu tempat, yaitu menambah pengetahuan tentang suatu tempat, bertemu orang baru, mempelajari budaya baru, mengurangi stres, belajar beradaptasi, menjadi pribadi yang lebih mandiri. dalam segala hal, lebih percaya diri, lebih sabar, melatih keterampilan komunikasi dan belajar bahasa baru. Saat ini banyak alat transportasi yang dapat digunakan untuk berwisata ke suatu tempat, sehingga memudahkan wisatawan untuk melakukan perjalanan.

Menurut sebuah artikel [3], agen pemasaran di sebuah perusahaan wisata biasanya akan menawarkan promosi atau rekomendasi mengenai berbagai kategori objek wisata secara acak kepada pelanggannya. Promo atau rekomendasi tersebut memuat kategori objek wisata yang sering dikunjungi dan mendapat penilaian yang baik dari banyak pelanggan. Namun karena perusahaan tidak begitu mengetahui dan memahami karakteristik atau minat setiap pelanggan, terkadang ada beberapa promo yang tidak sesuai dengan minat mereka, sehingga mereka tidak tertarik untuk mengambil promo yang ditawarkan.

Sudah ada beberapa makalah yang membahas rekomendasi pariwisata, namun hanya fokus pada 1 kategori tempat wisata atau atraksi wisata: Restoran [4, 5], Museum [6-10], Resor Spa [11-13], Pantai [14], Taman [15], Aktivitas Berbasis Alam [16, 17], Hotel [18, 19], Film [20, 21].

Berdasarkan penelitian yang telah dilakukan oleh peneliti sebelumnya diketahui bahwa rekomendasi hanya terdiri dari satu destinasi, biasanya wisatawan mengunjungi beberapa tempat dalam 1 sesi kunjungan tempat wisata. Oleh karena itu, penulis berencana melakukan segmentasi minat wisatawan pada kategori objek wisata dengan membandingkan metode K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), dan menerapkan metode Particle Swarm Optimization (PSO) untuk mengoptimalkan metode clustering. .

Segmentasi ini akan membahas rekomendasi paket wisata yang lebih spesifik menurut rating. Karakter atau kesamaan minat antara 1 wisatawan dengan wisatawan lainnya akan dikelompokkan menjadi 1 cluster. Dari setiap cluster yang dapat dibentuk dapat memudahkan perusahaan untuk melihat kategori objek wisata mana yang diminati atau. Setiap wisatawan dapat menawarkan promo atau rekomendasi paket wisata sesuai minatnya. Misalnya, di cluster 0, 15 wisatawan menyukai pantai, restoran, dan mal. Jadi, promo atau rekomendasi paket wisata yang diberikan perusahaan kepada 15 wisatawan tersebut lebih diarahkan pada 3 kategori objek wisata di atas dan objek wisata lainnya yang tidak perlu ditawarkan.

Tujuan yang ingin dicapai adalah mengembangkan model clustering minat wisatawan menggunakan metode K-Means dan DBSCAN berdasarkan review data rating destinasi wisata, menganalisis kesamaan karakter setiap cluster yang diperoleh terhadap destinasi wisata yang direkomendasikan, dan menentukan optimasi metode yang akan diterapkan pada metode K-Means dalam hal segmentasi minat wisatawan.

Manfaat yang ingin dicapai adalah melihat dan memahami minat kategori objek wisata masing-masing wisatawan sehingga memudahkan pihak pemasaran di perusahaan travel untuk menawarkan paket wisata atau liburan dengan tepat, dapat meningkatkan keuntungan suatu perusahaan karena meningkatkan minat wisatawan dalam menggunakan jasanya, menambah wawasan mengenai metode yang digunakan untuk clustering, menentukan metode optimasi yang tepat untuk metode K-Means, dan dapat menganalisis kesamaan dari setiap cluster yang terbentuk.

Masalah yang akan dibahas dalam makalah ini adalah metode yang digunakan untuk clustering adalah metode Partitional dan Density Based. Metode Partitional and Density Based yang digunakan adalah metode K-Means dan DBSCAN, metode yang digunakan untuk optimasi adalah metode PSO (Particle Swarm Optimization), data ini hanya berisi informasi tentang kategori objek wisata dan tidak dilengkapi dengan nama-nama wisatawan objek-objek di negara-negara Eropa sehingga pemeringkatannya hanya bersifat umum, misalnya keseluruhan restoran di negara-negara Eropa diberi rating rata-rata 3,42 oleh wisatawan A, dan data ini hanya membahas kawasan Eropa.

Oleh karena itu penulis menggunakan metode clustering untuk menentukan kriteria/minat setiap wisatawan untuk dapat menawarkan promo atau rekomendasi paket wisata sesuai minat wisatawan. Karakter atau kesamaan minat antara 1 wisatawan dengan wisatawan lainnya akan dikelompokkan menjadi 1 cluster. Setiap cluster yang dapat dibangun dapat memudahkan perusahaan untuk melihat kategori objek wisata apa saja yang diminati atau dibuat oleh masing-masing wisatawan. Penulis membaca beberapa makalah dan menemukan bahwa metode clustering yang digunakan untuk teknik segmentasi adalah K-Means dan DBSCAN. Oleh karena itu, penulis membandingkan perbandingan kedua metode tersebut dalam hal segmentasi minat wisatawan terhadap kategori objek wisata dilihat mana yang terbaik. Dari beberapa makalah yang penulis baca, masing-masing teknik optimasi memiliki kegunaan yang berbeda-beda. Dalam hal segmentasi minat wisatawan pada kategori objek wisata ini, penulis memilih menggunakan teknik PSO.

Related Litrev

Ada beberapa makalah yang membahas tentang teknik segmentasi. Berikut ringkasan dari masing-masing makalah:

Dengan menggunakan pendekatan design science research, penelitian ini bertujuan untuk merancang dan mengevaluasi metode 'big data analytics' untuk mendukung pengambilan keputusan strategis dalam pengelolaan destinasi pariwisata. Teknik pengelompokan, bernama P-DBSCAN, diterapkan pada data geografis untuk mengidentifikasi area populer yang diminati. Ini memperhitungkan jumlah foto dan jumlah wisatawan, yang memastikan bahwa lokasi yang diidentifikasi sebenarnya memiliki banyak wisatawan yang telah dikunjungi untuk kepentingan tertentu. Keunggulan P-DBSCAN telah ditunjukkan dalam penelitian terbaru untuk mengidentifikasi lokasi yang dikunjungi wisatawan. K-means clustering diterapkan untuk membangun kosakata visual. Kata visual didefinisikan sebagai pusat kelompok, dan nilai k menentukan jumlah kata visual yang tersedia [22].

Sebuah studi komparatif multinasional menyoroti motivasi perjalanan siswa dan tren wisata. Data dianalisis menggunakan dua analisis komponen utama (PCA), kombinasi dari dua metode pengelompokan: Metode Ward, dan metode solusi optimal, metode K-Means. Muncul tujuh klaster berdasarkan atraksi/aktivitas wisata yaitu explorer, soft explorer, wisatawan, newty seeker, evaders, function seeker, dan tourism lovers. Temuan dari penelitian ini menunjukkan bahwa persepsi atraksi/kegiatan wisata berbeda di setiap negara meskipun ada beberapa kesamaan [23].

Data besar dalam penelitian pariwisata: Sebuah tinjauan literatur. Mengenai metode berbasis centroid, K-Means adalah algoritma yang populer dalam penelitian pariwisata. Metode partisi berbasis centroid seperti itu memiliki satu kelemahan nyata yang membutuhkan pencarian mendalam untuk menemukan pusat cluster terbaik, yang tidak efisien dalam lingkungan yang kaya data; sebaliknya, pengelompokan berbasis kepadatan mungkin lebih cocok untuk foto data besar, yang membutuhkan pengetahuan domain minimum dan secara efektif menyaring penjahat bahkan di hadapan titik kebisingan. Oleh karena itu, klasterisasi aplikasi dengan noise (DBSCAN) berbasis densitas spasial dan variannya telah dikembangkan dan diterapkan secara luas pada pengelompokan foto-foto pariwisata. Untuk membangun hierarki cluster, cluster berbasis konektivitas (juga dikenal sebagai Hierarchical Clustering), algoritma yang fleksibel dan cepat berdasarkan matriks ketidaksetaraan selain dari ruang metrik juga telah diperkenalkan untuk mendeteksi atraksi wisata dan mengelompokkan kota yang dikunjungi oleh wisatawan yang sama [24]. analisis wisatawan yang berpikiran nilai: Wisata berbasis alam di Arktik. Analisis klaster dilakukan untuk mengklasifikasikan wisatawan yang berpartisipasi menggunakan pengelompokan hierarkis dengan kuadrat jarak Euclidean dan hubungan Ward. Ini mengungkapkan solusi tiga cluster untuk interpretasi terbaik, dan prosedur algoritma cluster K-means digunakan untuk mengklasifikasikan tiga segmen pariwisata berdasarkan dimensi nilai yang dirasakan [17].

Pendekatan jaringan untuk segmentasi wisata melalui konten yang dibuat pengguna. Untuk mengurangi dimensi, mereka menggabungkan kata kunci yang cenderung muncul bersama dengan analisis klaster hierarkis. Metode rata-rata pengelompokan hierarki hierarki dalam kelompok dilakukan dengan menggunakan pendekatan agglomerative (bottom-up) dan phi linkage (korelasi analog Pearson untuk data biner) [25].

Penggunaan media sosial oleh wisatawan: Pendekatan pengelompokan. Analisis klaster mencakup dua fase. Pertama, tiga algoritma hirarkis diterapkan, yaitu linkage lengkap, linkage rata-rata dan metode Ward [26].

Menggunakan teknik data mining untuk membuat profil pelanggan hotel yang menguntungkan: Aplikasi analisis RFM. K-means adalah salah satu algoritma yang paling populer digunakan dalam analisis cluster, sebagian besar di bidang data mining, dan analisis data statistik. Untuk melakukan analisis klaster pada penelitian ini, nomor klaster harus dieksplorasi menggunakan SOM untuk kemudian digunakan dalam algoritma K-means [27].

Ada beberapa makalah yang membahas metode optimasi. Berikut ringkasan dari masing-masing makalah:

Analisis cluster yang optimal menggunakan hybrid K-Means dan Ant Lion Optimizer. K-Means adalah metode analisis klaster populer yang bertujuan untuk mempartisi sejumlah titik data ke dalam klaster K. Ini telah berhasil diterapkan pada sejumlah masalah. Namun, efisiensi K-Means tergantung pada inisialisasi pusat cluster. Teknik kecerdasan swarm yang berbeda diterapkan pada masalah pengelompokan untuk meningkatkan kinerja. Dalam karya ini pendekatan clustering hybrid berdasarkan K-means dan Ant Lion Optimization telah dipertimbangkan untuk analisis cluster yang optimal. Ant Lion Optimization (ALO) adalah model optimasi global stokastik [28].

Modifikasi Bee Colony Optimization (MBCO) dan hibridisasinya dengan k-means untuk aplikasi pengelompokan data. Dalam makalah ini, pendekatan Modified BCO (MBCO) diusulkan untuk pengelompokan data. Dalam MBCO yang diusulkan, karakteristik remisi lebah dan memberikan peluang yang adil untuk lebah yang dapat dipercaya dan tidak dapat dipercaya sedang dirawat. Untuk memvalidasi algoritma yang diusulkan, tujuh set data standar dipertimbangkan. Dari perhitungan persentase kesalahan klasifikasi, terlihat bahwa algoritma yang diusulkan memiliki kinerja yang lebih baik dari beberapa algoritma yang ada. Hasil simulasi menyimpulkan bahwa algoritma yang diusulkan dapat digunakan secara efisien untuk pengelompokan data [29].

Seleksi Fitur menggunakan Algoritma Genetika K-Means untuk Optimasi Multi-Objektif. Dalam makalah sebelumnya, optimasi multi-tujuan pada lingkungan belajar menggunakan algoritma genetika k-means (NLMOGA), diusulkan dan diterapkan pada beberapa set kehidupan nyata dalam data. Di NLMOGA, solusi dipilih dari repositori populasi global dan kemudian lingkungan belajar dibuat untuk mempromosikan evolusi setiap tujuan untuk solusi yang dipilih. Efektivitas pendekatan ini dievaluasi dengan berbagai kehidupan nyata dalam set data ekspresi gen benchmark [30].

Clustering Menggunakan Kombinasi Particle Swarm Optimization dan K-means. Makalah ini menerapkan kombinasi optimasi massa partikel dan K-means untuk pengelompokan data. Pendekatan yang diusulkan mencoba untuk meningkatkan kinerja teknik pengelompokan partisi tradisional seperti K-means dengan menghindari persyaratan awal jumlah cluster atau centroid untuk pengelompokan. Pendekatan yang diusulkan dievaluasi menggunakan berbagai dataset primer dan dunia nyata. Selain itu, makalah ini juga menyajikan perbandingan hasil yang dihasilkan oleh pendekatan yang diusulkan dan dengan K-means berdasarkan langkah-langkah validitas pengelompokan seperti jarak antar dan cluster, kesalahan kuantisasi, indeks siluet, dan indeks Dunn. Perbandingan hasil menunjukkan bahwa dengan bertambahnya ukuran kumpulan data, pendekatan yang diusulkan menghasilkan peningkatan yang signifikan dalam teknik pengelompokan partisi K-means [31].

Hybrid K-Means dan Peningkatan Self-Adaptive Particle Swarm Optimization untuk Pengelompokan Data. Algoritma K-Means saat ini merupakan salah satu teknik pengelompokan yang paling populer, karena kesederhanaan dan skalabilitasnya. Namun, kinerja K-Means sangat dipengaruhi oleh pilihan pusat cluster awal, yang dapat mengarah pada solusi suboptimal. Dalam makalah ini, algoritma pengelompokan hybrid baru diusulkan, bernama IDKPSOck, berdasarkan peningkatan yang dapat disesuaikan secara otomatis dalam Particle Swarm Optimization (PSO) dan K-Means, yang menggunakan operator crossover untuk meningkatkan kemampuan PSO untuk melarikan diri dari minimum lokal. titik dari ruang masalah. Untuk mengevaluasi kinerja pendekatan yang diusulkan, eksperimen dilakukan pada enam belas set data benchmark yang diperoleh dari UCI Machine Learning Repository. Evaluasi eksperimental, yang dilakukan dengan menggunakan uji hipotesis Friedman dalam kaitannya dengan empat metrik pengelompokan, telah menunjukkan efektivitas model yang diusulkan dalam kaitannya dengan algoritma perbandingan [32].

Hybrid Particle Swarm Optimization dan Analisis K-Means untuk Pengelompokan Jembatan Berdasarkan Data Inventarisasi Jembatan Nasional. Dalam makalah ini disajikan pendekatan optimasi berbasis hybrid dari metaheuristic particle crowd optimization dan k-means method (KPSO) dalam mengelompokkan data. Tujuannya adalah untuk mengelompokkan jembatan dengan atribut defisiensi struktural yang serupa dengan meminimalkan jumlah kesalahan kuadrat yang terkait dengan menetapkan titik data ke setiap cluster dan menentukan jumlah cluster yang paling tepat. Pendekatan yang disajikan dibandingkan dengan versi dasar optimasi partikel segerombolan (PSO) dan metode pengelompokan k-means tradisional. Algoritma diuji menggunakan database National Bridge Inventory (NBI). Hasil penelitian menunjukkan bahwa KPSO memberikan hasil yang lebih baik dalam hal fungsi tujuan dan menunjukkan peluang untuk menerapkan teknik optimasi untuk analisis data pada sistem infrastruktur sipil [33].

Berikut kesimpulan dari studi literatur yang telah dilakukan:

Penulis tidak menggunakan metode collaborative filtering karena metode ini memerlukan profil dari wisatawan lain, artinya merekomendasikan suatu item berdasarkan kesesuaian profil wisatawan yang satu dengan profil wisatawan lainnya. Sebagai contoh, sebagian besar wisatawan yang menyukai produk A juga menyukai produk B, jadi jika wisatawan lain menyukai produk A, kami dapat merekomendasikan produk B. Sedangkan metode content-based filtering memerlukan profil suatu barang dan riwayat aktivitasnya pada barang tersebut seperti peringkat atau suka atau tidak suka. Memerlukan profil berarti memerlukan fitur atau karakteristik item seperti tema, genre, penulis, tahun terbit, lokasi, atau nama tempat. Namun dalam hal segmentasi minat wisatawan pada kategori objek wisata ini, tidak ada kecocokan antara profil wisatawan yang satu dengan yang lain dan tidak ada ciri atau ciri seperti di atas. Oleh karena itu penulis menggunakan metode clustering untuk mengetahui karakteristik/minat setiap wisatawan untuk dapat menawarkan promo atau rekomendasi paket wisata sesuai minat wisatawan. Karakteristik atau kesamaan minat antara 1 wisatawan dengan wisatawan lainnya akan dikelompokkan menjadi 1 cluster. Dari masing-masing cluster yang terbentuk dapat memudahkan perusahaan untuk mengetahui kategori objek wisata apa saja yang diminati atau disukai oleh setiap wisatawan.

Penulis membaca beberapa makalah dan menemukan metode clustering yang digunakan untuk teknik segmentasi adalah K-Means dan DBSCAN. Oleh karena itu penulis melakukan perbandingan kedua metode tersebut dalam menangani kasus segmentasi minat wisatawan pada kategori objek wisata untuk mengetahui metode mana yang terbaik.

Dari beberapa makalah yang penulis baca, masing-masing teknik optimasi memiliki kegunaan yang berbeda-beda. Dalam hal segmentasi minat wisatawan pada kategori objek wisata ini, penulis memilih menggunakan teknik PSO.

Conclusion

Untuk mengetahui dan menentukan segmentasi minat wisatawan digunakan dua metode clustering yaitu metode K-Means dan DBSCAN. Kesimpulan yang diperoleh dari penelitian ini adalah metode K-Means dilakukan dengan terlebih dahulu menentukan jumlah cluster. Dalam hal segmentasi minat wisatawan, penulis menentukan nilai K adalah 5 dan metode DBSCAN dilakukan dengan menentukan jumlah minimum anggota dalam 1 cluster dan untuk dapat membuat 1 cluster diperlukan jarak minimum antar lingkungan atau berapa nilai ambang batas minimum. Misalnya kita membuat threshold sebesar 0.5 artinya jika > 0.5 tidak dianggap sebagai Neighborhood, sedangkan jika < 0.5 dianggap sebagai 1 Neighborhood, maka kita hitung ulang jarak dari titik pusat ke titik selanjutnya. Contoh lain, cluster 0 akan terbentuk jika jumlah anggota minimal terpenuhi, misalnya 15 orang.

Jumlah minimum anggota dalam DBSCAN sering disebut sebagai Min-Pts (Poin Minimum) dan jarak minimum antar lingkungan atau ambang batas minimum sering disebut sebagai Epsilon. Semua objek yang tidak cocok dengan cluster mana pun dianggap noise. Dalam hal segmentasi minat wisatawan, penulis menentukan nilai epsilon sebesar 3,0 dan nilai Min-Pts sebesar 100. Analisis kesamaan karakter setiap cluster yang terbentuk memanfaatkan kesamaan karakteristik atau minat antara 1 wisatawan dengan wisatawan lainnya akan dikelompokkan menjadi 1 cluster, dimana cluster-cluster yang terbentuk akan diberi nama yang sesuai dan menggambarkan karakter dalam 1 cluster, sehingga memudahkan pihak pemasaran di perusahaan travel untuk mengetahui dan memahami minat dari masing-masing kategori objek wisata. Hal ini dapat membuat wisatawan lebih tertarik menggunakan jasanya karena promo paket liburan atau wisata yang ditawarkan sesuai dengan minatnya.

Dari segi performansi, metode K-Means menghasilkan performansi Silhouette Coefficient yang lebih baik dibandingkan dengan metode DBSCAN karena semakin tinggi nilai koefisien siluet maka semakin baik metode clustering yang digunakan dan performansi DBI Coefficient yang dihasilkan metode DBSCAN lebih baik dibandingkan dengan K- Berarti metode karena nilai terkecil dari hasil perhitungan indeks Davies-Bouldin menyiratkan nilai untuk algoritma clustering yang lebih baik. Pada metode K-Means Clustering diperoleh Silhouette Coefficient sebesar (0,193) dan Coefficient DBI sebesar (1.740). Pada metode DBSCAN Clustering diperoleh Silhouette Coefficient sebesar (-0,054) dan Coefficient DBI sebesar (1,408).

Penulis tidak menggunakan metode collaborative filtering karena metode ini membutuhkan profil dari wisatawan lain, artinya merekomendasikan suatu item berdasarkan kecocokan antara profil wisatawan yang satu dengan profil wisatawan lainnya. Sebagai contoh, sebagian besar wisatawan yang menyukai produk A juga menyukai produk B, jadi jika wisatawan lain menyukai produk A, kami dapat merekomendasikan produk B. Sedangkan metode content-based filtering memerlukan profil suatu item dan riwayat aktivitasnya pada item tersebut seperti rating atau suka atau tidak suka. Memerlukan profil berarti memerlukan fitur atau karakteristik item seperti tema, genre, penulis, tahun penerbitan, lokasi, atau nama tempat. Namun dalam hal segmentasi minat wisatawan pada kategori objek wisata ini, tidak ada kecocokan antara profil wisatawan yang satu dengan yang lain dan tidak ada ciri atau ciri seperti di atas. Oleh karena itu penulis menggunakan metode clustering untuk mengetahui karakteristik/minat setiap wisatawan untuk dapat menawarkan promo atau rekomendasi paket wisata sesuai minat wisatawan.

Karakteristik atau kesamaan minat antara 1 wisatawan dengan wisatawan lainnya akan dikelompokkan menjadi 1 cluster. Dari masing-masing cluster yang terbentuk dapat memudahkan perusahaan untuk mengetahui kategori objek wisata apa saja yang diminati atau disukai oleh setiap wisatawan. Dari beberapa makalah yang penulis baca, masing-masing teknik optimasi memiliki kegunaan yang berbeda-beda. Dalam hal segmentasi minat wisatawan terhadap kategori objek wisata ini didapatkan bahwa teknik optimasi yang cocok untuk Metode K-Means adalah teknik PSO (Particle Swarm Optimization) karena yang ingin dioptimasi pada Metode K-Means adalah titik pusat awal. Biasanya titik centroid awal ditentukan secara manual dan acak, dengan teknik optimasi, penulis dapat mengetahui titik awal centroid mana yang terbaik dalam hal ini.

Teknik PSO ini telah membantu penulis untuk menemukan titik awal centroid terbaik oleh sistem, bukan secara manual. Manfaat dan kegunaan metode PSO adalah metode PSO dapat mengatasi kelemahan metode clustering tradisional yaitu pemilihan pusat cluster awal. Dampak Metode PSO terhadap Metode K-Means adalah dapat meningkatkan kinerja metode K-Means. Hal ini dikarenakan nilai DBI Coefficient metode K-Means kurang baik jika dibandingkan dengan metode DBSCAN. Berdasarkan hasil penelitian dari metode PSO K-Means didapatkan bahwa terdapat peningkatan nilai Silhouette Coefficient dan DBI Coefficient yang lebih baik pada metode ini dibandingkan dengan metode K-Means sebelum optimasi. Nilai Silhouette Coefficient yang dihasilkan adalah (0,189) dan Koefisien DBI (1,892). Hasil pengelompokan cluster dan analisis kemiripan karakter pada metode K-Means sama dengan hasil pengelompokan cluster dan analisis kemiripan karakter pada metode K-Means PSO.

Hasil analisis karakteristik cluster yang terbentuk dari metode PSO K-Means Clustering lebih sesuai dengan data eksternal mengenai alasan mengapa banyak wisatawan yang menjadikan negara-negara Eropa sebagai salah satu destinasi liburannya daripada metode DBSCAN. Untuk pengembangan lebih lanjut dari penelitian ini, penulis menyarankan agar lebih mengoptimalkan karakteristik wisatawan agar tidak hanya mengetahui rating kategori daya tarik wisata tetapi juga mengetahui genre, usia, dan hobi masing-masing wisatawan. Hal ini memungkinkan perusahaan pemasaran untuk mengenal pelanggannya lebih baik. Dapat menjangkau karakteristik wisata kategori daya tarik wisata di negara lain, tidak hanya di negara-negara Eropa.

An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation With Correlated Variables by HONG-HAO ZHAO 1, XI-CHUN LUO 2, RUI MA3, AND XI LU1

Bisnis omnichannel menjadi topik hangat karena perkembangan pesat di e-commerce dan pengenalan pelanggan dengan mode belanja multichannel. Berbagai organisasi bisnis memiliki ulai mengerjakan masalah bisnis omnichannel untuk memenuhi tren baru permintaan pelanggan dan cenderung mengabdikan upaya mereka untuk bisnis online dan ine. Dengan demikian, tidak diragukan lagi bahwa memahami perilaku belanja pelanggan online sangat penting untuk bisnis omnichannel. Model RFM (kebaruan, frekuensi, moneter) dan metode pengelompokan k-means biasanya digunakan untuk mengekstrak informasi pelanggan dan segmen pelanggan, masing-masing. Untuk memperluas model RFM, kami membagi frekuensi total dan informasi moneter menjadi data tingkat mingguan, dan sebagai akibatnya, jumlah variabel yang sesuai dengan satu pelanggan meningkat secara signifikan, yang mengarah ke masalah analisis dimensi tinggi. Untuk mengatasi masalah ini, dalam makalah ini kami memperluas metode pengelompokan k-means reguler dengan norma-L1 untuk kasus independen ke metode pengelompokan dengan penalti bersih elastis dengan fokus pada variabel berkorelasi. Hasil simulasi kami menunjukkan bahwa metode yang diusulkan berkinerja lebih baik daripada metode k-means standar dengan memberikan tingkat kesalahan yang lebih rendah dan dapat memilih variabel secara bersamaan di bawah 4 skenario yang berbeda. Contoh nyata dari pengecer online disajikan untuk menggambarkan penggunaan metode yang diusulkan dan menyoroti potensi tinggi dalam pengelompokan aplikasi berdimensi tinggi. Secara khusus, jumlah variabel berkurang dari 108 menjadi 98 tanpa kehilangan akurasi pengelompokan.

Bisnis omnichannel mulai menarik perhatian masyarakat karena aplikasinya yang luas dan potensinya yang besar di industri modern. Seperti diketahui bahwa kegiatan bisnis tradisional fokus pada transaksi ine antara organisasi bisnis dan pelanggan dan omnichannel mengintegrasikan bisnis tradisional dengan bisnis elektronik (E-business) dan mengambil keuntungan dari kedua belah pihak. Dengan pesatnya perkembangan teknologi internet saat ini, E-business mengalami pertumbuhan yang pesat [1].

Terdapat tren yang jelas bahwa pelanggan mengalihkan mode belanja mereka dari gerai tradisional ke internet, menghilangkan kendala waktu dan geografis dalam berbelanja [2], [3]. Selanjutnya, dengan berbelanja online, pelanggan dapat memperoleh informasi produk dari berbagai saluran dan cenderung beralih di antara saluran yang berbeda untuk mendapatkan pengalaman berbelanja yang lebih baik. Untuk mengakomodasi pola belanja baru pelanggan dan memberikan layanan yang lebih baik, organisasi bisnis menerapkan strategi yang berbeda untuk mencapai bisnis omnichannel dan salah satu contoh yang populer adalah mode bisnis online-to-of ine (O2O).

Jelas bahwa keberhasilan implementasi bisnis omnichannel bergantung pada pemahaman yang benar tentang preferensi belanja online pelanggan. Segmentasi pelanggan justru merupakan proses mengidentifikasi kelompok pelanggan yang berbeda sesuai dengan perilaku belanja mereka [4] [6]. Tidak seperti segmentasi pelanggan tradisional yang bergantung pada variabel seperti demografi dan sikap [7], catatan transaksi memainkan peran penting dalam segmentasi pelanggan online Untungnya, perkembangan modern dalam teknologi informasi memungkinkan organisasi bisnis untuk melacak data transaksi pelanggan secara tepat dan dengan demikian mengelompokkan pelanggan [8]. Untuk menerapkan segmentasi pelanggan, identifikasi variabel potensial yang dapat mengekspresikan perilaku belanja pelanggan menjadi tugas penting [9]. Model RFM yang diusulkan oleh Hughes [10] adalah model yang terkenal untuk mengkarakterisasi perilaku belanja pelanggan [11]. RFM merumuskan karakteristik pelanggan dengan mengekstrak informasi seperti kebaruan pembelian terakhir, frekuensi pembelian dan nilai moneter pembelian dari catatan transaksi pelanggan. Model RFM digunakan dalam banyak keadaan karena kesederhanaannya, dan aplikasi yang sukses dapat ditemukan di Bult dan Wansbeek [12], Newell [13], Migkautsch [14], Wu dan Lin [15] dan Lee dan Park [16] .Berdasarkan variabel yang diidentifikasi oleh model RFM, teknik data mining dapat diterapkan untuk mengelompokkan pelanggan. Clustering adalah salah satu jenis teknik data mining yang membagi sekumpulan objek ke dalam beberapa kategori dimana objek-objek serupa dikelompokkan bersama ([17], [18]). K-means adalah algoritma clustering yang terkenal yang pertama kali diusulkan oleh MacQueen [19]. Ini telah terbukti menjadi teknik pengelompokan yang efektif yang dapat menangani masalah pengelompokan secara efisien [20], meskipun banyak pendekatan lain telah dilaporkan dalam literatur [21] [25]. Contoh sukses juga dapat ditemukan di Wei et al. [26], Mesforoush dan Tarokh [20], Cheng dan

Chen [1] dan Kuo dkk. [27]. Meskipun model RFM dan metode k-means standar telah memperoleh hasil yang menarik, masalahnya cenderung berbeda di bawah keadaan bisnis modern. Tercatat bahwa model RFM mengkonstruksi tiga variabel yaitu resensi, frekuensi, moneter. Ambil contoh frekuensi belanja, model RFM hanya menghitung jumlah total frekuensi belanja setiap pelanggan. Pelanggan dengan total frekuensi belanja yang sama tidak perlu memiliki perilaku belanja yang sama, melainkan distribusi frekuensi belanja dari waktu ke waktu mencerminkan pola dan perilaku belanja secara lebih tepat. Dalam keadaan seperti itu, untuk mensegmentasi pelanggan online secara tepat, diperlukan frekuensi belanja dengan skala waktu yang lebih kecil. Dengan demikian model RFM tradisional menjadi tidak efektif karena skala waktu yang digunakannya. Untuk mencapai skala waktu yang lebih kecil, salah satu solusi yang mungkin dilakukan adalah dengan menguraikan data tahunan menjadi data level mingguan dengan mengekstrak catatan data transaksi pelanggan, yang berbeda dari model RFM tradisional yang menerapkan jumlah total untuk menggambarkan frekuensi belanja dan pengeluaran untuk setiap pelanggan. Selanjutnya, data aliran klik lain yang tersedia di lingkungan e-niaga juga dapat diperkenalkan. Akibatnya, jumlah variabel yang mencirikan perilaku pelanggan meningkat secara signifikan dan mengarah pada masalah pengelompokan dimensi tinggi.

Akibatnya, metode k-means standar menjadi kurang efektif dalam kondisi dimensi tinggi [28]. Metode k-means terreguler yang menambahkan istilah penalti ke metode k-means standar dapat menjadi pendekatan yang efektif untuk menangani masalah berdimensi tinggi. Aplikasi yang berhasil dalam kasus cluster berdimensi tinggi dengan variabel independen dibahas oleh Witten dan Tibshirani [29] dan oleh Sun dan Wang [30].

Aplikasi yang berhasil dalam kasus cluster berdimensi tinggi dengan variabel independen dibahas oleh Witten dan Tibshirani [29] dan oleh Sun dan Wang [30]. Namun, variabel tidak dapat sepenuhnya independen dalam aplikasi nyata dan sedikit penelitian telah dilakukan di sepanjang arah ini. Secara umum dalam lingkungan e-commerce modern, segmentasi pelanggan yang tepat membutuhkan lebih banyak variabel yang mencirikan perilaku pelanggan. Ketika data skala waktu kecil didekomposisi dari jumlah total, kemungkinan jumlah fitur pelanggan lebih banyak daripada jumlah pelanggan terutama untuk usaha kecil dan platform dengan jumlah pelanggan yang sangat terbatas. Akibatnya, fitur yang digunakan untuk segmentasi pelanggan berkorelasi dan masalah menjadi masalah clustering berdimensi tinggi. Akibatnya, metode pengelompokan yang efektif untuk pengelompokan dimensi tinggi dengan variabel berkorelasi diinginkan. Oleh karena itu, dalam makalah ini, kami memperluas metode pengelompokan k-means reguler dari fokus pada variabel independen ke fokus pada variabel berkorelasi dengan memperkenalkan penalti bersih elastis. Metode yang diusulkan dinilai di bawah kondisi pengelompokan dimensi tinggi dengan variabel berkorelasi. Hasil simulasi kami menunjukkan bahwa metode yang diusulkan berkinerja lebih baik daripada metode k-means standar dengan memberikan tingkat kesalahan yang lebih rendah dan dapat memilih variabel secara bersamaan di bawah 4 skenario yang berbeda. Selanjutnya, jumlah variabel berkurang dari 108 menjadi 98 tanpa kehilangan akurasi pengelompokan dalam aplikasi pada contoh nyata.

K-MEANS CLUSTERING TERregulasi

Untuk mengatasi kelemahan metode k-means standar,

kami memperkenalkan metode k-means reguler di bagian ini.

Gagasan regularisasi dikemukakan oleh Tibshirani [32]

di mana istilah penalti ditambahkan ke kuadrat terkecil asli

fungsi untuk mendapatkan perkiraan untuk koefisien dalam linear

masalah regresi. Istilah hukuman tambahan dapat mengecilkan

koefisien ke nol ketika koefisien tersebut mendekati nol.

Akibatnya, metode regularisasi menangani model tting

dan pemilihan variabel secara bersamaan. Sifat rinci dari

metode regularisasi dapat ditemukan di Zou dan Hastie [33].

Dalam makalah ini, kami menerapkan pendekatan serupa dengan menambahkan

istilah regularisasi untuk Persamaan. (1) mirip dengan Sun dan Wang [30].

Masalah optimasi baru diharapkan dapat memecahkan dimensi tinggi

masalah pengelompokan dengan mengambil keuntungan dari istilah regularisasi, dan bentuk rinci diberikan oleh

Menyelesaikan

Hasil pengelompokan metode k-means regularized adalah solusi optimal untuk Persamaan. (2). Mirip dengan kmean standar, solusi optimal untuk k-means reguler adalah tidak mudah untuk dicapai. Oleh karena itu, kami masih menerapkan iterative pendekatan untuk memecahkan Persamaan. (2).

Mirip dengan pendekatan iteratif untuk menyelesaikan metode k-means standar, kami masih menghitung cluster Bj dan pusat Cj secara terpisah. Ketika Cj dikalikan, Bj dapat dengan mudah dihasilkan dengan menugaskan Xi ke yang terdekat Tengah. Ketika cluster Bj di-xed, Cj dapat diperoleh dalam mode komponen seperti yang dibahas dalam Sun andWang [30]