Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi

Vol. 5, No. 1, Januari 2009 **ISSN 0216 - 0544**

PENGGUNAAN *CLUSTER-BASED SAMPLING* UNTUK PENGGALIAN KAIDAH ASOSIASI MULTI OBYEKTIF

[*] Febriana Santi Wahyuni, Daniel O Siahaan dan Chastine Fatichah
Jurusan Teknik Informatika, Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember (ITS) Surabaya
Gedung Teknik Informatika, Kampus ITS, Jl. Raya ITS, Sukolilo, Surabaya, 60111
E-Mail: [*] vbryana@yahoo.com

Abstract

The exploration of association models is used to identify association model of item combination in each part, to detect co-occur attributes in common frequency, and to conduct a kind of model of the groups. Furthermore, to explore a single association, the quality measurement of the risen models is frequently only based on one evaluative criteria, that is confidence factor or predictive accuracy. In the previous research, it was used a measurement criteria comprehensibility, interestingness and confidence factor or predictive accuracy. Therefore, it results in a model describing the real condition. Cluster based sampling is used in this research to carry out clustering. This is purposed that the data used in exploring association models have been clustered well. The use of the sampling technique can increase the model resulted. Algorithm K-means and Fuzzy C Means is used in this research in the course of clustering process. The examining shows that the clustered data can result in model having a higher average value confidence, that is 85 % at the minimum support 20 and 40, than data without clustering, clustering process data using K-means results in a higher average value confidence than Fuzzy C-Means (83%) at the minimum support 20 and 40.

Key words: Cluster-Based Sampling, Assosiation Rule, Multiobjective.

PENDAHULUAN

Data mining berkembang menjadi alat bantu untuk mencari pola-pola yang berharga dalam suatu basisdata yang sangat besar jumlahnya, sehingga tidak memungkinkan dicari secara manual. Beberapa teknik *data mining* dapat diklasifikasikan ke dalam kategori berikut, meliputi klasifikasi, *clustering*, penggalian kaidah asosiasi, analisa pola sekuensial, prediksi, visualisasi data dan lain sebagainya. *Association mining* diusulkan pertama kali oleh [1], yang selanjutnya berperan utama dalam mendukung penelitian, pembangunan dan aplikasi dari teknik-teknik *data mining* selanjutnya. Beberapa teknik dari *Association mining* ini telah dikembangkan sampai saat ini. *Association rule mining* dipergunakan untuk mencari kaidah asosiasi antara suatu kombinasi item. Mendeteksi kumpulan-kumpulan atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut. Contoh, 90% dari orang yang berbelanja di suatu *supermarket* yang membeli roti juga membeli selai, dan 60% dari semua orang yang berbelanja membeli keduanya.

Kebanyakan dari permasalahan-permasalahan di dunia nyata merupakan permasalahan multi obyektif yang seharusnya secara bersama-sama dioptimalkan untuk memperoleh hasil yang terbaik dari permasalahan tersebut. Demikian halnya dengan masalah-masalah penggalian kaidah asosiasi. Menemukan sebuah solusi tunggal untuk sebuah masalah multi obyektif sulit

untuk dilakukan. Sehingga merupakan hal yang umum untuk mencari sekumpulan solusi berdasarkan pada kriteria-kriteria yang tidak dominan. Sebuah pendekatan untuk menyelesaikan masalah multi obyektif disarankan oleh Vilfredo Pareto. Teknik optimasi yang berdasarkan pendekatan ini disebut dengan teknik optimasi Pareto. Pada penelitian yang dilakukan oleh [2], digunakan tiga pengukuran meliputi *support count*, *comprehensibility* dan *interestingness*, untuk mengevaluasi sebuah kaidah sehingga dapat dipikirkan sebagai obyektif-obyektif yang berbeda dari masalah penggalian asosiasi. *Support count* adalah banyaknya *record* yang memenuhi semua kondisi sebelumnya dari suatu kaidah. *Comprehensibility* adalah banyaknya atribut yang terlibat dalam sebuah kaidah dan mencoba untuk menentukan kemampuan memahami dari kaidah-kaidah tersebut. Dan *interestingnes*s adalah seberapa pentingnya sebuah kaidah. Jika jumlah dari atribut-atribut yang terlibat dalam bagian *antecedent* lebih sedikit, maka kaidah tersebut lebih komprehensif. Sebuah kaidah yang memiliki nilai *support count* sangat tinggi, akan diukur sebagai kurang menarik.

Pada umumnya data yang digunakan untuk penggalian kaidah asosiasi sangatlah besar dan terdapat variasi data yang sangat tinggi. Hal ini dapat mengurangi kualitas dari kaidah yang dihasilkan. Oleh karena itu, pada penelitian ini digunakan *cluster-based sampling* untuk melakukan *clustering data*, agar data yang digunakan untuk penggalian kaidah asosiasi sudah terklaster dengan baik. Kemudian dari data yang terklaster tersebut, dilakukan pengambilan sampel yang digunakan untuk penggalian kaidah asosiasi. Dengan menggunakan teknik sampling ini diharapkan dapat meningkatkan kualitas dari kaidah yang dihasilkan.

Permasalahan yang dirumuskan dalam penelitian ini adalah bagimana membandingkan kualitas kaidah-kaidah yang dihasilkan pada proses penggalian kaidah asosiasi multi obyektif antara data yang di-cluster dengan data yang tidak di-*cluster*, serta bagaimana mengetahui algoritma yang lebih baik antara algoritma K-*Means* dan *Fuzzy* C *Means* untuk penggalian kaidah asosiasi multi obyektif.

Adapun tujuan dari penelitian ini adalah penggunaan *cluster-based sampling* untuk penggalian kaidah asosiasi multi obyektif untuk membandingkan kualitas kaidah-kaidah yang dihasilkan pada proses penggalian kaidah asosiasi multi obyektif antara data yang di-*cluster* dengan data yang tidak di-*cluster*. Dalam hal ini kualitas dari kaidah yang lebih baik adalah yang mempunyai rata-rata nilai *confidence* yang lebih tinggi. Serta membandingkan algoritma K-*Means* dan *Fuzzy* C-*Means* untuk penggalian kaidah asosiasi multi obyektif.

Manfaat penelitian yang diajukan adalah bahwa data yang di-*cluster* akan menghasilkan kaidah-kaidah yang lebih berkualitas dibandingkan dengan data yang tidak melalui proses *clustering*. Dan mengetahui algoritma *clustering* yang lebih baik antara K-*Means* dan *Fuzzy* C-*Means* untuk penggalian kaidah asosiasi multi obyektif.

Penggalian Kaidah Asosiasi

Data mining muncul di saat analisis data menjadi sangat komplek dalam memajukan manajemen bisnis, dimana *data mining* dapat membantu penggunanya untuk mengetahui pola dan keteraturan alam himpunan data yang sifatnya tersembunyi. *Data mining* diartikan sebagai proses ekstraksi informasi yang berguna dan potensial dari sekumpulan data yang terdapat secara implisit dalam suatu bisnis data. Terdapat banyak istilah dari *data mining* yang dikenal luas seperti *Knowledge Mining From Database*, *Knowledge Extraction*, *Data Archeology, Data Dredging* dan lain sebagainya [3].

Semakin berkembangnya kebutuhan manusia untuk mengolah basisdata sehingga memicu perkembangan dari metode-metode *data mining*. Beberapa metode yang dikenal di dalam *data mining* yaitu penggalian kaidah sekuensial, klasifikasi data dan korelasi data serta kaidah asosiasi.

Penggalian kaidah asosiasi adalah salah satu teknik *data mining* untuk menemukan kaidah asosiasi antara suatu kombinasi *item* [1]. Sebagai contoh berdasarkan basisdata penjualan dari sebuah pasar swalayan, dimana *record* menggambarkan transaksi pembelian yang dilakukan oleh para pelanggan dan atribut-atributnya menggambarkan barang-barang yang disediakan atau dijual. Dari kaidah asosisasi yang diperoleh dari analisa pembeliannya dapat diketahui seberapa besar

kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Misalnya terdapat kaidah asosiasi {roti, mentega}→ {susu}, dengan nilai *support* nya 40% dan nilai *confidence*-nya 50%. Artinya bahwa seorang pelanggan yang membeli roti dan mentega mempunyai kemungkinan 50% untuk juga membeli susu. Aturan ini cukup signifikan karena mewakili 40% dari catatan transaksi selama ini. Dengan pengetahuan tersebut pengelola pasar swalayan dapat mengatur untuk promosi pemasaran dengan menggunakan kupon diskon untuk beberapa kombinasi barang tertentu, peletakan barang dan lain-lain. Terdapat banyak daerah aplikasi untuk teknik-teknik penggalian kaidah asosiasi, termasuk rancangan katalog, rancangan toko, pembagian pelanggan, diagnosa alarm telekomunikasi dan lain sebagainya.

Dalam menentukan suatu kaidah asosiasi, terdapat suatu *interestingness measure* (ukuran kepercayaan) yang didapatkan dari hasil pengolahan data dengan perhitungan tertentu. Umumnya ada dua ukuran, yaitu:

1. *Support*: suatu ukuran yang menunjukkan seberapa besar tingkat dominasi suatu *item/itemset* dari keseluruhan transaksi. Ukuran ini akan menentukan apakah suatu *item/itemset* layak untuk dicari *confidence*-nya (misal, dari seluruh transaksi yang ada, seberapa besar tingkat dominasi yang menunjukkan bahwa *item A* dan *B* dibeli bersamaan) dapat juga digunakan untuk mencari tingkat dominasi *item* tunggal.
2. *Confidence*: suatu ukuran yang menunjukkan hubungan antar 2 *item* secara *conditional* (misal, seberapa sering *item B* dibeli jika orang membeli *item* A).

Sebuah kaidah asosiasi adalah sebuah implikasi $A \rightarrow B$, dimana sekumpulan *item A* dan *B* tidak saling beririsan (*intersect*). Masing-masing kaidah asosiasi mempunyai dua kualitas pengukuran yaitu *support* dan *confidence* yang didefinisikan sebagai berikut:

Support :

$$supp(A \rightarrow B) = prob\{A \cup B\} \quad (1)$$

Confidence:

$$conf(A \rightarrow B) = supp\{A \cup B\} / supp\{A\} \quad (2)$$

Kedua ukuran ini nantinya akan berguna dalam menentukan *interestingness* kaidah asosiasi, yaitu untuk dibandingkan dengan *threshold* (batasan) yang ditentukan. Batasan tersebut umumnya terdiri dari *min_support* dan *min_confidence*.

Kedua ukuran ini nantinya akan berguna dalam menentukan *interestingness* kaidah asosiasi, yaitu untuk dibandingkan dengan *threshold* (batasan) yang ditentukan. Batasan tersebut umumnya terdiri dari *min_support* dan *min_confidence*.

Metodologi dasar penggalian asosiasi terbagi menjadi dua tahap meliputi:

1. *Frequent itemset generation.* Pada tahapan ini dilakukan pencarian kombinasi *item* yang memenuhi syarat minimum dari nilai *support* dalam basisdata.
2. *Rule Generation.* Setelah semua kaidah frekuensi tinggi ditemukan, selanjutnya mencari turunan asosiasi yang memenuhi syarat minimum *confidence* dengan menghitung *confidence* asosiasi $A \rightarrow B$ dari *support* kaidah frekuensi tinggi *A* dan *B* dengan menggunakan rumus (2).

Terdapat dua proses utama yang dilakukan pada algoritma Apriori meliputi:

1. *Join (*penggabungan). Untuk menemukan L_k, C_k dibangkitkan dengan melakukan proses join L_{k-1} dengan dirinya sendiri, $Ck = L_{k-1} * L_{k-1}$, kemudian C_k diambil hanya yang terdapat dalam L_{k-1}.
   Untuk menemukan L_k, C_k dibangkitkan dengan melakukan proses join L_{k-1} dengan sendirinya
2. *Prune* (pemangkasan). Menghilangkan anggota C_k yang memiliki *support count* l; lebih kecil dari min support supaya tidak dimasukkan ke L_k

Tahapan yang dilakukan algoritma apriori untuk membangkitkan *large itemset* adalah:

1. Menelusuri seluruh *record* pada basis data transaksi dan menghitung *support count* dari tiap *item*.
2. Large 1 *itemset* L_1 dibangun dengan menyaring C_1 dengan *support count* yang lebih besar atau sama dengan *min support* untuk dimasukkan ke dalam L_1.
3. Untuk membangun L_2 algoritma apriori menggunakan proses *join* untuk menghasilkan C_2.
4. Dari C_2, *itemset* yang memiliki *support count* lebih besar atau sama dengan *min support* akan disimpan dalam L_2.
5. Proses ini diulang sampai tidak ada lagi kemungkinan k-*itemset.*

Contoh proses pembangkitan kandidat untuk dijadikan *itemset* dan *large itemset* dapat dilihat pada Gambar 1.

Penggalian Kaidah Asosiasi Multi Obyektif

Permasalahan-permasalahan penggalian kaidah asosiasi dapat dianggap sebagai sebuah masalah yang multi obyektif, karena masalah-masalahnya kompleks. Terdapat beberapa pendekatan untuk permasalahan multi obyektif ini. Salah satunya adalah pendekatan optimasi Pareto. Dalam definisi optimasi Pareto, sebuah solusi S_1 dikatakan menjadi dominan pada solusi yang lain S_2, jika dan hanya jika solusi S_1 tersebut adalah jelas lebih baik daripada paling tidak satu dari kriteria dan tidak lebih buruk daripada, S_1 dalam evaluasi dari keseluruhan kriteria. Sebuah solusi dinyatakan sebagai solusi tidak dominan jika solusi tersebut adalah lebih unggul dari semua solusi yang lain dalam semua kriteria dari optimasi [4]. Sebuah pendekatan untuk menyelesaikan masalah multi obyektif disarankan oleh Vilfredo Pareto. Teknik optimasi yang berdasarkan pendekatan ini disebut dengan teknik optimasi Pareto.

Pada penelitian yang dilakukan oleh [2], digunakan tiga pengukuran meliputi *support count*, *comprehensibility* dan *interestingness*, untuk mengevaluasi sebuah kaidah sehingga dapat dipikirkan sebagai obyektif-obyektif yang berbeda dari masalah penggalian asosiasi. *Support count* adalah banyaknya *record* yang memenuhi semua kondisi sebelumnya dari suatu kaidah. *Comprehensibility* adalah banyaknya atribut yang terlibat dalam sebuah kaidah dan mencoba untuk menentukan kemampuan memahami dari kaidah-kaidah tersebut. Untuk menghitung nilai *Comprehensibility* sebuah *rule* digunakan Persamaan (3).

$$Comprehensibility=Log(1+|C|/log(1+|A\cup C|) \quad (3)$$

Dimana $|C|$ adalah jumlah atribut yang terlibat dalam bagian *consequent*, dan $|A\cup C|$) adalah total kaidah.

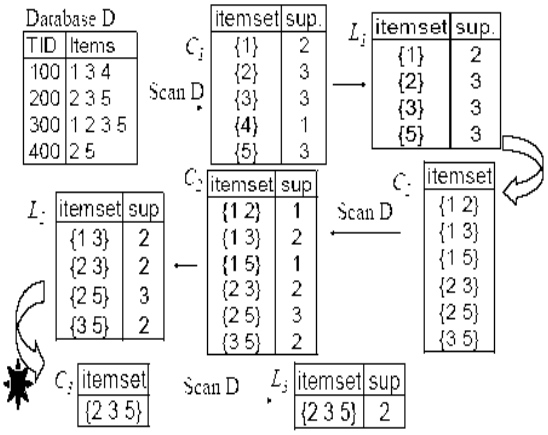
Dan *interestingness* adalah seberapa pentingnya sebuah kaidah. Jika jumlah dari atribut-atribut yang terlibat dalam bagian *antecedent* lebih sedikit, maka kaidah tersebut lebih komprehensif. Sebuah kaidah yang memiliki nilai *support count* sangat tinggi, akan diukur sebagai kurang menarik. Nilai *interestingness* dari sebuah kaidah diperoleh dengan menggunakan Persamaan (4).

$$Interestingness = [SUP(A\cup C)/SUP(A)x[SUP(A\cup C)/SUP(C)]x[1-(SUP(A\cup C)/+|D|] \quad (4)$$

Dimana : $|D|$= jumlah *record* dalam basisdata.

Cluster-Based Sampling

Clustering merupakan salah satu metode *data mining* yang bersifat tanpa arahan (*unsupervised)*. *Clustering* data adalah proses dari pengelompokkan data berdasarkan similaritas atau kesamaan antara data. Similaritas *clustering* dapat diaplikasikan untuk beberapa bidang, misalnya di bidang penelitian pasar, *clustering* digunakan untuk membagi populasi umum dari konsumen-konsumen ke dalam segmen pasar, pembagian pasar dan menentukan sasaran pasarnya.



Gambar 1. Pembangkitan Kandidat *Itemset* dan *Large Itemset*.

Terdapat beberapa pendekatan dari *clustering*, salah satunya adalah untuk basisdata yang besar, dimana digunakan untuk *sampling* dan *compression*. Metode pengelompokan *clustering* adalah kumpulan obyek data dimana jika obyek data yang terletak di dalam *cluster* memiliki kemiripan sedangkan yang tidak berada dalam satu *cluster* tidak memiliki kemiripan. Jika terdapat *n* obyek pengamatan dengan *p* variabel maka sebelum dilakukan pengelompokkan data atau obyek, terlebih dahulu ditentukan ukuran kedekatan sifat antar data yang ada. Ukuran kedekatan data yang biasa digunakan adalah jarak *euclidius* (*euclidean distance*) antara dua obyek dari p

dimensi pengamatan. Jika obyek pertama yang diamati adalah X=[x1,x2,x3,…,xp] dan Y=[y1,y2,y3,…,yp] maka perhitungan jarak dengan menggunakan *euclidean distance* untuk satu vektor digunakan Persamaan (5).

$$D_{L_2}(x_2 - x_1) = \|x_1 - x_2\| = \sum_{j-1}^{p} |x_{2j} - x_{1j}| \qquad (5)$$

Sampling adalah proses pemilihan unsur-unsur (item-item) yang mewakili suatu populasi (seluruh unsur/item yang ada) secara sistematis dengan tujuan mempelajari unsur/item tersebut. Pada penelitian ini digunakan sampling berbasis klaster, dimana teknik clustering yang digunakan adalah algoritma K-*Means* dan *Fuzzy* C *Means*.

K-*Means*

Algoritma K-*Means* merupakan metode yang umum digunakan pada teknik *clustering*. Menurut Mac Queen [5], K-*Mean* adalah salah satu algoritma *unsupervised learning* yang paling sederhana yang dikenal dapat menyelesaikan permasalahan *clustering* dengan baik. Ide utamanya adalah mendefinisikan *centroid* sejumlah *k*, untuk masing-masing klaster. *Centroid* ini harus diletakkan dengan cara yang cerdik pada satu tempat, karena lokasi yang berbeda akan menyebabkan hasil yang berbeda pula. Maka sebaiknya meletakkan sebisa mungkin berjauhan satu dengan yang lain. Langkah berikutnya adalah mengambil masing-masing titik kepunyaan sekumpulan data tertentu dan menghubungkannya ke *centroid* yang terdekat. Ketika tidak ada lagi titik yang belum dihubungkan, maka langkah pertama terlengkapi dan satu pengelompokan awal telah dilakukan. Dalam posisi ini perlu dihitung kembali *k centroid-centroid* baru sebagai *barycenters* dari hasil klaster-klaster pada langkah sebelumnya. Setelah mempunyai *centroid* baru, satu keterikatan harus dilakukan antara titik-titik sekumpulan data yang sama dengan *centroid* yang baru. Satu pengulangan telah dilakukan, sebagai hasil dari pengulangan ini terlihat bahwa *k centroid* mengubah lokasi mereka secara bertahap sampai tidak ada lagi perubahan yang dilakukan. Dengan kata lain *centroid* tidak bergerak/berubah lagi.

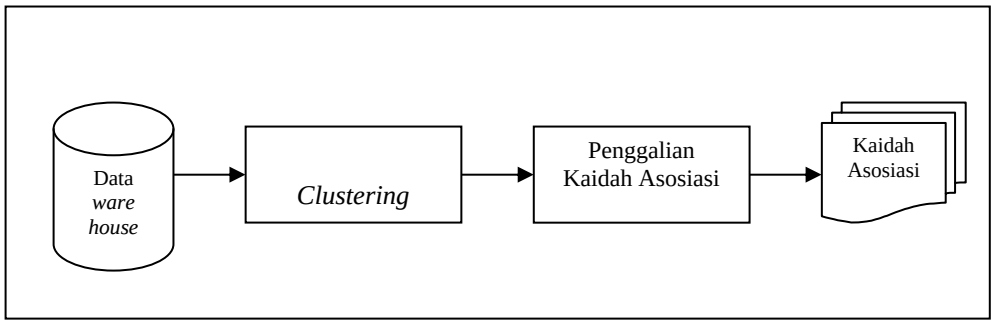
Fuzzy C -Means

Adalah sebuah metode *clustering* yang mengijinkan satu data menjadi milik dua atau lebih *cluster*. Metode ini sering digunakan dalam pengenalan pola (*pattern recoqnition*). Metode *Fuzzy* C-*Means* adalah salah satu metode *clustering* yang mengalokasikan kembali data kedalam masing-masing *cluster* dengan memanfaatkan teori *Fuzzy*. Dalam metode *Fuzzy* C-*Means* dipergunakan variabel *membership function* i_{ku}, yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota ke dalam suatu *cluster* [6].

Pada metode ini juga digunakan suatu variabel *m* yang merupakan *weighting exponent* dari *membership function*. Variabel ini dapat mengubah besaran pengaruh dari *membership function* i_{ku} dalam proses *clustering*. Variabel *m* mempunyai wilayah nilai $m > 1$, sampai pada saat ini tidak ada ketentuan yang jelas berapa besar nilai *m* yang optimal dalam melakukan optimasi suatu permasalahan *clustering*. Nilai *m* yang umum digunakan adalah 2. *membership function* untuk suatu data ke suatu *cluster* tertentu dihitung menggunakan Persamaan (6).

$$u_{ik} = \sum_{j=1}^{c} \left(\frac{D(x_k, v_i)}{D(x_k, v_j)} \right)^{\frac{2}{m-1}} \qquad (6)$$

Dimana: u_{ik} = *membership function* data ke-k ke *cluster* ke-i
v_i = nilai *centroid cluster* ke-i
m = *weighting exponent*



Gambar 2. Diagram Blok Sistem

Secara mendasar terdapat dua cara pengalokasian data kembali kedalam masing-masing *cluster* pada saat proses iterasi *clustering*. Yang pertama adalah pengalokasian dengan cara tegas (*hard*). Dimana data *item* secara tegas dinyatakan sebagai anggota *cluster* yang satu dan tidak menjadi anggota *cluster* yang lain. Yang kedua dengan menggunakan nilai *Fuzzy* dimana masing-masing data *item* diberikan nilai kemungkinan untuk bisa bergabung ke setiap *cluster* yang ada.

Pada K-*Means* pengalokasian data kembali didasarkan pada perbandingan jarak antara data dengan *centroid* setiap *cluster* yang ada. Pada *Fuzzy* C-*Means* pengalokasian kembali data kedalam masing-masing *cluster* dipergunakan variabel *membership function* u_{ik} yang merujuk pada seberapa besar kemungkinan suatu data bisa menjadi anggota dalam satu *cluster*. Selain itu juga digunakan variabel *m* yang merupakan *weighting exponent* dari *membership function.*

PERANCANGAN SISTEM

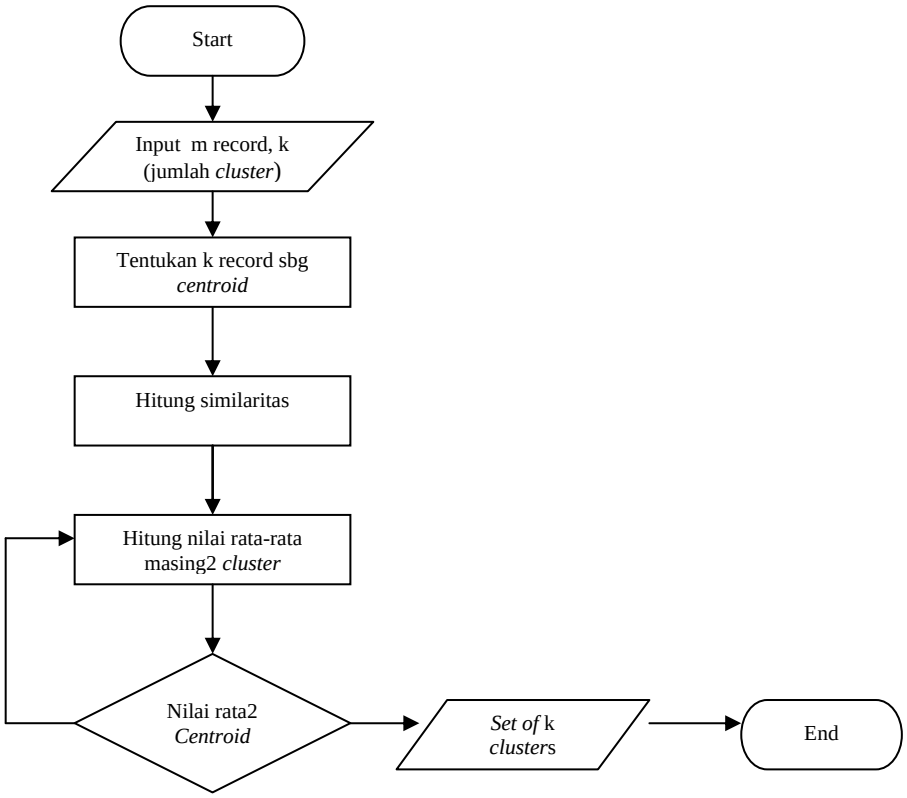
Langkah-langkah yang dilakukan di dalam penelitian ini meliputi Perancangan algoritma, Implementasi algoritma, Uji coba, Evaluasi. Adapun diagram blok dari sistem ditunjukkan pada Gambar 2. Terdapat dua proses utama dalam sistem yaitu proses *clustering* dan proses penggalian kaidah asosiasi. Data transaksi yang berasal dari data *warehouse* akan melalui proses klasterisasi terlebih dahulu sebelum dilakukan proses penggalian kaidah asosiasinya.

Dalam perancangan algoritma, terdapat dua proses utama yaitu proses klasterisasi data *input* dan proses penggalian kaidah asosiasi multi obyektif. Dari proses klasterisasi akan diperoleh data sampel yang sudah dikelompok-kelompokkan, yang selanjutnya digunakan sebagai data input untuk proses penggalian kaidah asosiasi. Di dalam penelitian ini digunakan dua metode *clustering* yaitu K-*Means* dan *Fuzzy* C *Means*, *flowchart* dari masing-masing metode tersebut berturut-turut ditunjukkan pada Gambar 3 dan Gambar 4.

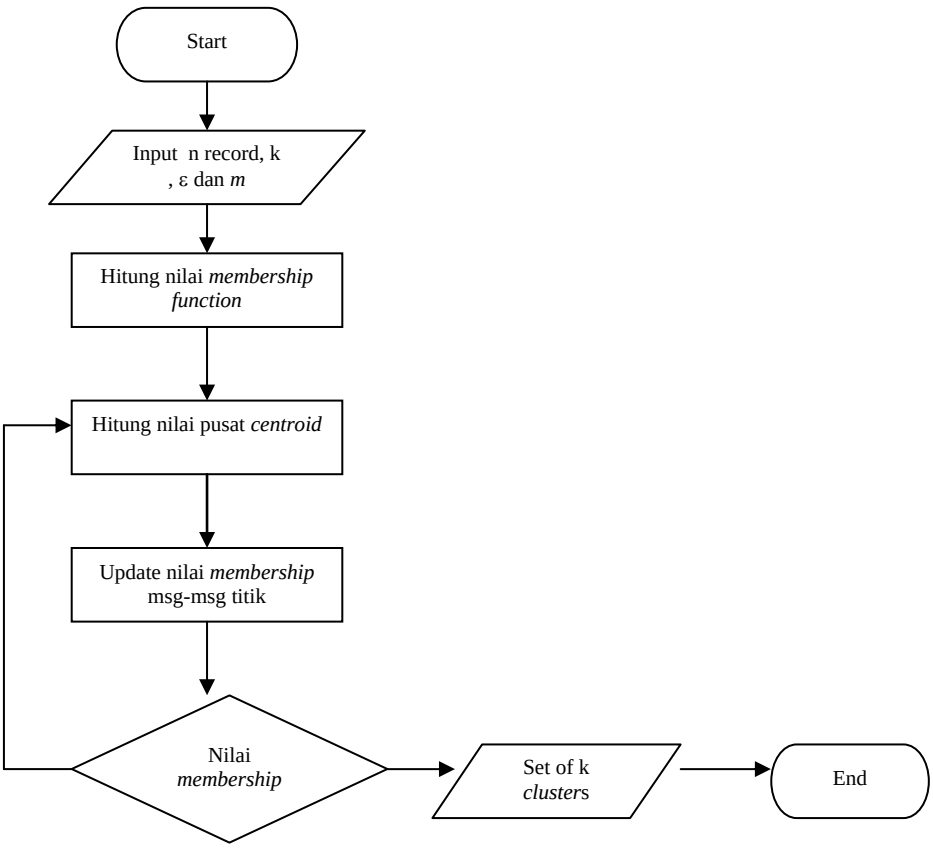
Tahapan yang dilakukan dalam melakukan *clustering* dengan metode K-*Means* adalah pertama dipilih *k record* secara *random* sebagai *centroid* atau pusat *cluster* untuk masing-masing *k cluster*. Selanjutnya menghitung jarak dari masing-masing *record* dengan *record* yang lain dengan menggunakan Persamaan (5). Dengan menggunakan jarak tersebut, selanjutnya adalah menentukan *record* r_i ke sebuah *cluster*, sedemikian sehingga jarak r_i dan *centroid cluster* adalah terkecil di antara *cluster* yang lain. Berikutnya adalah menghitung *kembali nilai rata-rata dari centroid untuk masing-masing cluster berdasarkan record* yang menjadi anggota dari *cluster* tersebut. Adapun untuk menghitung nilai rata-rata *centroid* digunakan Persamaan (7). Proses tersebut diulangi sampai nilai rata-rata dari masing-masing *centroid* tidak berubah atau mendekati tetap.

$$v_{i,j} = \frac{\sum_{k-1}^{N_i} x_{kj}}{N_i} \qquad (7)$$

Dimana *Ni* adalah jumlah data yang menjadi anggota *cluster i.*



Gambar 3. *Flowchart* K-*Means*.



Gambar 4. *Flowchart Fuzzy* C *Means.*

Tabel 1. Hasil Uji Coba dengan Nilai *Min_Supp* 20.

Min_Conf	Non *Cluster*	K-*Means*				*Fuzzy* C *Means*			
		2	3	4	5	2	3	4	5
60	74	73	76	83	82	78	74	77	79
75	79	89	92	87	88	88	92	82	87
Rata-rata	76.5	81	84	85	85	80	83	79.5	83

Tabel 2. Hasil Uji Coba dengan Nilai *Min_Supp* 40.

Min_Conf	Non *Cluster*	K-*Means*				*Fuzzy* C *Means*			
		2	3	4	5	2	3	4	5
60	69	75	80	84	83	73	77	78	77
75	90	90	91	87	87	90	93	91	72
Rata-rata	79.5	82.5	85.5	85.5	85	81.5	85	84.5	74.5

Adapun tahapan yang dilakukan untuk melakukan proses *clustering* dengan menggunakan metode *Fuzzy* C *Means* adalah pertama dimulai dengan menentukan jumlah *cluster* yang diinginkan, menentukan nilai *m* (*weighting exponent)* dan menentukan nilai ε (*threshold*). Selanjutnya menghitung nilai *membership* masing-masing titik dari data dengan menggunakan Persamaan (6). Dilanjutkan dengan menghitung nilai pusat *centroid*, nilai ini diperoleh dengan menggunakan Persamaan (7). Nilai dari semua titik diperbaiki berdasarkan dari nilai pusat *centroid* yang baru. Penghitungan nilai pusat *centroid* dengan nilai *membership* dilakukan sampai nilai *membership* lebih kecil dari nilai ε (*threshold*) yang telah ditentukan diawal.

Penggalian kaidah asosiasi dilakukan dengan menggunakan Algoritma Apriori, yang merupakan algoritma yang umum digunakan dalam penggalian kaidah asosiasi. Untuk penggalian kaidah asosiasi obyektif tunggal hanya digunakan pengukuran *support count* saja. Sedangkan untuk penggalian kaidah asosiasi multi obyektif digunakan tiga pengukuran yaitu nilai *support count, comprehensibility* dan *interestingness*.

HASIL DAN PEMBAHASAN

Dataset yang digunakan sebagai bahan uji coba adalah *dataset retail* dari sebuah *supermarket*, diambil dari http://fimi.cs.helsinki.fi/retail. Dataset terdiri dari 1968 *record* transaksi pembelian dari sebuah *supermarket*. Satu *record* paling sedikit memuat satu item pembelanjaan. Item-item dikonversikan ke dalam angka-angka yang menunjukkan kode barang yang dibeli pelanggan. *Dataset* tersebut selanjutnya dilakukan *clustering* dengan menggunakan metode K-*Means* dan *Fuzzy* C *Means* dengan jumlah *cluster* 2, 3, 4 dan 5. Berikutnya adalah melakukan proses penggalian kaidah asosiasi untuk *dataset* yang tidak di-*cluster* dan data yang di-*cluster*. Untuk uji coba ditentukan nilai *minimum support* adalah 20, 40 dan 60 dan nilai *minimum confidence* 60% dan 75%. Hasil uji coba dengan nilai minimum support 20, 40 dan 60 ditunjukkan berturut-turut pada Tabel 1, Tabel 2 dan Tabel 3.

Pada Tabel 1 dan Tabel 2 ditunjukkan bahwa nilai rata-rata dari data yang di-*cluster*, dengan menggunakan K-*Means* dan *Fuzzy* C *Means*, menghasilkan nilai rata-rata *confidence* yang lebih besar dibandingkan dengan data yang tidak di-*cluster*. Hal ini menunjukkan bahwa proses *clustering* akan menghasilkan kaidah-kaidah asosiasi dengan kualitas yang lebih baik.

Pada tabel yang sama menunjukkan bahwa nilai rata-rata *confidence* yang dihasilkan dari penggalian kaidah asosiasi multi obyektif dengan menggunakan metode *clustering* K-*Means* akan menghasilkan kaidah-kaidah dengan nilai rata-rata *confidence* yang lebih besar pada jumlah *cluster* yang sama dibandingkan jika menggunakan metode *Fuzzy* C *Means*. Nilai *support* yang dianjurkan untuk digunakan adalah 20 dan 40, karena akan menghasilkan nilai *confidence* yang lebih besar dibandingkan dengan nilai support 60 seperti yang ditunjukkan pada Tabel 3, dimana nilai rata-rata *confidence* data yang tidak ter-*cluster* lebih besar dibandingkan dengan data yang ter-*cluster*, baik dengan metode K-Means maupun *Fuzzy* C *Means*. Sesuai dengan tujuan penelitian ini yaitu membandingkan hasil penggalian kaidah asosiasi multi obyektif, dengan menggunakan sampel yang dilakukan proses *clustering* terlebih dahulu akan menghasilkan kaidah-kaidah asosiasi yang lebih baik, hal ini ditunjukkan dengan nilai rata-rata yang diperoleh mempunyai nilai yang lebih besar dibandingkan data yang tidak melalui proses *clustering* terlebih dahulu. Demikian juga dengan perbandingan metode *clustering* yang digunakan yaitu K-*Means* dan

Fuzzy C-*Means*. Metode K-*Means* lebih baik daripada *Fuzzy* C-*Means*, hal ini ditunjukkan dengan nilai rata-rata *confidence* yang dihasilkan dengan menggunakan metode K-*Means* mempunyai nilai yang lebih besar dibandingkan apabila menggunakan metode *Fuzzy* C-*Means*.

SIMPULAN DAN SARAN

Dari hasil uji coba yang sudah dilakukan dapat diambil simpulan sebagai berikut:

1. Pada nilai *minimum suppo*rt 20 dan 40 nilai rata-rata *confidence* dari kaidah yang dihasilkan dari data yang di-*cluster* menunjukkan angka lebih baik daripada data yang tidak di-*cluster*.
2. Pada nilai *minimum support* 20 dan 40 nilai rata-rata *confidence* dari kaidah yang di-*cluster* menggunakan K-*Means* lebih tinggi daripada *Fuzzy* C-*Means* untuk masing-masing *cluster*.

Saran yang dapat diberikan berkaitan dengan penelitian ini untuk pengembangan selanjutnya adalah:

1. Pada penelitian ini algoritma, hanya digunakan untuk penggalian kaidah asosiasi multi obyektif dimensi tunggal, sehingga bisa dikembangkan untuk menyelesaikan permasalahan penggalian kaidah asosiasi multi obyektif dimensi banyak.
2. Menggunakan teknik *sampling* yang lain, misalnya *random sampling* atau *regression-based sampling*.

DAFTAR PUSTAKA

[1] Agrawal R, Imielinski T and Swami T. *Mining* Association Rules between Sets of Items in Large Databases. *Proceedings of ACM SIGMOD International Conference on Management of Data (SIGMOD'93*). 207-216. 1993.

[2] Ghosh A and Nath B. Multi-Objective Rule Mining using Genetic Algorithms. *Information Sciences.* 163: 123-133. 2004.

[3] Han J and Kamber M. *Data Mining: Concept and Techniques.* San Fransisco, CA: Morgan Kaufman Publishers. 2000.

[4] Freitas A. *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* New York: Springer-Verlag. 2002.

[5] MacQueen JB. Some Methods For Classification And Analysis Of Multivariate Observations. *Proc. of 5th Berkeley Symposium on Mathematical Statistic and Probability*. 1: 281-297. 1967.

[6] Bezdek JC. *Pattern Recognition with Fuzzy Objective Function Algorithm.* New York: Plenum Press. 1981.