Pemanfaatan *Clustering* Algoritma McKmeans untuk *Business Intelligence*

Fajrizal[1)] dan Khairani Djahara[2)]
Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Lancang Kuning
Jl. Yos Sudarso KM 08 Rumbai Pekanbaru - Riau, Telp. 0761- 7745164
e-mail: fajrizal@unilak.ac.id[1)], khairani.djahara@unilak.ac.id[2)]

Abstrak

Business Intelligence membutuhkan teknik data mining yang tepat untuk menemukan solusi bisnis yang terbaik, clustering merupakan salah satu cara yang dapat digunakan. Permasalahan clustering adalah memperkirakan jumlah k yang tepat sehingga diperlukan suatu algoritma seperti McKmeans yang dapat mengestimasi k terbaik. Estimasi k terbaik diperoleh dari kestabilan cluster dengan melihat selisih nilai mean terbesar antara MCA (maximum clustering assignments) Index McKmeans dengan MCA Index Random Prototype Baseline. Dataset dresses yang digunakan diperoleh dari UCI Machine Learning dengan menggunakan Nearest Neighbor untuk melengkapi missing value dan blanks dengan 180 data dan 13 atribut. Hasil clustering terbaik diperoleh k=6 dengan hasil identifikasi cluster yang paling diminati adalah dresses dengan style casual; neckline berbentuk *o-neck*; *sleevelength* dengan pilihan *sleeveless* atau *short; waiseline* yang *natural*; dan *pattern type solid. Selain itu, hasil clustering k=6 juga dapat dimanfaatkan untuk penentuan model baju baru di masa mendatang; penentuan material yang digunakan yang terkait dengan harga; serta melihat keterkaitan antara musim dengan penentuan model baju.*

Kata kunci: *Business Intelligence, Clustering, Data Mining, McKmeans*

1. Pendahuluan

Clustering merupakan teknik pengolahan data mining yang termasuk pada kategori *unsupervised learning* yaitu pembelajaran tanpa guru atau tanpa menggunakan label kelas seperti pada klasifikasi. *Clustering* pada algoritma K-Means mencari nilai kemiripan antara data dengan teknik penghitungan jarak salah satunya menggunakan *euclidean distance*, permasalahan yang selalu terjadi adalah bagaimana memperkirakan jumlah *k* terbaik, sehingga suatu evaluasi kestabilan pengulangan *cluster* diperlukan untuk memperkirakan jumlah *cluster* (*k)* yang tepat. Evaluasi kestabilan *clustering* pada [1] dilakukan dengan melihat selisih nilai *mean* terbesar antara MCA (*maximum clustering assignments*) dari *clustering* menggunakan McKmeans dengan MCA dari *random prototype clustering* sehingga diperoleh nilai *k clustering* terbaik.

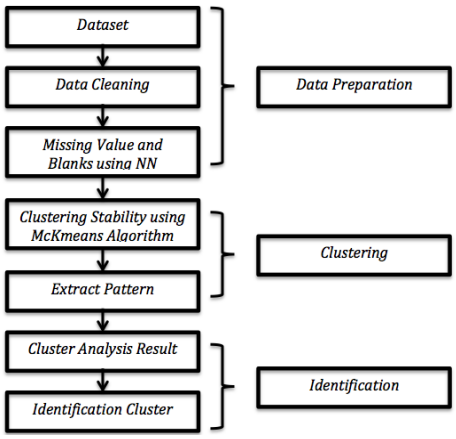
Business Intelligence merupakan analisa dengan menggunakan suatu *software* untuk menemukan solusi bagi *enterprise's user* dengan melakukan proses pengumpulan dan analisa informasi yang ada untuk membuat keputusan yang terbaik dalam menjalankan bisnis. Analisa informasi bagi *enterprise's user* diperoleh dari *retailers* yang langsung berhubungan dengan *customer*. *Retail data mining* [2] akan membantu menemukan pola belanja *customer* dan kecenderungan lainnya yang akan meningkatkan efisiensi dan keuntungan bagi *enterprise's user*. Data mining juga dapat membantu *enterprise's user* untuk menentukan produk terbaru apa [3] yang akan diluncurkan yang diperkirakan *customer* akan tertarik untuk membeli. Penggabungan antara asosiasi dengan klasifikasi menggunakan algoritma ID3 dan C4.5 pernah dilakukan oleh [4] untuk menganalisa pola pasar untuk menentukan manakah barang yang paling banyak terjual pada klasifikasi pagi atau malam hari penjualan. Pada penentuan pola pasar, *Customer Relationship Management* (CRM) perlu diperhatikan untuk menentukan teknik data mining yang tepat seperti hasil yang dipaparkan oleh [5] pada dimensi *customer identification* teknik data mining yang tepat secara berututan adalah *classifcation; clustering; regression* dimana *classification* dan *clustering* memilik selisih satu penelitian.

Perpaduan antara *clustering* dan *business intelligence* salah satunya diterapkan oleh [6] pada *customer clustering* untuk menentukan *high-profit; high-value* dan *low-risk customer*. Hasil dari *clustering* ini adalah untuk menentukan beberapa strategi yang akan digunakan untuk meningkatkan level *customer* seperti penjualan silang antara barang yang laku dengan yang tidak. Pada penelitian ini

pemanfaatan kestabilan *clustering* pada algoritma McKmeans [1] bertujuan untuk mendapatkan *k clustering* terbaik untuk menentukan keputusan terbaik dalam menentukan strategi apa yang harus dilakukan untuk meningkatkan keuntungan atau sebagai landasan bagi *enterprise's user* untuk menentukan produk terbaru apa yang akan dilluncurkan yang diperkirakan akan diminati oleh *customer*.

2. Metode Penelitian

Metode penelitian memiliki tiga tahapan yaitu *data preparation; clustering* dan *identification*. Gambar 1 merupakan keseluruhan tahapan metode penelitian yang dilakukan.



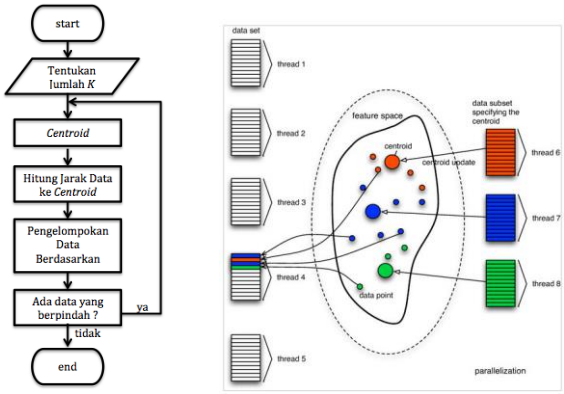
Gambar 1. Proses *Clustering*

2.1. *Data Preparation*

Ada tiga langkah yang dilakukan pada tahapan *data preparation* yaitu pemilihan *dataset; data cleaning ; missing value and blanks using nearest neighbor. Dataset* yang dipilih adalah data yang terkait dengan *business intelligence*. *Data cleaning* adalah proses penghapusan data yang tidak dibutuhkan. *Missing value and blanks using NN Algorithm* adalah proses penggantian nilai yang hilang atau kosong pada *dataset* dengan melihat kedekatan jarak dengan data yang lain menggunakan algoritma *Nearest Neighbor* (*NN*). Jarak yang digunakan adalah *euclidean distance*.

2.2. *Clustering*

Setelah melalui proses *data preparation* artinya data siap untuk diolah ke tahap *clustering*. Pada tahapan *clustering* ada dua langkah yang harus dilewati yaitu *clustering stability using McKmeans Algorithm* dan *Extract Pattern. Clustering stability using McKmeans Algorithm* adalah proses pemanfaatan algoritma McKmeans [1] untuk mencari nilai kestabilan *cluster.* Algoritma McKmeans merupakan perluasan dari Kmeans, dimana McKmeans memanfaatkan keparalelan dan pembagian memori dalam pencarian jarak terdekat setiap data ke *centroid* dan pencarian *centroid* baru untuk menghemat penggunaan waktu pencarian *k cluster* terbaik. Gambar 2 merupakan konsep algoritma Kmeans dan McKmeans.



Gambar 2. *Flowchart* Kmeans (kiri) dan Konsep Dasar McKmeans (kanan) [1]

Gambar 2 menjelaskan bagaimana proses *k-means* dilakukan pada dasarnya. Kemudian melakukan paralesasi menjadi McKmeans dengan memparalelkan terhadap pencarian jarak terkecil antara data ke *centroid* dengan membagi data ke beberapa bagian *thread*; dan penghitungan *update centroid* dimana setiap *centroid* terbentuk memiliki *thread* tersendiri.

Sementara pencarian jumlah *k* yang stabil (terbaik) adalah dengan membandingkan selisih nilai mean terbesar antara *MCA (maximum cluster assignments) index McKmeans* dengan *MCA index random prototype baseline*. *MCA index McKmeans* adalah nilai maksimum *k cluster* yang terbentuk dari algoritma McKmeans, sedangkan *MCA index random protoype baseline* adalah batas bawah *cluster* yang seharusnya terbentuk jika menggunakan data acak sebagai pembentuk *cluster*. Sedangkan *extract pattern* adalah proses pengambilan label kelas yang telah terbentuk pada saat *clustering*.

2.3. *Identification*

Ada dua langkah *identification* yaitu *cluster analysis result* dan *identification cluster*. *Cluster analysis result* adalah tahapan analisa hasil *cluster* dengan memperhatikan pola-pola yang telah terbentuk. *Identification cluster* adalah tahapan pengambilan kesimpulan dari pola-pola yang telah terbentuk tersebut.

3. Hasil dan Pembahasan

Hasil dan pembahasan diperoleh setelah menerapkan proses *clustering* yang merupakan metode penelitian yang dipakai pada penelitian kali ini.

3.1. *Dataset*

Dataset yang digunakan adalah data penjualan baju (*dresses*) diperoleh dari UCI *Machine Learning Dataset*. Data penjualan baju pada awalnya berjumlah 500 data dengan 13 atribut. Kemudian dilakukan *cleaning* data dengan menghapus data yang memiliki penjualan (*sales*) selama 3 bulan berturut-turut (Agustus sampai dengan September 2013) di bawah seribu, sehingga jumlah data menjadi 380 data. Proses *cleaning* selanjutnya adalah menghapus data yang memiliki *missing value* (atau data yang berbentuk tanda tanya) dan *blanks* (atau data kosong) sehingga jumlah data menjadi 180 data. Sedangkan data dengan *missing value* dan *blanks* berada pada satu atribut saja dilakukan *preprocessing* menggunakan algoritma NN (*Nearest Neighbor*) yaitu mencari tetangga terdekat dari data yang memiliki atribut *complete* dengan menggunakan *eculidean distance* sebagai pengukur jaraknya. Setiap atribut memiliki beberapa kategori yang nilainya diganti dengan numerikal. Tabel 1 menjelaskan atribut yang digunakan pada dataset *dresses* berikut kategorinya.

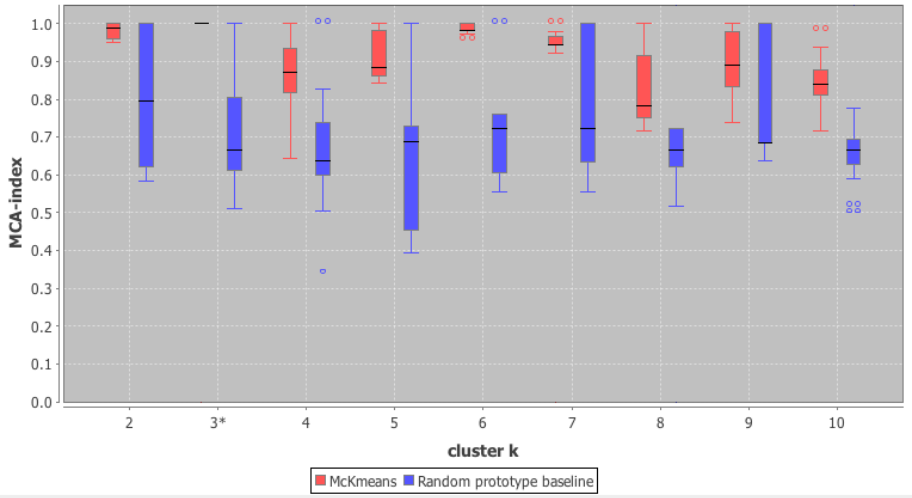
Tabel 1. Atribut *Dresses* dan Kategori

No	Nama Atribut	Kategori Atribut
1	Style	1.Bohemian; 2.Brief; 3.Casual; 4.Cute; 5.Falre; 6.Novelty; 7.Party; 8.Sexy; 9.Vintage; 10.Work
2	Price	1.Low; 2.Average; 3.Medium; 4.High; 5.Very High
3	Rating	1.0; 2.3.5; 3.3.6; 4.3.7; 5.4; 6.4.1; 7.4.2; 8.4.3; 9.4.4; 10.4.5; 11.4.6; 12.4.7; 13.4.8; 14.4.9; 15.5
4	Size	1.S; 2.M; 3.L; 4.XL; 5.Free
5	Season	1.Autumn; 2.Spring; 3.Summer; 4.Winter
6	NeckLine	1.Backless; 2.Boat-neck; 3.Bowneck; 4.Halter; 5.O-neck; 6.Open; 7.Peterpan-collor; 8.Ruffled; 9. Scoop; 10.Slash-neck; 11.Sqare-collor; 12.Sweetheart; 13.Turndown-collor; 14.V-neck
7	SleeveLength	1.Butterfly; 2.Capsleeves; 3.Full; 4.Halfsleeve; 5.Petal; 6. Short; 7.Sleeveless; 8.Threequarter; 9.Turndowncolor
8	WaiseLine	1.Dropped; 2.Empire; 3.Natural; 4.Princess
9	Material	1.Acrylic; 2.Cashmere; 3.Chiffonfabric; 4.Cotton; 5.Knitting; 6.Lace; 7.Linen; 8.Lycra; 9.Microfiber; 10.Milksilk; 11.Mix; 12.Nylon; 13.Other; 14.Polyster; 15.Rayon; 16.Shiffon; 17.Silk; 18.Spandex; 19.Viscos; 20.Wool
10	FabricType	1.Batik; 2.Broadcloth; 3.Chiffon; 4.Corduroy; 5.Dobby; 6.Flannel; 7.Jersey; 8.Lace; 9.Organza; 10.Other; 11.Poplin; 12.Satin; 13.Shiffon; 14.Tulle; 15.Wollen; 16.Worsted
11	Decoration	1. Applique; 2.Beading; 3.Bow; 4.Button; 5.Crystal; 6.Draped;

		7.Embroidary; 8.Flowers; 9.Hollow out; 10. Lace; 11.None; 12.Plain; 13.Pleat; 14. Pockets; 15.Rivet; 16.Ruched; 17.Ruffles; 18. Sashes; 19.Sequined; 20.Tassel; 21.Tiered
12	PatternType	1. Animal; 2.Character; 3.Dot; 4.Floral; 5.Geometrix; 6.Leopard; 7.None; 8.Patchwork; 9.Print; 10.Solid; 11.Splice; 12.Striped
13	Recommendation	1.No; 2.Yes

3.2. Kestabilan *Cluster*

Kestabilan *cluster* diukur dengan memanfaatkan algoritma McKmeans dengan seratus kali percobaan; dengan seribu kali maksimum iterasi; maksimum *cluster* untuk diestimasikan antara *k=2* sampai dengan *k*=10. Kestabilan *cluster* diperoleh pada saat selisih nilai *mean MCA Index McKmeans* lebih besar dari nilai *mean MCA Index Random Prototype Baseline* yang dapat dilihat pada gambar 3.



Gambar 3. *MCA Index McKmeans* dan *MCA Index Random Prototype Baseline*

Tabel 3 menjelaskan bagaimana penentuan *MCA Index McKmeans* dan *MCA Index Random Prototype Baseline* dalam memilih *cluster* terbaik.

Tabel 3. Penentuan *Cluster* Terbaik dengan *MCA Index*

No	*Cluster (k)*	*Mean MCA Index McKmeans*	*Mean MCA Index Random Prototype Baseline*	Selisih
1	*k = 2*	0.994	0.953	0.17
2	*k = 3*	0.998	0.824	0.174
3	*k = 4*	0.859	1	0.141
4	*k = 5*	0.91	1	0.09
5	*k = 6*	0.984	1	0.016
6	*k = 7*	0.965	1	0.035
7	*k = 8*	0.838	0.724	0.114
8	*k = 9*	0.88	0.89	0.01
9	*k = 10*	0.833	1	0.167

Dari tabel 3 di atas estimasi *cluster* terbaik antara *k=2* sampai *k=10* diperoleh bahwa selisih nilai *mean* terbesar dari *MCA Index* berada pada *k=3*.

Sementara untuk menentukan estimasi *k* terbaik diuji dengan seratus kali percobaan kemudian dipilih *k* dengan jumlah terbanyak. Tabel 4 menentukan jumlah *k* yang paling banyak diestimasi sebagai *k* terbaik.

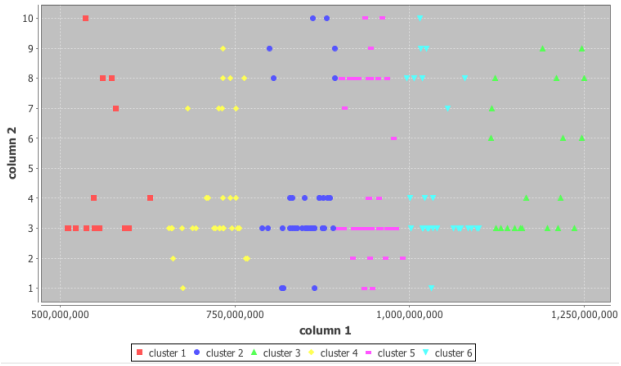
Tabel 4. Penentuan *k* Terbaik yang Paling Banyak di Estimasi

No	*Cluster (k)*	Banyak Estimasi
1	*k = 2*	5
2	*k = 3*	15
3	*k = 4*	10
4	*k = 5*	11
5	*k = 6*	20
6	*k = 7*	18
7	*k = 8*	6
8	*k = 9*	9
9	*k = 10*	6
Total Percobaan		**100**

Dari tabel 4 di atas didapat bahwa *k* terbaik yang paling banyak diestimasi adalah *k=6* sebanyak 20 kali kemunculan dari 100 kali percobaan. Nilai *k=6* ini akan digunakan untuk identifikasi hasil *cluster*.

3.2. Identifikasi Hasil *Cluster*

Hasil *cluster* dengan *k*=6 sebaran diagramnya dilihat pada gambar 4 dan identifikasi hasil *cluster* pada tabel 5.



Gambar 4. Penyebaran *Clustering* dengan *k=6* Terbaik

Tabel 5. Identifkasi Hasil *Cluster* (*k=6*)

No	*Cluster (k=6)*	Distribusi	Identifkasi *Cluster*
1	*Cluster 1*	16 data	Style=Casual; Price=Low; Rating=4.7; Size=Free; Season= Summer; Neckline=O-neck; SleeveLength=Sleeveless; Waiseline=Natural; Material=Chiffonfabric; FabricType=Chiffon; Decoration=Bow; PatternType=Solid; Recommendation=No
2	*Cluster 2*	40 data	Style=Casual; Price=Average; Rating=4.6; Size=Free; Season= Summer; Neckline=O-neck; SleeveLength=Short; Waiseline=Natural; Material=Cotton; FabricType=Chiffon; Decoration=Sashes; PatternType=Solid; Recommendation=No
3	*Cluster 3*	20 data	Style= Casual; Price=Average; Rating=0; Size=Free; Season= Summer; Neckline=O-neck; SleeveLength=Sleeveless; Waiseline=Natural; Material=Polyster; FabricType=Chiffon; Decoration=Hollow out; PatternType=Solid; Recommendation=No
4	*Cluster 4*	35 data	Style=Casual; Price=Average; Rating=4.7; Size=M; Season= Spring, Winter; Neckline=O-neck; SleeveLength=Sleeveless; Waiseline=Natural; Material=Cotton; FabricType=Chiffon; Decoration=Lace; PatternType=Solid; Recommendation=No
5	*Cluster 5*	40 data	Style=Casual; Price=Average; Rating=4.8; Size=Free; Season= Winter; Neckline=O-neck; SleeveLength=Sleeveless; Waiseline=Natural; Material=Polyster; FabricType=Chiffon; Decoration=Lace,Sashes; PatternType=Solid; Recommendation=No
6	*Cluster 6*	29 data	Style= Casual; Price=Average; Rating=4.7; Size=M; Season= Winter; Neckline=O-neck; SleeveLength=Sleeveless; Waiseline=Natural; Material=Polyster; FabricType=Chiffon; Decoration=Lace; PatternType=Solid; Recommendation=No

Dari tabel 5, *cluster* 1, 2 dan 3 berada pada *season* yang sama yaitu *summer*. Pada *summer* kecendrungan pembeli memilih baju dengan *style casual* dan *sleevelength* atau panjang lengan baju *short* (pendek) atau *sleeveless* (tanpa lengan). Perbedaan *cluster* terletak pada harga baju yang dibeli yaitu *low* atau *average* serta *decoration* baju yang dipilih. Harga baju murah (*low*) dikarenakan pemilihan material pembuat baju yaitu menggunakan *chiffon fabric* dengan *decoration bow* sedangkan harga baju menjadi *average* karena material yang digunakan adalah *cotton* atau *polyster* dengan *decoration sashes* atau *hollow out. Cluster* 5,6 berada pada *season* yang sama yaitu *winter*. Pada *winter*, pembeli cenderung membeli baju yang berada pada kisaran harga *average*, hal ini dikarenakan material yang digunakan adalah *polyster*. Perbedaan *cluster* terletak dari *decoration* yang digunakan adalah *lace* atau *sashes*. *Cluster* 4 pengelompokan baju yang bisa digunakan pada musim *spring* maupun *winter*. Pembeli pada *cluster* ini membeli baju dengan kisaran harga *average* dengan material *cotton* dan *decoration lace*.

4. Simpulan

Pemanfaatan algoritma McKmeans dalam penentuan estimasi *cluster* terbaik dapat digunakan untuk membuat suatu keputusan bisnis dalam upaya untuk meningkatkan keuntungan perusahaan. Berdasarkan *clustering* yang dilakukan model baju yang paling diminati oleh pembeli adalah *style casual; neckline* berbentuk *o-neck*; *sleevelength* dengan pilihan *sleeveless* atau *short; waiseline* yang *natural*; dan *pattern type solid*, sehingga dengan informasi ini pembuat keputusan dapat memperhatikan unsur-unsur tersebut sebagai bahan pertimbangan dalam peluncuran model baju terbaru. Selain itu, pembuat keputusan mengetahui kemampuan dari pembeli pada kisaran harga *low* atau *average* sehingga dapat memperhatikan kombinasi material dan dekorasi yang digunakan, seperti kombinasi material *chiffonfabric* dengan *decoration bow* untuk harga murah (*low*) dan kombinasi material *polyster* atau *cotton* dengan *decoration sashes*, *hollow out* atau *lace* untuk harga rata-rata (*average*). Pembuat keputusan juga dapat mempertimbangkan musim (*winter, spring, summer, autumn*) sebagai unsur pembentuk ide dalam pembuatan model baju.

Peluang penelitian yang dapat dilakukan adalah menganalisa penjualan dengan memanfaatkan teknik asosiasi dengan menggunakan algoritma *association rule* maupun *FP-Growth.* Penggabungan teknik antara *clustering* dengan asosiasi dapat juga dilakukan untuk mempertajam keputusan yang dibuat.

Daftar Pustaka

[1] Kraus J M and Kestler H A. A highly efficient multi-core algorithm for clustering extremely large datasets. *BMC Bioinformatics*. 2010; 11(1): 169.

[2] Folorunso O , Ogunde A O, et.al. Data Mining for Business Intelligence in Distribution Chain Analytics. *International Journal of Computer, the Internet and Management*. 2010; 18(1): 15-26.

[3] Gupta G and Agrawal H. Improving Customer Relationship Management using Data Mining. *International Journal of Machine Learning and Computing*. 2012; 2(6): 874-877.

[4] Sabitha B, Amma N G B, et.al. Implementation of Data Mining Technique to Perform Market Analysis. *International Journal of Innovative Research in Computer and Communication Engineering* (IJIRCCE). 2014; 2(11): 7003-7008.

[5] Rodpysh K V, Aghai A. Applying Data Mining in Customer Relationship Management. *International Journal of Information Technology Control and Automation (IJITCA)*. 2012; 2(3): 15-25.

[6] Rajagopal S. Customer Data Clustering using Data Mining Technique. *International Jounal of Database Management Systems* (IJDBMS). 2011; 3(4): 1-11.