KLASIFIKASI DINAMIS DENGAN MODIFIKASI ALGORITMA FUZZY C-MEANS (FCM)

Herlawati[1], Rahmadya Trias Handayanto[2]
[1] Program Studi Sistem Informasi STMIK Nusa Mandiri Jakarta,
[2] Program Studi Teknik Komputer Universitas Islam "45" Bekasi
Email: [1] herlawati@nusamandiri.ac.id, [2] rahmadya_handayanto@unismabekasi.ac.id

ABSTRACT

Aside forecasting, classification is an important process in the data mining field. Nowdays, the classification usually use soft computing algorithms, such as Fuzzy Inference System (FIS), Neural Networks (NNs), and Genetic Algorithms (GAs). Different from K-Means, the fuzzy-based classification is sometimes is said soft clustering. Some dynamic method has been research using K-Means for obtaining the optimal number of cluster. This paper try to implement this method for FCM algoritms because this algorithms run better than K-Means. Similar to Dynamic Clustering using K-Means, for FCM every elements of cluster are counted the distance from the center.

Key Workds : Fuzzy C-Means Clustering (FCM), Cluster Quality, Dynamic Classification

ABSTRAK

Selain peramalan, klasifikasi merupakan salah satu proses penting dalam bidang data mining. Saat ini klasifikasi banyak dilakukan dengan algoritma-algoritma yang berbasis *soft computing* seperti fuzzy, jaringan syaraf tiruan (JST) ataupun algoritma genetik. Berbeda dengan K-Means, klasifikasi berbasis fuzzy yang sering disebut *fuzzy C-Means* (FCM) merupakan klasifikasi halus (*soft clustering*). Beberapa metode dinamis dengan memodifikasi algoritma K-Means telah banyak dilakukan dan terbukti memiliki hasil yang optimal. Tulisan ini bermaksud menerapkan metode dinamis itu pada algoritma FCM mengingat FCM memiliki keunggulan tertentu dibanding K-Means. Seperti pada K-Means, klasifikasi dinamis pada FCM menunjukkan perbaikan pada nilai intra dan inter dimana nilai-nilai tersebut menunjukkan kedekatan antar elemen tiap kluster dan seberapa jauh jarak pisah antar pusat-pusat kluster.

Kata Kunci : *Fuzzy C-Means Clustering* (FCM), Kualitas Kluster, Klasifikasi Dinamis

1. Pendahuluan

Data mining merupakan cabang ilmu komputer yang sangat penting karena dapat meningkatkan nilai tambah suatu perusahaan/organisasi. Sesuai dengan namanya, data mining bermaksud menggali informasi berharga yang berasal dari kumpulan data yang telah ada sehingga membantu pengambil kebijakan dalam upaya meningkatkan performa suatu organisasi. Selain untuk peramalan (*forecasting*), salah satu aktivitas data mining yang tidak kalah penting adalah klasifikasi. Dengan klasifikasi, pengambil kebijakan dapat menerapkan tindakan khusus terhadap kelas-kelas yang telah dihasilkan.

Sebelum dicetuskannya konsep *soft computing*, *K-Means* dan *Decision Tree* telah menjadi algoritma andalan dalam proses klasifikasi. Karena menggunakan algoritma konvensional, sering disebut dengan klasifikasi kasar (*hard clustering*). Tetapi semenjak munculnya algoritma-algoritma *soft computing*, klasifikasi sudah banyak yang menggunakan metode-metode tersebut, dan sering disebut dengan klasifikasi halus (*soft clustering*).

Untuk meningkatkan performa, terkadang beberapa metode digabung menjadi satu sistem yang saling melengkapi. Riset-riset telah banyak dilakukan dalam rangka meningkatkan performa sistem yang dibuat. Jain *et al.* (1998) telah berusaha merangkum riset-riset yang menggabungkan Jaringan Syaraf Tiruan (JST), *Fuzzy Inference System* (FIS), dan algoritma genetik guna menghasilkan kinerja yang lebih baik. Zhang *et al.* (2010) mencoba menggabungkan JST dengan algoritma genetik untuk meningkatkan hasil pada kasus khusus yang rumit, terutama di bidang kimia.

Baik *K-Means*, *Decision Tree*, dan FCM, semua menerapkan jumlah kluster yang tetap. Terkadang hasil klasifikasi kurang baik jika menggunakan jumlah klaster yang tetap. Beberapa metode telah digunakan untuk memperbaiki kinerja klasifikasi dengan mencari jumlah kluster optimal berdasarkan simpangan baku (deviasi) yang dihasilkan pada tiap-tiap kluster. Shafeeq *et al.* (2012) telah mencoba melakukan klasifikasi dinamis pada algoritma K-Means.

Tulisan ini berlanjut dari pembahasan metode klasifikasi dinamis, pembuatan prototipe sistem, pengujian dan pembahasan, hingga diakhiri kesimpulan dan saran. Prototipe menggunakan bahasa pemrograman Matlab, juga dibahas perbandingan antara FCM dengan jumlah kluster tetap dengan yang bervariasi.

2. Bahan Dan Metode

2.1 Bahan

Data yang digunakan dalam penelitian ini adalah data yang dibangkitkan dari fungsi bilangan random menggunakan aplikasi matlab sejumlah 2720 dengan tiga buah variabel, yaitu nilai transaksi, transaksi terakhir dan umur pelanggan. Data harus berukuran besar karena berdasarkan data tersebut akan dicari jumlah kluster optimum.

Penentuan k dinamis merupakan salah satu tahapan praproses guna memberikan gambaran ke pengguna mengenai jumlah k optimum dari data kasar tersebut. Proses berikutnya adalah menentukan titik pusat kluster mana saja yang kurang penting dibanding titik pusat kluster yang lainnya.

2.2 Metode Penelitian

Klasifikasi mengharuskan kita menentukan jumlah kelas yang akan dibentuk. Namun terkadang pemaksaan terhadap jumlah kelas ini menghasilkan klasifikasi yang tidak sesuai dengan harapan. Oleh karena itu perlu mengetahui kualitas klasifikasi yang telah dihasilkan. Kualitas kurang baik diketahui dari jarak antar pusat kluster yang terlampau dekat atau deviasi antar anggota kluster dengan pusatnya yang terlalu besar (Shafeq *et al.*, 2011).

Algoritma dirancang dengan menggunakan bahasa pemrograman Matlab dengan bantuan toolbox dan fungsi-fungsi yang telah tersedia. Penggunaan toolbox selain mempermudah pembuatan prototipe juga dapat digunakan oleh periset lain sebagai alat ukur dalam membandingkan satu metode dengan metode yang lain karena

pengujian satu metode dengan bahasa yang berbeda tentu saja kurang akurat mengingat *compiler* satu dengan yang lain memiliki kecepatan dan keandalan yang berbeda. Versi Matlab yang digunakan adalah versi 7 (Matlab 7). Untuk lebih adil dalam proses klasifikasi, di sini digunakan data acak dengan bantuan fungsi random yang ada di Matlab dengan cara menuliskan fungsi *rand(size)*. Untuk menentukan inter dan intra, digunakan berturut-turut persamaan (7) dan (8) pada kode Matlab setelah proses perhitungan dengan FCM.

Penelitian terdiri dari 2 tahapan, yaitu klasifikasi dengan *fuzzy C-Means* (FCM) dan penentuan hasil klasifikasi untuk beragam jumlah kelas k yang kemudian ditentukan jumlah kelas k yang optimum berdasarkan kualitas klasifikasi. Harga awal diambil dua sebagai syarat klasifikasi, kemudian perhitungan berlanjut dengan k+1 dan seterusnya hingga diperoleh penurunan kualitas inter dan intra yang berarti terdapat nilai k optimum berdasarkan iterasi yang telah terjadi.

2.2.1 Klasifikasi dengan *Fuzzy C-Means* (FCM)

Berikut ini prosedur yang digunakan untuk mencari pusat kluster dengan FCM (Miyamoto, 2008):

1. Menyiapkan data:

   Matriks X yang merupakan data yang akan dicluster, berukuran k x j, dengan k adalah jumlah data yang akan di-cluster dan j adalah jumlah variabel/atribut (kriteria).

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1j} \\ X_{21} & X_{22} & \dots & X_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{kj} \end{bmatrix} \quad \text{.....(1)}$$

2. Menentukan nilai awal :
   a. Jumlah cluster yang akan dibentuk (n >c ≥ 2).
   b. pembobot (w > 1).
   c. Maksimum iterasi (max n).
   d. Kriteria penghentian/*treshold* (ε = nilai positif yang sangat kecil).
   e. Menentukan fungsi obyektif awal (P_0).
3. Membentuk matriks partisi awal U (derajat keanggotaan dalam *cluster*) dengan ukuran k x i; matriks partisi biasanya dibuat acak, dengan k adalah jumlah data yang akan di-*cluster* dan i adalah jumlah *cluster*.

$$\begin{bmatrix} U_{11} & U_{12} & \dots & U_{1i} \\ U_{21} & U_{22} & \dots & U_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ U_{k1} & U_{k2} & \dots & U_{ki} \end{bmatrix} \quad \text{.....(2)}$$

4. Menghitung pusat *cluster* (V) untuk setiap *cluster*, menggunakan rumus :

$$V_{ij} = \frac{\sum_{k=1}^{n} (\mu_{ik})^w x_{kj}}{\sum_{k=1}^{n} (\mu_{ik})^w} \quad \text{.....(3)}$$

   Dengan V_{ij} adalah pusat *cluster* pada *cluster* ke-i dan atribut ke-j. μ_{ik} adalah data partisi (pada matriks U) pada *cluster* ke-i dan data ke-k. X_{kj} adalah data (pada matriks U) pada atribut ke-j dan data ke-k dan w adalah pembobot.
5. Menghitung nilai obyektif (P_n) dengan menggunakan persamaan 4.

$$P_n = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ik})^w (d_{ik})^2 \quad \text{.....(4)}$$

Dengan μ_{ik} adalah data partisi (pada matriks U) pada *cluster* ke-i dan data ke-k. d_{ik} adalah fungsi ukuran jarak untuk jarak Euclidean pada pusat *cluster* ke-i dan data ke-k. w adalah pembobot, dan P_n adalah nilai obyektif pada iterasi ke-n.

6. Perbaiki derajat keanggotaan setiap data pada setiap cluster (perbaiki matriks partisi) menggunakan persamaan 5.

$$\mu_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(w-1)}\right]^{-1} \quad \text{.....(5)}$$

dengan :

$$d_{ik} = d(x_k - v_i) = \left[\sum_{j=1}^{m}(x_{kj} - v_{ij})\right]^{1/2} \quad \text{.....(6)}$$

Dengan μ_{ik} adalah data partisi (pada matriks U) pada pusat *cluster* ke-i dan data ke-k. d_{ik} adalah fungsi ukuran jarak untuk jarak Euclidean pada pusat cluster ke-i dan data ke-k. d_{jk} adalah fungsi ukuran jarak untuk jarak Euclidean pada pusat *cluster* ke-j dan data ke-k. w adalah pembobot, dan X_{kj} adalah data (pada matriks U) pada atribut ke-j dan data ke-k.

7. Menghentikan iterasi jika pusat *cluster* V tidak berubah. Alternatif kriteria penghentian adalah jika perubahan nilai *error* kurang dari *treshold* |Pn - Pn-1| <ε. Alternatif adalah ketika perulangan melebihi maksimum iterasi (n > max n). Jika iterasi belum berhenti, kembali ke langkah 4.
8. Jika iterasi berhenti, ditentukan *cluster* dari tiap-tiap data. Cluster dipilih berdasarkan nilai matriks partisi terbesar.

2.2.2 Menentukan kualitas hasil klasifikasi

Walaupun kualitas klasifikasi sudah cukup baik, terkadang kualitas jumlah kluster juga menentukan pengelompokan yang kita lakukan. Terkadang suatu data lebih cocok dikelompokkan dalam tiga kategori daripada dua kategori.

Penentuan kualitas pertama adalah kekompakkan data antara anggota kluster dengan pusatnya. Sering diistilahkan dengan intra. Kekompakkan di sini ditandai dengan jarak yang cenderung dekat antara anggota kluster dengan pusat klusternya. Perhitungannya dapat menggunakan persamaan 7.

$$\sqrt{\frac{1}{n-1}\sum_{1=1}^{n}(X_i - X_m)^2} \quad \text{.....(7)}$$

Prinsip persamaan 7 adalah mengukur jarak antara titik data $X_i^{(j)}$ dengan pusat kluster c_j. Kemudian dihitung rata-rata jarak tersebut sesuai dengan jumlah datanya. Jarak yang digunakan menggunakan persamaan normal *Euclidean*.

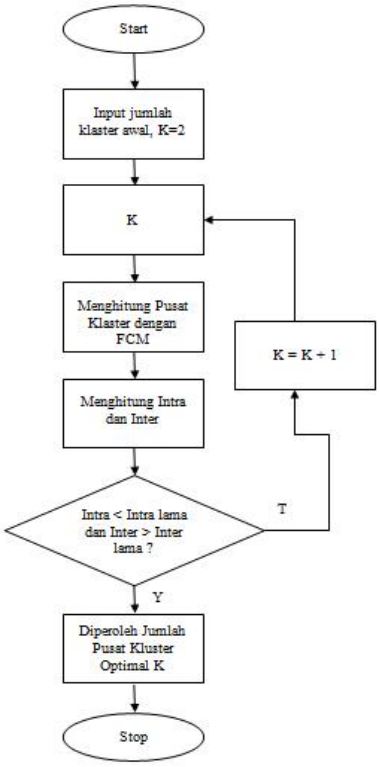
Penentuan kualitas berikutnya adalah jarak minimum antar pusat kluster yang diistilahkan dengan inter. Persamaan yang digunakan adalah persamaan 8.

Inter = min { $||m_k - m_{kk}||$} dengan k= 1,2,..,K-1 dan kk = k+1, k+2, …,K. .....(8)

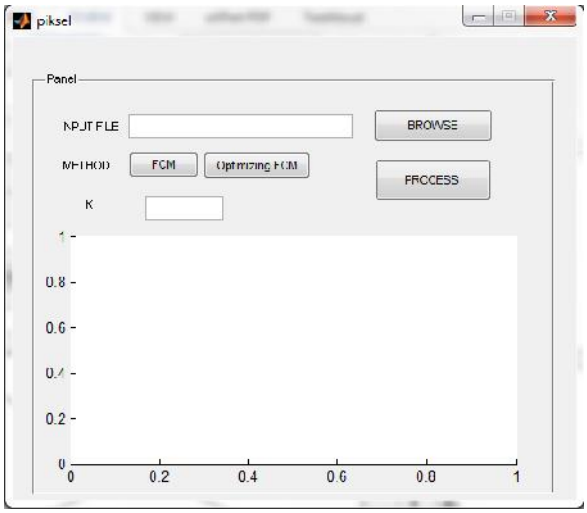
Jadi jika hasil klasifikasi memiliki intra yang kecil dan inter yang besar maka hasil klasifikasinya dikatakan baik. Artinya hasil klasifikasi memiliki tingkat kekompakan yang baik tiap kluster dan jarak yang terpisah jauh antara satu pusat kluster dengan pusat kluster lainnya.

Pada Gambar 1. memperlihatkan algoritma untuk klasifikasi dinamis yang ditambahkan pada algoritma FCM (persamaan 8). Harga kelas awal diambil nilai dua, berarti minimum dua kelas. Jika nilai inter lebih kecil dari nilai inter terdahulu dan nilai intra lebih besar dari terdahulu berarti iterasi berhenti, jika tidak maka jumlah kluster harus ditambah untuk menghasilkan nilai intra dan inter optimal.

Untuk melakukan pengujian dibuat prototipe sistem dengan menggunakan bahasa pemrograman Matlab. Untuk mempermudah pemakaian, digunakan *Graphic User Interface (GUI)* bawaan Matlab (Gambar 2).



Gambar 1. Diagram Alir Klasifikasi Dinamis dengan FCM



Gambar 2. Rancangan Pengujian FCM Dinamis

3. Hasil Dan Pembahasan

Pengujian dilakukan dengan memasukan data sampel pada prototipe sistem hingga diperoleh nilai inter dan intra untuk beragam jumlah kluster k. Antara satu kelompok data dengan kelompok data yang lain akan memiliki nilai optimum yang berbeda, namun demikian pasti akan diperoleh nilai yang optimum. Untuk data yang kami ujikan diperoleh hasil pada tabel 1.

Dari Tabel 1. diperoleh informasi bahwa kualitas klasifikasi dipengaruhi juga oleh variasi dari jumlah kluster. Bertambahnya jumlah kluster mempengaruhi naik turunnya intra dan inter. Di sini kami menggunakan nilai minimal kluster adalah dua, karena jika diambil harga satu, tidak ada fungsi klasifikasi.

Tabel 1. Hasil Pengujian dengan Variasi Jumlah Kluster

Jumlah Kluster (k)	Inter	Intra
2	0.420	0.117
3	0.589	0.435
4	0.721	0.594
5	0.466	0.594
6	0.687	0.569

Sesuai dengan diagram alir pada Gambar 1, bahwa inter dibuat sebesar mungkin sementara intra sekecil mungkin. Tabel 1. memperlihatkan jumlah kluster yang optimal adalah empat karena pada titik tersebut inter menurun sedangkan intra tidak berubah (atau setidaknya menurun). Jadi untuk data tersebut sebaiknya klasifikasi dibuat dalam empat kelas.

Penelitian sebelumnya yang dilakukan oleh Shafeq *et al*. (2010) menggunakan K-Means menuai banyak kritik dibanding dengan metode-metode perbaikannya yaitu K-Midoids dan yang saat ini banyak digunakan oleh periset-periset yang lain yaitu berbasis *fuzzy*, *Fuzzy C-Means* yang berbasis *Soft Clustering* dengan algoritma Soft Computing, seperti yang dilakukan oleh Zhang et al (2010) dengan algoritma Jaringan Syaraf Tiruan (JST). Tentu saja antara JST dengan FCM memiliki keunggulan dan kelebihan masing-masing dari sisi kecepatan dan akurasi.

4. Kesimpulan dan Saran

4.1. Kesimpulan

Berdasarkan pengujian terhadap variasi jumlah kluster diperoleh hasil bahwa antara jumlah kluster yang satu dengan yang lain memiliki kualitas klasifikasi yang berbeda. Hal ini sangat bermanfaat bagi pengambil keputusan apakah membagi data menjadi dua kelas, tiga kelas, dan sebagainya.

4.2. Saran

Untuk penelitian berikutnya dapat menerapkan teknik ini untuk jenis klasifikasi lainnya seperti SVM, algoritma genetik, dan teknik-teknik *soft computing* lainnya.

Daftar Pustaka

Jain, Lakhmi C., N.M. Martin. 1998. Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications. New York: CRC Press, CRC Press LLC.

Miyamoto,Sadaaki, Hidetomo Ichihashi, Katsuhiro Honda. 2008. Algorithms for Fuzzy Clustering. Berlin: Springer-Verlag Berlin Heidelberg.

Shafeeq, Ahamed B M, Hareesha K S. 2012. Dynamic Clustering of Data with Modified K-Means Algorithm. International Conference on Information and Computer Networks (ICICN 2012).

Zhang, Yu, Jingliang Xu, Zhenhong Yuan, Huijuan Xu, Qiang Yu. 2010. Artificial neural network-genetic algorithm based optimization for the immobilization of cellulase on the smart polymer Eudragit L-100. *Bioresource Technology* 101 : 3153–3158.