CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Jurnal Informatika

Jurnal Informatika, Vol. 12, No. 1, Mei 2014, 33-39
ISSN 1411-0105
DOI: 10.9744/informatika.12.1.33-39

KLASTERISASI DAN ANALISIS TRAFIK INTERNET MENGGUNAKAN FUZZY C MEAN DENGAN EKSTRAKSI FITUR DATA

Adi Suryaputra P.[1], Febriliyan Samopa[2], Bekti Cahyo Hindayanto[3]
[1,2,3] Progam Studi Sistem Informasi, Insitut Sepuluh Nopember Surabaya
Kampus **ITS** Keputih, Sukolilo, Surabaya 60111
E-mail: asuryaputra@gmail.com,samopa@gmail.com, bekticahyo@its-sby.edu

Abstrak: Fasilitas internet merupakan salah satu bagian penting dari infrastruktur kampus pada saat ini. Fasilitas internet merupakan penunjang dari kegiatan belajar mengajar yang ada. Bagian penting dari fasilitas internet adalah besarnya bandwidth yang disediakan, dimana seringkali bandwidth tersebut dirasa kurang bagi jurusan tertentu pada jam-jam tertentu terutama jam perkuliahan aktif. Untuk mengatasi hal tersebut perlu adanya sebuah analisa dan klasterisasi terhadap trafik internet di tiap-tiap titik tempat pembagian bandwidth dilakukan sehingga pada akhirnya bisa disediakan informasi yang bisa menjadi pendukung keputusan pemberian bandwidth di tiap-tiap titik yang ada. Salah satu algoritma untuk klasterisasi yang biasa digunakan adalah algoritma Fuzzy C-Mean, dimana pada proses awal sebelum klasterisasi data penggunaan bandwidth internet yang ada dalam satu periode akan dikumpulkan untuk menjadi inputan pada algoritma Fuzzy C-Mean untuk dilakukan pembagian klaster terhadap penggunaan bandwidth yang ada berdasarkan aplikasi yang digunakan dan pemakai jaringan internet. Tetapi dataset awal yang ada pada Fuzzy C Mean belum optimal, sehingga perlu dilakukan suatu optimasi dataset dengan menggunakan ekstraksi fitur data sehingga klaster yang dihasilkan oleh algoritma Fuzzy C Mean memiliki output akurat. Hasil yang akan didapat dari penelitian ini adalah ekstraksi fitur data yang paling tepat untuk melakukan klasterisasi dan analisis trafik internet berdasarkan aplikasi pengguna dan besarnya kapasitas yang dipakai oleh pengguna, dimana informasi hasil klasterisasi tersebut bisa digunakan untuk optimasi bandwidth internet.

Kata kunci: Trafik, Internet, Fuzzy C-Mean, Klasterisasi, ekstraksi, fitur.

Abstract: *Internet facilities is one important part of the infrastructure of the campus at this time. Internet facility is a part of teaching and learning activities. Important part of the internet facility is the internet bandwidth, which is often deemed less bandwidth for certain majors at certain hours of lecture hours especially active. To overcome this there needs to be an analysis and clustering of the internet traffic at each point where the distribution of bandwidth is done so that in the end can provide information that can support decision granting bandwidth at each point there. One algorithm for clustering algorithms used are Fuzzy C-Mean, in which the clustering process before the beginning of the internet bandwidth usage data that exists in one period will be collected to be input to the Fuzzy C-Mean algorithm for the distribution of clusters on the use of existing bandwidth based applications that use the internet and network users. But the initial dataset that of the Fuzzy C Mean is not optimal, so we need some optimization dataset using feature extraction data so that the resulting clusters by Fuzzy C Mean algorithm has the accurate output. Results to be obtained from this study is the extraction of feature data that is most appropriate to perform clustering and analysis of Internet traffic based on user applications and the amount of capacity used by the user, which information the clustering results can be used to optimize internet bandwidth*

Keywords*: Traffic, Internet, Fuzzy C-Mean, Clustering, Extraction, feature*

PENDAHULUAN

Fasilitas internet merupakan salah satu bagian penting dari infrastruktur kampus pada saat ini, dengan adanya internet yang handal maka proses pengolahan data, pencarian materi pendukung perkuliahan, proses belajar mengajar secara online bisa berjalan dengan baik. Bagian penting dari fasilitas internet adalah besarnya bandwidth yang disediakan, tetapi seringkali bandwidth tersebut dirasa kurang bagi jurusan tertentu dan pada jam-jam tertentu terutama pada jam perkuliahan aktif, sedangkan ada jurusan yang merasa bandwidth yang disediakan sudah cukup untuk memenuhi kebutuhan jurusan tersebut bahkan tidak terpakai secara maksimal. Metode yang bisa digunakan untuk mengoptimalkan bandwidth yang ada agar bisa tercapai suatu koneksi internet yang handal dan stabil adalah dengan melakukan klasifikasi terhadap trafik. Secara umum metode klasifikasi trafik dapat digolongkan ke dalam metode *Port-Based*, *Payload-*

Based, Protocol Behavior or Heuristic Based Classification dan Analisa Klasifikasi berdasarkan data statistika. Akan tetapi seiring dengan berkembangnya aplikasi yang menggunakan port yang tidak tetap dan banyaknya aplikasi yang berusaha menghindari metode klasifikasi *port-based* dan *payload-based*, metode yang biasa dilakukan adalah dengan melakukan identifikasi jenis aplikasi yang menggunakan bandwidth yang tersedia. Salah satu metode identifikasi adalah dengan menggunakan algoritma Machine Learning [1]. Contoh penelitian klasifikasi internet yang pernah dilakukan adalah dengan menggunakan Algorima Self Organizing Map (SOM) yang dilakukan di Monash University. Pada penelitian tersebut penggunaan bandwidth internet dibagi menjadi beberapa klaster berdasarkan volume penggunaan bandwidth internet [2]. Penelitian klasifikasi traffik internet yang lain adalah dengan menggunakan ekstraksi fitur-fitur data terhadap data traffic internet yang akan diproses dengan menggunakan algoritma Naïve Bayessian. Ekstraksi fitur data tersebut dilakukan agar klaster yang dihasilkan bisa memiliki anggota yang memiliki fitur yang sama. Pada bagian kesimpulan penelitian ini disampaikan bahwa metode ekstraksi fitur data bisa menghasilkan sebuah performa yang bagus untuk pendeteksian penggunaan traffik internet dengan kompleksitas yang masih sederhana. Penelitian tersebut hanya menggunakan data penggunaan traffik internet yang umum, seperti browser, messenger, FTP, email, dan dns, penggunaan traffik internet yang lain seperti database, game, dan serangan seperti worm dan virus tidak diperhitungkan [3].

Penelitian klasifikasi trafik internet dengan mengambil data penggunaan trafik internet secara menyeluruh dilakukan oleh Chengjie GU, Shunyi ZHANG, dan Xiaozhen XUE, pada bulan april 2011. Pada penelitian tersebut digunakan Algoritma Fuzzy K Mean dengan melakukan perubahan pada Kernel Algoritma. Pada penelitian tersebut dikatakan bahwa Algoritma Fuzzy K Mean tidak dapat melakukan optimasi karakteristik dari data menjadi inputan dan juga pada Fuzzy K Mean semua fitur data dianggap memiliki kontribusi yang sama terhadap klaster yang akan dihasilkan. Hal inilah yang menyebabkan tingkat akurasi dari klasterisasi yang dihasilkan kurang akurat dan masih perlu ditingkatkan akurasinya. Pada kesimpulan penelitian ini dikatakan bahwa masih perlu dilakukan sebuah penelitian untuk menemukan fitur-fitur apa saja yang cocok dan tepat untuk meningkatkan akurasi dari klasifikasi trafik internet [4].

Berdasarkan penelitian-penelitian terdahulu tersebut, bisa dilihat ada sebuah peluang penelitian untuk klasterisasi traffik internet dengan menggunakan algoritma machine learning. Algoritma klasterisasi yang memenuhi ini adalah algoritma Fuzzy C Mean. Salah satu keunggulan algoritma ini adalah jumlah klaster tidak perlu ditentukan dari awal seperti algoritma Fuzzy K Mean, dengan demikian diharapkan agar klaster yang terbentuk dapat merepresentasikan data yang nyata. Akan tetapi Fuzzy C Mean memerlukan sebuah ekstraksi fitur data agar data penggunaan traffik internet yang memiliki korelasi yang sama bisa masuk ke dalam klaster yang sama. Pada Algoritma Fuzzy C Mean, jumlah klaster yang akan dibentuk tidak perlu ditentukan terlebih dahulu, sehingga jumlah klaster yang nantinya terbentuk akan menunjukkan pengelompokkan data yang terjadi. Pada penelitian yang dilakukan oleh Xizhao Wang, Yadong Wang, Lijuan Wang [5] dinyatakan bahwa Algoritma Fuzzy C Mean sangat bergantung pada pemilihan matriks awal untuk proses klasterisasi. Dimana Algoritma Fuzzy C Mean juga bergantung pada fitur bobot yang mempengaruhi jarak antar klaster yang terbentuk. Sehingga pada penelitian yang dilakukan tersebut dinyatakan perlu adanya suatu penyesuaian fitur bobot pada Algoritma Fuzzy C Mean, hal ini dikuatkan oleh penelitian yang dilakukan oleh Ingunn Bergeta, BjZrn-Helge Mevikc, Tormod Næsb pada tahun 2007 [6].

Pada penelitian terbaru di tahun 2012 yang dilakukan oleh Xiaojun LOU, Junying LI, dan Haitao LIU masih dinyatakan bahwa Fuzzy C Mean secara umum memiliki kelemahan untuk hasil output partisi/ klaster untuk dataset yang sama. Pada penelitian ini juga dinyatakan bahwa meskipun sudah banyak penelitian yang dilakukan untuk memperpendek jarak antar klaster dengan klaster pusat, namun penelitian telah ada ini tidak sepenuhnya memperhitungkan distribusi data dan tidak sepenuhnya menggunakan bentuk/karakter dari dataset dalam upaya untuk membuat hasil klaster dari Fuzzy C Mean lebih akurat untuk sebuah dataset yang sama [7]. Setelah melihat penelitian Fuzzy C Mean yang terakhir, dapat dilihat bahwa ada sebuah kontribusi yang dapat dilakukan dalam penelitian klasterisasi bandwidth internet dengan menggunakan algoritma Fuzzy C Mean. Fokus penelitian ini adalah melakukan perhitungan untuk pendistribusian data dengan memperhitungkan karakter dataset yang ada. Dimana proses perhitungan dilakukan dengan menggunakan ekstraksi fitur pada data pengunaan bandwidth internet sebelum proses klasterisasi pada algoritma Fuzzy C Mean diterapkan.

FUZZY C-MEAN

Fuzzy C-Means adalah suatu teknik klasterisasi yang mana keberadaannya tiap-tiap titik data dalam suatu klaster ditentukan oleh derajat keanggotaan. Teknik ini pertama kali diperkenalkan oleh Jim Bezdek pada tahun 1981 [6]. Konsep dari *Fuzzy C-*

Means pertama kali adalah menentukan pusat klaster, yang akan menandai lokasi rata-rata untuk tiap-tiap klaster. Pada kondisi awal, pusat klaster ini masih belum akurat. Tiap-tiap titik data memiliki derajat keanggotaan untuk tiap-tiap klaster. Dengan cara memperbaiki pusat klaster dan derajat keanggotaan tiap-tiap titik data secara berulang, maka akan dapat dilihat bahwa pusat klaster akan bergerak menuju lokasi yang tepat. Perulangan ini didasarkan pada minimasi fungsi obyektif yang menggambarkan jarak dari titik data yang diberikan kepusat *cluster* yang terbobot oleh derajat keanggotaan titik data tersebut. *Output* dari *Fuzzy C-Means* merupakan deretan pusat klaster dan beberapa derajat keanggotaan untuk tiap-tiap titik data. Informasi ini dapat digunakan untuk membangun suatu *fuzzy inference system.*

Langkah-langkah algoritma FCM secara lengkap adalah sebagai berikut (Zimmerman, 1991); (Yan, 1994); (Ross, 2005) [5]:

1. Tentukan :
   a. Matriks X berukuran n x m, dengan n = jumlah data yang akan di cluster; dan m = jumlah variabel (kriteria).
   b. Jumlah variabel klaster (variabel C) yang akan dibentuk, dimana untuk awal mula klaster disetting bernilai lebih besar sama dengan 2 (C ≥ 2)
   c. Tentukan besar variabel pembobot (variabel w), pada fase inisialisasi nilai variabel w diberikan lebih dari 1
   d. Jumlah maksimum iterasi
   e. Kriteria penghentian yang diberikan pada variabel ε (ε = nilai positif yang sangat kecil)
2. Bentuk matriks partisi awal U (derajat keanggotaan dalam klaster); matriks partisi awal biasanya dibuat secara acak

$$U = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_2) & \mu_{1n}(x_n) \\ \mu_{21}(x_1) & \mu_{22}(x_2) & \mu_{2n}(x_n) \\ . & . & . \\ . & . & . \\ . & . & . \\ \mu_{c1}(x_1) & \mu_{c2}(x_2) & \mu_{cn}(x_n) \end{bmatrix}$$

3. Hitung pusat cluster V untuk setiap klaster

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

4. Perbaiki derajat keanggotaan setiap data pada setiap klaster (perbaiki matriks partisi)

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

dengan

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

5. Tentukan kriteria penghentian iterasi, yaitu perubahan matriks partisi pada iterasi sekarang dan iterasi sebelumnya

$$\Delta = \| U^{t} - U^{t-1} \|$$

Apabila $\Delta < \varepsilon$ maka iterasi dihentikan

EKSTRAKSI FITUR DATA

Fitur adalah sebuah perhitungan karakter secara statistika yang dihitung dari berbagai informasi dari sebuah obyek. Menggunakan semua fitur data untuk menjadi parameter dalam sebuah Algoritma *Machine Learning* tidak selalu menjadi pilihan yang tepat, karena tidak semua fitur relevan dengan *Machine Learning* tersebut, untuk itu diperlukan pemilihan dan ekstraksi fitur yang efisien dengan proses klasifikasi, dengan mengabaikan fitur-fitur yang tidak relevan dan redundant, proses ini dinamakan dengan pemilihan fitur atau ekstraksi fitur, dimana proses ini memainkan peranan penting di dalam algoritma *Machine Learning* untuk bisa meningkatkan akurasi dari hasil klasifikasi yang ada dan tidak sekedar memilih fitur apa yang cocok dengan klasifikasi yang diinginkan [8]. Fitur-fitur data yang ada pada penggunaan *bandwidth internet* terlihat pada Tabel 1.

Tabel 1. Contoh Fitur Data Penggunaan Bandwidth Internet

Number	Feature
1	Flow metrics (duration, packet-count, total bytes)
2	Packet inter-arrival time (mean, variance, 1st and 3rd quartiles, median, minimum, maximum...)
3	Size of TCP/IP control fields (mean, variance, 1st and 3rd quartiles, median, minimum, maximum...)
4	Total packets (in each direction and total for flow)
5	Payload size (mean, variance, 1st and 3rd quartiles, median, minimum, maximum...)
6	Effective bandwidth based upon entropy
7	Ranked list of top-ten Fourier-transform components of packet inter-arrival times
8	Numerous TCP-specific values derived from tcptrace
9	Total number of ACK packets seen carrying SACK information, minimum observed segment size

Metode yang bisa dipakai untuk melakukan ekstraksi fitur data antara lain adalah *Correlation-based Feature Selection* (CFS). Metode CFS merupakan bagian evaluasi heuristik yang memperhatikan manfaat fitur individu untuk prediksi kelas bersama-

sama dengan level antar-korelasi di antara mereka. CFS menempatkan skor tinggi sebagai subset data yang mengandung fitur dengan korelasi tinggi dengan kelas tetapi memiliki antar-korelasi rendah satu dengan yang lain. [8]. CFS mengevaluasi sebuah nilai subset dari atribut dengan mempertimbangkan kemampuan prediktif individu masing-masing fitur data dan tingkat redundansinya. Koefisien korelasi tersebut digunakan untuk memperkirakan hubungan antara subset dari atribut dan kelas, serta korelasi antara fitur. Relevansi dari sebuah kelompok fitur bertambah dengan korelasi antara fitur dan kelas fitur dan akan semakin berkurang dengan bertambahnya inter-korelasi. CFS digunakan untuk menentukan subset fitur terbaik dan biasanya dikombinasikan dengan strategi pencarian seperti pemilihan ke depan, eliminasi mundur, dua arah pencarian dan pencarian genetic. Rumus dari Correlation-based Feature selection adalah seperti berikut ini [9]

$$r_{sk} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

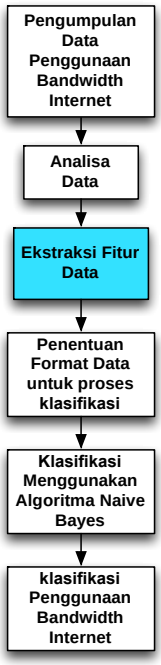
$$CFS = max_{Sk}\left[\frac{r_{cf1}+r_{cf2}+\cdots+r_{cfk}}{\sqrt{k+2(r_{f1f2}+\cdots+r_{fiji}+\ldots+r_{fkf1})}}\right]$$

Variabel r_{sk} adalah korelasi antara subset fitur dijumlahkan dan variabel kelas, K adalah jumlah fitur subset, variabel r_{cf} adalah rata-rata korelasi antara subset fitur variabel kelas, dan variabel r_{ff} adalah rata-rata antar-hubungan antara fitur bagian. Sedangkan CFS digunakan untuk menentukan kriteria-kriteria dari fitur yang akan dilakukan seleksi, tujuannya adalah agar korelasi antar kriteria fitur yang ada bisa maksimal.

Melalui proses ekstraksi fitur data ini kriteria-kriteria dari setiap fitur data penggunaan bandwidth akan ditentukan terlebih dahulu, setelah kriteria ditentukan akan dilakukan proses perhitungan korelasi antar fitur data pengunaan bandwidth internet. Dimana pada proses perhitungan akan didapatkan data penggunaan bandwith mana saja yang secara fitur berkorelasi untuk nantinya akan dikumpulkan menjadi sebuah subset data yang sama, proses ini akan diulang sampai ditemukan subset fitur data yang terbaik. Setelah subset data yang terbaik ditentukan maka dataset yang terbentuk ini akan diproses untuk klasteriasai dengan Algoritma Fuzzy C Mean, dengan adanya dataset yang sudah dilakukan ekstraksi fitur klasterisasi yang dihasilkan akan akurat. Pada proses ini terlihat bahwa ekstraksi fitur data akan melakukan pra proses di Algoritma Fuzzy C Mean, dimana pra proses ini digunakan mengatasi kekurangan dari Algoritma Fuzzy C Mean dalam menentukan fitur bobot pada saat menentukan matriks awal klaster yang akan terbentuk.

METODOLOGI PENELITIAN

Metodologi penelitian yang digunakan terlihat pada Gambar 1. Kontribusi dari penulis terlihat pada gambar yang berwarna biru yaitu melakukan ekstraksi fitur data pada dataset yang nantinya akan diimplementasikan pada algoritma fuzzy c-mean untuk mendapatkan cluster dari data penggunaan bandwidth internet berdasarkan penggunaan aplikasi pada perangkat lunak RapidMiner



Gambar 1. Diagram Alir Metodologi Penelitian

Kontribusi ilmiah pada penelitian ini adalah melakukan ekstraksi fitur data penggunaan bandwidth internet untuk bisa menjadi dataset awal yang optimal algoritma Fuzzy C Mean agar algoritma ini bisa menghasilkan output yang konsisten, langkah-langkah untuk melakukan optimasi adalah sebagai berikut:

1. Memahami karakteristik data bandwidth internet yang ada beserta fitur-fitur apa saja yang ada di dalam data tersebut
2. Melihat fitur-fitur mana yang kira-kira relevan untuk dilakukan ekstraksi
3. Melakukan ekstraksi fitur data dengan mencari fitur-fitur apa saja yang saling berkorelasi dan bisa menghasilkan klasterisasi yang paling akurat, metode yang akan digunakan untuk melakukan ekstraksi fitur adalam metode Correlation-based Feature Selection (CFS) yang sudah dimodifikasi, dimana pada ekstraksi fitur ini tidak hanya menggunakan port based dan pay-load based yang sudah pernah dilakukan di penelitian-penelitian terdahulu, pada ekstraksi fitur-fitur ini akan dihasilkan sebuah dataset yang anggotanya saling

berkorelasi sehingga membuat jarak antara klaster yang nantinya dihasilkan dengan klaster pusat bisa semakin pendek dan optimal

IMPLEMENTASI PENELITIAN

Penelitian yang dilakukan menggunakan data moore set yang biasa digunakan untuk penelitian bandwidth internet, data tersebut terdiri 244 attribut dan berisi 65036 record data, tahapan implementasi penelitian adalah sebagai berikut:

1. Ekstraksi Dataset Mooreset dengan melakukan implementasi Algoritma CFS dengan menggunakan tools WEKA
2. Implementasi Algoritma klasterisasi Fuzzy C Mean pada dataset yang sudah diekstraksi dengan menggunakan aplikasi RapidMiner
3. Penghitungan akurasi dan uji kinerja algoritma

Ekstraksi Dataset Mooreset

Dataset pada penelitian ini diambil dari http://www.cl.cam.ac.uk/research/srg/netos/nprobe/data/papers/sigmetrics/. Data tersebut juga digunakan oleh Chengjie GU, Shunyi ZHANG, dan Xiaozhen XUE dalam penelitiannya pada bulan april 2011. Dataset yang digunakan pada penelitian ini adalah dataset 10, dataset 10 tersebut terdiri dari 244 atribut dan berisi 65536 record data diekstraksi dengan metode Correlation Feature Selection (CFS) dengan menggunakan aplikasi Weka. Hasil dari ekstraksi ini terdapat tujuh Fitur yang saling berkorelasi. Ketujuh fitur ini akan dihapus dari dataset karena fitur-fitur tersebut saling berkorelasi dan bisa menyebabkan akurasi dari algoritma klasterisasi. Ketujuh atribut yang dihapus bisa terlihat pada Gambar 2.

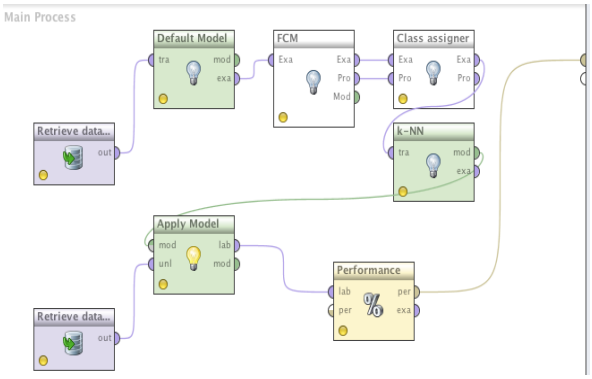
```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 1479
        Merit of best subset found:    0.886

Attribute Subset Evaluator (supervised, Class (nominal): 249 266):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,79,98,101,125,136,169 : 7
                     4
                     91
                     110
                     113
                     137
                     148
                     182
```

Gambar 2. Hasil Ekstraksi Fitur Data

Dari Gambar 2. terlihat bahwa ketujuh atribut yang dideteksi memiliki korelasi dan akan dihapus dari dataset adalah atribut nomor 1, 79, 98, 101, 125, 136, 169.

Implementasi Algoritma klasterisasi Fuzzy C Mean pada RapidMiner

Pada fase ini dataset yang sudah diekstraksi akan diimplementasikan pada algoritma Fuzzy C Mean, implementasi tersebut akan dimodelkan pada aplikasi Rapid Miner seperti pada Gambar 3.



Gambar 3. Model Implementasi Algoritma Fuzzy C Mean pada Rapid Miner

Pada Gambar 3 terlihat bahwa ada 2 buah data yang menjadi inputan untuk algoritma FCM, dimana data tersebut berupa data training dan data lengkap. Data training digunakan oleh Algoritma Fuzzy C Mean mengenali pola data yang akan diklasterisasi, sedangkan data lengkap akan digunakan oleh algoritma tersebut untuk melakukan klasterisasi dan bisa dilakukan perhitungan akurasi dari klasteriasi tersebut. Perhitungan akurasi dari klasterisasi yang dihasilkan oleh Fuzzy C Mean akan dilakukan oleh komponen Performance. Selain akurasi komponen tersebut juga akan menghitung Class Recall dan Class Precisison dari klasterisasi yang sudah dihasilkan. Rumus perhitungan untuk Akurasi, Class Recall dan Class Precision adalah seperti berikut.

$$accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \times 100\%$$

$$precision = \frac{TP}{TP + FP} \times 100\%$$

$$recall = \frac{TP}{TP + FN} \times 100\%$$

True Positive (TP) adalah jumlah dari data yang terklasifikasi di kelas yang benar. False Positive (FP) adalah jumlah data yang dianggap berada di kelas yang salah oleh aplikasi padahal seharusnya data tersebut sudah berada di kelas yang benar. False Negatif adalah jumlah data yang berada di kelas yang salah.

Perhitungan Hasil Akurasi dan Uji Kinerja Algoritma

Setelah proses klasterisasi dilakukan melalui aplikasi Rapid Miner didapatkan perhitungan akurasi, Class Recall, Class Precision. Hasil perhitungan dapat dilihat pada Tabel 2.

Tabel 2. Perbandingan Tingkat Akurasi Algoritma

Algoritma	Akurasi
Fuzzy C-Mean	83.73%
Fuzzy C-Mean dengan ekstraksi Fitur Data	88.10 %

Tabel 3. Perbandingan Class Recall Kedua Algoritma

Class	Fuzz C Mean	Fuzzy C Mean + Ekstraksi Fitur Data
WWW	99.82 %	94.72 %
P2P	18.58 %	12.98 %
MAIL	0 %	80.37 %
SERVICE	0 %	0 %
FTP-PASSIVE	0 %	0 %
ATTACK	0 %	0 %
IOTERACTIVE	0 %	0 %
DATABASE	0 %	0 %
FTP-CONTROL	0 %	35.08 %
FTP-DATA	0 %	55.07 %
GAMES	0 %	0 %

Tabel 4. Perbandingan Class Precision Kedua Algoritma

Class	Fuzz C Mean	Fuzzy C Mean + Ekstraksi Fitur Data
WWW	83.92%	92.37 %
P2P	41.13%	75.70 %
MAIL	0 %	68.44 %
SERVICE	0 %	0 %
FTP-PASSIVE	0 %	0 %
ATTACK	0 %	0 %
IOTERACTIVE	0 %	0 %
DATABASE	0 %	0 %
FTP-CONTROL	0 %	4.92 %
FTP-DATA	0 %	41.79 %
GAMES	0 %	0 %

KESIMPULAN DAN SARAN

Kesimpulan

Kesimpulan yang dapat diambil adalah sebagai berikut:

1. Algoritma Fuzzy C Mean bisa digunakan untuk melakukan klasterisasi bandwidth internet
2. Hasil dari algoritma Fuzzy C Mean dengan ekstraksi fitur data bisa digunakan untuk analisa bandwidth internet karena akurasi yang dihasilkan adalah 88.10%
3. Penggunaan ekstraksi fitur data dengan menggunakan Correlation Feature Selection (CFS) terbukti bisa meningkatkan akurasi dalam menentukan klasterisasi dari algoritma Fuzzy C-Mean.

Saran Pengembangan

Berikut adalah saran pengembangan terhadap penelitian yang akan datang:

1. Metode ekstraksi fitur data yang lain bisa diimplementasikan di algoritma Fuzzy C Mean untuk diuji apakah bisa meningkatkan tingkat akurasi dari algoritma Fuzzy C Mean.
2. Dapat dilakukan uji coba dengan melakukan modifikasi algoritma Fuzzy C Mean dan ekstraksi fitur data untuk meningkatkan class precision dan class recall.

DAFTAR PUSTAKA

1. Mohd Babiker Abuagla, Dr. Sulaiman bin Mohd Nor. Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization, *International Journal of Computer Science and Security*, Volume 3, Issue 2, 2009
2. Xiaozhe Wanga, Ajith Abrahamb, Kate A. Smitha. Intelligent web traffic mining and analysis, *Journal of Network and Computer Applications* 28 (2005): 147–165
3. Junghun Parka, Hsiao-Rong Tyanb, and C.-C. Jay Kuoa, *Internet Traffic Classification For Scalable Qos Provision*, IEEE Multimedia Conference and Expo, 2006
4. Chengjie GU, Shunyi ZHANG, Xiaozhen XUE., Internet Traffic Classification based on Fuzzy Kernel K-means Clustering models, *International Journal of Advancements in Computing Technology*, (3) 3, April 2011.
5. Xizhao Wang, Yadong Wang, Lijuan Wang, *Improving fuzzy c-means clustering based on feature-weight learning*, Science Direct (2004) 1123–1132
6. Berget Ingunn, Mevik Bjrnhelge, Ns Tormod. New Modifications and Applications of Fuzzy C-means Methodology, *Computational Statistics and Data Analysis*, (52) 5, 2008, pp. 2403-2418.
7. Xiaojun LOU, Junying LI, Haitao LIU, Improved Fuzzy C-means Clustering Algorithm Based on Cluster Density, *Journal of Computational Information Systems*, (8) 2, 2012, pp. 727-737.
8. ZHAO Jing-Jing, HUANG Xiao-Hong, SUN Qion, MA Yan, Real time feature selection in traffic classification, *The Journal of China Universities of Posts and Telecommunication*, 2008.

9. Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram, Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection, International Journal of Information Technology and Knowledge Management, (2) 2, 2010, pp. 271-277
10. Emha Taufiq Luthfi. *Fuzzy C-Means Untuk Clustering Data (Studi Kasus: Data Performance Mengajar Dosen)*, Seminar Nasional Teknologi 2007 (SNT 2007) ISSN: 1978 – 9777.
11. Sueli A. Mingoti, Joab O. Lima. Comparing SOM neural network with Fuzzy c-means,K-means and traditional hierarchical clustering algorithms, *European Journal of Operational Research*, 174, 2006, pp. 1742–1759.
12. Linquan Xie, Ying Wang, Liping Chen, and Guangxue Yue1. *An Anomaly Detection Method Based on Fuzzy C-means Clustering Algorithm*, Proceedings of the Second International Symposium on Networking and Network Security 2-4, April. 2010, pp. 089-092, ISBN 978-952-5726-09-1
13. T. Velmurugan, T, Santhanam, Performance Evaluation of K-Means and Fuzzy C-Means Clustering Algorithm for Statistical Distributions of Input Data Points, *European Journal of Scientific Research*, (46) 3, 2010, pp. 320-330.
14. Liu Jingwei, Xu Meizhi, Kernelized Fuzzy Attribute C-means Clustering Algorithm, Fuzzy Sets and Systems, (159) 18, 2008, pp. 2428-2445.
15. Bao Rong Chang , Hsiu Fen Tsai, *Improving network traffic analysis by foreseeing data-packet-flow with hybrid fuzzy-based model prediction*, Science Direct (2009) 6960–6965, 2004.
16. SUN Mei-feng, CHEN Jing-tao, Research of the traffic characteristics for the real time online traffic classification, *The Journal of China Universities of Posts and Telecommunications*, June 2011
17. Maurizio Dusi, Francesco Gringoli, Luca Salgarelli, Quantifying the accuracy of the ground truth associated with Internet traffic traces, http://www.elsevier.com/locate/comnet, Sciencedirect, November 2011