ALGORITMA PENGELOMPOKAN MENGGUNAKAN SELF-ORGANIZING MAP DAN K-MEANS PADA DATA SUMBER DAYA MANUSIA PROVINSI INDONESIA

Hardika Khusnuliawati
Program Studi Teknik Informatika, Fakultas Teknik,
Universitas Sahid Surakarta
Jl. Adi Sucipto 154, Jajar, Surakarta, 57144, Telp. (0271) 743493, 743494
Email: hardika.khusnulia@gmail.com

Abstrak

Unsupervised data is a data type that will be encountered a lot in real-world problems. Example of unsupervised data is data about the condition of a region based on geographic or demographic information. Unsupervised methods that have been tested and studied are clustering methods.

The combination of the clustering algorithm has been well studied. The combined algorithm that *has been implemented is Self Organizing Map (SOM) and K-Means algorithm where K-Means algorithm is used to clarify the visualization result of SOM algorithm. That combined algorithm is implemented on a dataset of human resource information from 33 provinces in Indonesia with evaluation of clustering experiments using the silhouette algorithm. From the experiment results, it can be seen that the provinces in Indonesia can be grouped based on information owned human resources.*

Keywords: unsupervised data, clustering, Self Organizing Map, K-Means

Pendahuluan

Latar belakang

Suatu informasi tentang kondisi suatu wilayah yang meliputi kondisi geografis maupun demografis memiliki peran yang cukup penting dalam analisis masalah maupun pengambilan keputusan. Melihat kepentingan tersebut, proses penggalian data menjadi metode yang menjanjikan. Pengelompokan data atau *clustering* merupakan salah satu metode penggalian data yang populer hingga saat ini. Analisis pengelompokan data ini telah banyak diterapkan dalam berbagai bidang seperti untuk analisis data medis, analisis pasar, analisis gambar maupun video, dan bidang-bidang lainnya. *Clustering* merupakan penggalian data yang bersifat *unsupervised* yaitu analisis pada data masukan tanpa label sehingga dapat mengetahui suatu pola tersembunyi dari data tersebut (The MathWorks, n.d.). Data yang bersifat *unsupervised* saat ini merupakan jenis data yang banyak ditemui dalam dunia nyata terutama menyangkut data informasi tentang kondisi suatu wilayah.

Berbagai metode *clustering* telah diusulkan antara lain yang banyak digunakan yaitu K-means, *Hierarchical clustering*, dan Self Organizing Map (SOM). Setiap algoritma tersebut memiliki karakteristik masing-masing. Penggabungan beberapa algoritma juga telah dilakukan dan dipelajari. Salah satu algoritma gabungan untuk pengelompokan data yang telah diterapkan yaitu penggabungan algoritma Self Organizing Map dengan algoritma K-Means. Pada (Farsadnia dkk., 2014), penggabungan dari kedua algoritma tersebut diterapkan pada data yang bersifat unsupervised tentang kondisi geografis suati wilayah. Metode K-Means yang diimplementasikan digunakan untuk memperjelas hasil dari pengelompokan data menggunakan algoritma SOM. Pada paper ini, diterapkan penggabungan kedua algoritma tersebut untuk diterapkan pada dataset tentang informasi demografis suatu wilayah yaitu informasi sumber daya manusia dari 33 provinsi sebagai indikator wilayah tertinggal.

Permasalahan

Bagaimana menerapkan metode pengelompokan gabungan antara algoritma SOM dan K-Means pada data sumber daya manusia di Indonesia.

Tujuan Penulisan

Tujuan penulisan dari penelitian ini yaitu menerapkan penggabungan dari algoritma pengelompokan SOM dengan K-Means pada data sumber daya manusia di Indonesia.

Tinjauan Pustaka

Pada penelitian ini, digunakan algoritma S-K sebagai algoritma clustering dengan penggabungkan dua algoritma yaitu Self-Organizing Map (SOM) dan K-Means. Implementasi dari algoritma SOM-KMEANS yang dilakukan terhadap data uji terdiri dari dua tahap yaitu tahap pertama digunakan algoritma SOM kemudian diikuti penggunaan algoritma K-Means.

Algoritma SOM

Algoritma SOM adalah salah satu algoritma *unsupervised* yang dapat diimplementasikan untuk pengelompokan data dengan pendekatan algoritma jaringan saraf atau *neural network* (Farsadnia dkk., 2014). Algoritma SOM memiliki dua *layer* yaitu *input layer* yang dibentuk dari sekumpulan *node* (atau *neuron* yang merupakan unit komputasi) dan *output layer* (kohonen *layer*) yang dibentuk dari *node* pada suatu bidang dua dimensi. Jumlah *output neurons* dari SOM merupakan faktor yang penting untuk dideteksi. Jumlah *output neurons* dapat dipilih dengan menggunakan aturan heuristik yang diusulkan Vesanto (Vesanto & Alhoniemi, 2000) dimana jumlah optimal

output neurons yaitu 5 x 1 dengan nilai N menyatakan jumlah data yang diuji.

Node pada *input layer* dengan *node* pada *output layer* saling terhubung dimana setiap *output node* memiliki vektor bobot yang berasosiasi dengan *input node*. Setiap kali suatu input node dengan vektor fitur X digunakan sebagai inputan dari SOM, maka setiap node pada *output layer* akan saling berkompetisi hinga terpilih satu *node* pemenang. Vektor bobot dari *node* pemenang beserta *node* tetangga diperbarui dengan menggunakan Persamaan 1.

$$w_{ij}(t+1) = w_{ij}(t) \propto (t).X_i(t) - w_{ij}(t) \quad \textbf{(1)}$$

Dimana $w_{ij}(t)$ menyatakan bobot antara node I pada *input layer* dengan *node* j pada *output layer* ketika iterasi ke-*t*. Sedangkan $\propto (t)$ menyatakan *learning rate* yang nilainya berkurang setiap iterasi *t*. Proses pembelajaran dari algoritma SOM berhenti ketika nilai vektor bobot telah stabil atau nilai iterasi yang ditetapkan telah terpenuhi.

Algoritma K-Means

Algoritma K-means merupakan algoritma pengelompokan data yang yang sederhana dan mudah diimplementasikan (Liao dkk., 2013). Pengelompokan data dengan algoritma K-Means dilakukan secara kasar dimana membagi suatu vektor inputan X_i dengan $i = 1, \dots, n$, n menyatakan jumlah data uji ke dalam sejumlah kelompok C_j, dengan $j = 1, \dots, k$, k menyatakan jumlah kelompok yang ditentukan. Kondisi dari C_j memenuhi beberapa syarat antara lain anggota dari C_j tidak boleh himpunan kosong, anggota dari setiap kelompok tidak boleh saling beririsan, dan setiap data uji hanya menjadi anggota dari satu kelompok saja.

Langkah dari algoritma K-Means yang pertama yaitu memilih inisialisasi pusat cluster secara random. Langkah kedua menempatkan setiap data uji ke pusat cluster yang memiliki jarak terdekat dengan data uji tersebut. Langkah ketiga melakukan penghitungan kembali pusat cluster berdasarkan rata-rata dari data uji yang menjadi anggota cluster. Kriteria konvergen dari algoritma K-Means ini dapat ditentukan jika pusat cluster tidak mengalami perubahan atau minimum nilai *mean square error* terpenuhi.

Algoritma Silhouette

Algoritma Silhoutte merupakan algoritma yang digunakan untuk mengevaluasi hasil pengelompokan data dengan mengukur tingkat kesamaan dari setiap data yang berada pada satu kelompok (The MathWorks, n.d.). Persamaan untuk mengukur tingkat kesamaan tersebut ditunjukkan pada Persamaan 2.

$$Si = (bi-ai)/\ max(ai,bi) \quad \textbf{(2)}$$

Dimana *Si* adalah nilai silhouette,*ai* menyatakan rata-rata jarak data ke-*i* terhadap lain pada cluster yang sama dan bi menyatakan rata-rata terkecil dari data ke-*i* terhadap data lain pada *cluster* yang berbeda. Nilai silhouette memiliki rentang dari -1 hingga +1. Semakin tinggi nilai silhouette menunjukkan semakin bagus kualitas *cluster*.

Metode Penelitian

Alur Algoritma

Pada penelitian ini, data uji akan dikelompokan menggunakan algoritma SOM terlebih dahulu. Kemudian dari hasil pembelajaran menggunakan algoritma SOM akan diperoleh visualisasi dari persebaran data pada *node* di *output layer*. Persebaran data

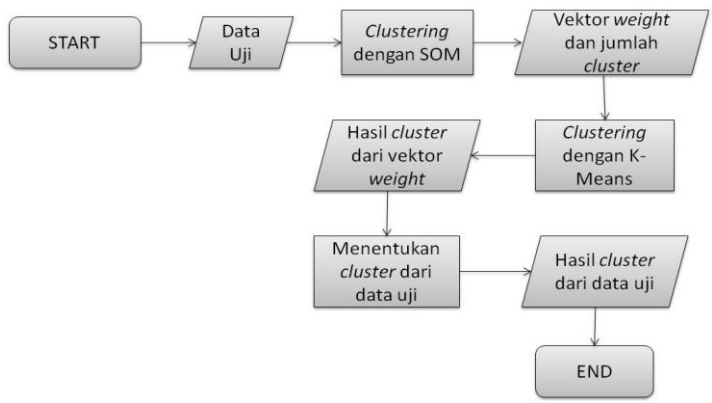
pada *node* di *output layer* tersebut merupakan hasil komputasi yang melibatkan vektor weight dari masing-masing fitur. Hasil vektor *weight* terakhir setelah melalui pembaruan nilai sesuai banyak iterasi pada algoritma SOM, ditetapkan menjadi inputan pada algoritma K-Means (Wang dkk., 2010). Jumlah *cluster* yang ditetapkan untuk mengelompokkan vektor *weight* ditentukan secara manual. Berikutnya hasil dari pengelompokan vektor *weight* oleh algoritma K-Means digunakan sebagai acuan untuk mengelompokan setiap data dari data uji sesuai jumlah *cluster* yang digunakan untuk mengelompokkan vektor *weight*. Diagram alir dari proses tersebut dapat dilihat pada Gambar. **1**.

***Dataset* yang Digunakan**

Simulasi pada penelitian ini dilakukan pada dataset dari Badan Pusat Statistik Indonesia pada tahun 2013 (Indonesia, 2014). Data yang diambil merupakan data sosial dan kependudukan dari 33 provinsi di Indonesia yang menjadi indikator dari kawasan terpencil dan daerah tertinggal. Indikator tersebut merupakan informasi sumber daya manusia meliputi data sebagai berikut.

1. Tingkat pengangguran terbuka
2. Tingkat partisipasi angkatan kerja
3. Indeks keparahan kemiskinan
4. Presentasi penduduk buta huruf
5. Presentasi partisipasi sekolah

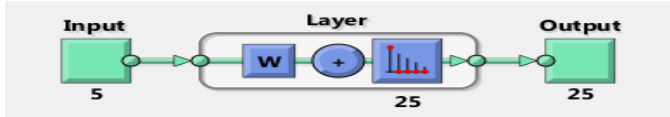
Kelima indikator tersebut merupakan fitur yang digunakan untuk mengelompokkan data uji.



Gambar. 1. Diagram Alir Proses Algortima SOM-KMEANS

Hasil dan Pembahasan

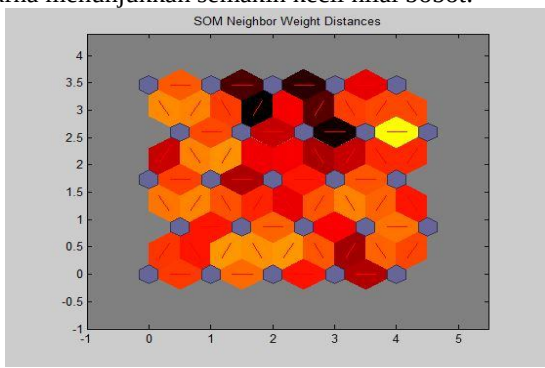
Uji coba terhadap data uji pada tahap pertama dilakukan dengan menggunakan algoritma SOM. Untuk algoritma SOM, jumlah *node* pada *input layer* adalah 5 *node* dan jumlah node pada *output layer* adalah 25 *node* sesuai metode yang diusulkan Vesanto (Vesanto & Alhoniemi, 2000). Gambar 2 menunjukkan model *network* dari hasil pembelajaran algoritma SOM.



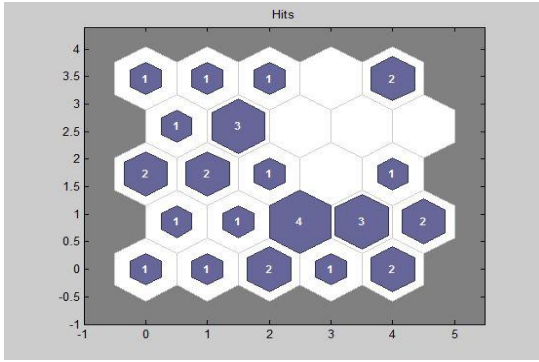
Gambar 2. Model network hasil pembelajaran algoritma SOM

Dari hasil pengelompokan menggunakan SOM diperoleh visualisasi jarak antara *node* dengan tetangganya yang dapat dilihat menggunakan U-Matrix dengan ukuran 5 x 5 bidang segi lima. Ilustrasinya ditunjukkan pada

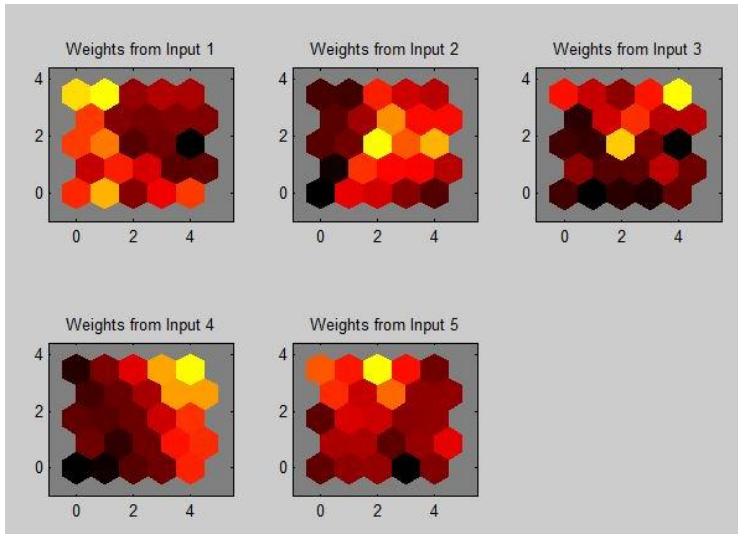
Gambar **3** Hexagon berwarna biru merepresentasikan node sedangkan garis merah menunjukkan koneksi antar node. Semakin gelap warna koneksi antar node menunjukkan jarak node yang semakin dekat begitu pula sebaliknya. Semakin terang warna koneksi antar node maka semakin jauh jarak dari node yang saling bertetangga tersebut. Untuk mengetahui bagaimana data uji tersebar pada *node* di *output layer* maka visualisasinya ditunjukkan pada Gambar 4. Angka yng ditunjukkan pada setiap node merupakan jumlah data uji yang berkaitan dengan *node* tersebut. Sedangkan hasil visualisasi dari persebaran weight pada output layer ditunjukkan pada Gambar 5. Semakin gelap warna menunjukkan semakin tinggi nilai bobot begitu sebaliknya semakin terang warna menunjukkan semakin kecil nilai bobot.



Gambar 3. Visualisasi U-Matrix dari pembelajaran dataset



Gambar 4. Visualisasi banyak data yang berasosiasi dengan suatu node pada *output layer*

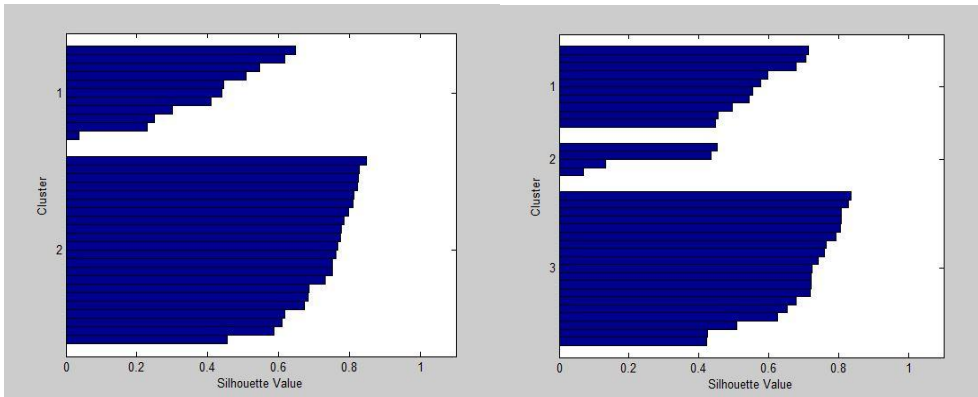


Gambar 5. Visualisasi persebaran bobot dari input layer terhadap output layer

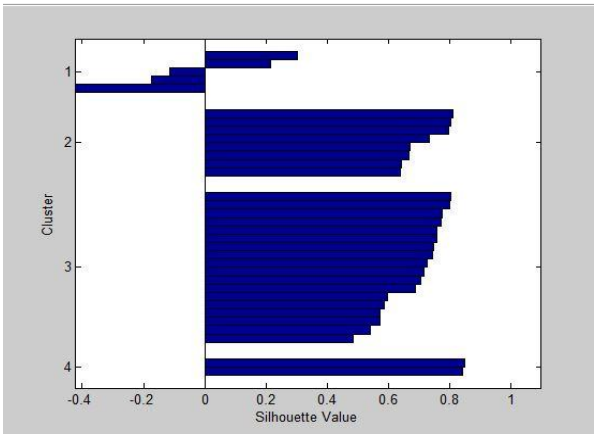
Tahap berikutnya melakukan *clustering* terhadap bobot hasil dari pembelajaran SOM menggunakan algoritma K-Means. Pengujian dilakukan dengan penentuan jumlah *cluster* yaitu 2, 3, dan 4 *cluster*. Dari hasil evaluasi menggunakan algoritma *silhouette*, hasil terbaik diperoleh ketika jumlah cluster yang ditentukan adalah 2 dan 3. Hasil evaluasi menggunakan silhouette value dari masing-masing jumlah cluster ditunjukkan pada Tabel. 1.

Tabel. 2. Hasil evaluasi jumlah *cluster* dengan menggunakan *silhouette value*

Jumlah *cluster*	***Silhoutte value***
2	0.6238
3	0.6116
4	0.5944



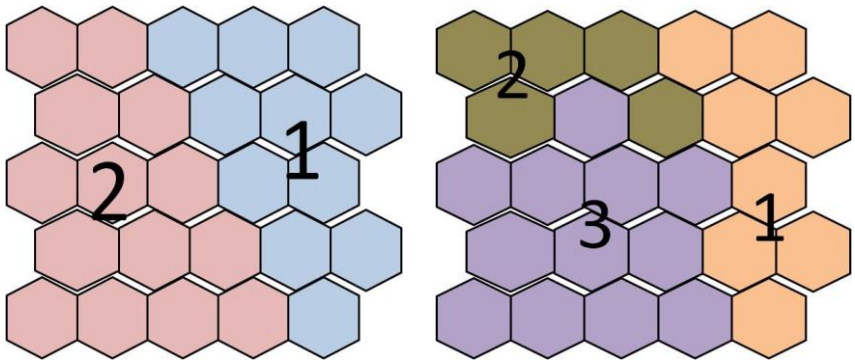
(a) (b)



(c)

Gambar 6. Grafik *silhouette value* (a) jumlah *cluster* = 2 (b) jumlah *cluster* = 3 (c) jumlah *cluster* = 4

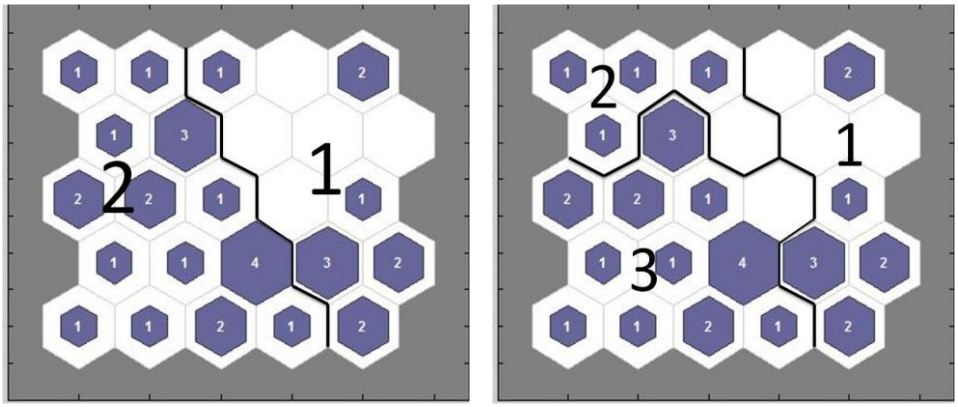
Setelah hasil pengelompokan dari vektor *weight* dengan algoritma K-Means maka diperoleh hasil pengelompokan *node* pada *output layer* sesuai jumlah *cluster* yang telah ditetapkan. Hasil pengelompokan node pada output layer ditunjukkan pada Gambar 7. Telah dijelaskan sebelumnya bahwa setiap data uji saling berkaitan dengan node pada *output layer*. Sesuai Gambar 4, maka dapat diperoleh *cluster* dari setiap data uji. Pembagian cluster untuk setiap data uji ditunjukkan pada Gambar 8.



(a) (b)

Gambar 7. Visualisasi hasil pengelompokan node pada output layer dengan menggunakan vektor weight hasil SOM (a) jumlah cluster = 2 (b) jumlah cluster = 3

Sesuai Gambar 8 dapat diperoleh jumlah provinsi sebagai data uji yang saling berasosiasi dengan node di *output layer*. Untuk pembagian provinsi dengan dua cluster maka diperoleh jumlah provinsi yang berasosiasi dengan *cluster* pertama adalah 11 provinsi sedangkan untuk *cluster* kedua adalah 22 provinsi Sedangkan untuk pembagian .provinsi dengan 3 *cluste*r maka diperoleh 10 provinsi yang masuk *cluster* pertama, 4 provinsi yang masuk *cluster* kedua dan 13 provinsi masuk ke *cluster* ketiga.



(a) (b)

Gambar 8. Visualisasi hasil pengelompokan node pada output layer beserta jumlah data uji yang berasosiasi dengan node (a) jumlah cluster = 2 (b) jumlah cluster = 3

Agar dapat mengetahui karakteristik dari masing-masing cluster, maka dapat digunakan acuan sesuai Gambar 5. Apabila provinsi sebagai data uji dibagi menjadi dua cluster, maka karakteristik untuk cluster pertama yaitu tingkat partisipasi sekolah yang sesuai rata-rata dan angka buta huruf yang jauh berbeda dengan rata-rata. Sedangkan karakteristik untuk cluster kedua yaitu angka buta huruf yang sesuai rata-rata dan tingkat pengangguran terbuka cukup berbeda dengan rata-rata. Sedangkan jika provinsi sebagai data uji dibagi menjadi tiga cluster maka karakteristik dari cluster pertama yaitu tingkat partisipasi sekolah yang sesuai rata-rata dan angka buta huruf yang jauh berbeda dengan rata-rata (memiliki karakteristik yang sama dengan cluster pertama jika data uji dibagi menjadi dua cluster). Karakteristik untuk cluster kedua yaitu tingkat pengangguran terbuka yang paling jauh berbeda dengan rata-rata, sedangkan untuk karakteristik cluster ketiga yaitu indeks keparahan kemisikinan yang dan tingkat buta huruf yang sesuai rata-rata.

Kesimpulan

Metode pengelompokan data yang merupakan gabungan dari SOM dan K-Means dapat diimplementasikan pada permasalahan dunia nyata untuk mengelompokan kondisi suatu wilayah berdasarkan infomasi demografisnya. Dalam paper ini, informasi demografis yang digunakan yaitu kualitas sumber daya manusia sebagai indicator wilayah tertinggal dari 33 provinsi di Indonesia. Algoritma SOM merupakan algoritma yang diimplementasikan pada tahap pertama untuk memperoleh visualisasi dari hasil pengelompokan provinsi di Indonesia. Sedangkan tahap kedua diujikan algoritma K-Means untuk memperjelas hasil pengelompokan data yang kemudian dievaluasi menggunakan algoritma Silhoutte. Dari uji coba yang dilakukan menunjukkan hasil pengelompokan terbaik dari 33 provinsi di Indonesia berdasarkan informasi sumber daya manusia yaitu sejumlah dua hingga tiga kelompok dengan rata-rata nilai silhouette value yang dihasilkan yaitu 0.6238 dan 0.6116.

Sebagai saran, untuk penelitian selanjutnya perlu menyertakan fitur lain yang dapat meningkatkan kualitas hasil pengelompokan data dari informasi sumber daya

manusia 33 provinsi di Indonesia serta membantu proses analisis perbedaan karakterisitik dari setiap masing-masing provinsi yang berhasil dikelompokan. Selain itu, perbaikan pada tahap preprocessing juga dapat dilakukan untuk meningkatkan kualitas hasil pengelompokan maupun analisis.

Daftar Pustaka

Farsadnia, F. dkk., 2014. Identification of homogeneous regions for regionalization of watersheds. *ELsevier: Journal of Hidrology*.

Indonesia, B.P.S.R., 2014. *Badan Pusat Statistik Republik Indonesia*. [Online] Available at: HYPERLINK "http://www.bps.go.id" http://www.bps.go.id .

Liao, K., Liu, G., Xiao, L. & Liu, C., 2013. A sample-based hierarchical adaptive K-means clustering method. *Knowledge-Based Systems*.

The MathWorks, I., n.d. *Mathworks Website*. [Online] Available at: HYPERLINK "http://www.mathworks.com/discovery/unsupervised-learning.html" http://www.mathworks.com/discovery/unsupervised-learning.html .

The MathWorks, I., n.d. *Mathworks website*. [Online] Available at: HYPERLINK "http://www.mathworks.com/help/stats/silhouette.html" http://www.mathworks.com/help/stats/silhouette.html .

Vesanto, J. & Alhoniemi, J., 2000. Clustering of the Self organizing Map. *Neural Network*.

Wang, H.-b., Yang, H.-l., Xu, Z.-j. & Yuan, Z., 2010. A clustering algorithm use SOM and K-Means in Intrusion Detection. *International Conference on E-Business and E-Government*.