Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)
Vol. 1, No. 1, April 2014, hlm. 1-4

PENGEMBANGAN METODE KLASIFIKASI BERDASARKAN K-MEANS DAN LVQ

Dian Eka Ratnawati[1], Marji[2], Lailil Muflikhah[3]

[1,2,3]Program Studi Ilmu Komputer, Universitas Brawijaya
Email: [1]dian_ilkom@ub.ac.id, [2]marji@ub.ac.id, [3]lailil@ub.ac.id

(Naskah masuk: 2 Desember 2013, diterima untuk diterbitkan: 17 Februari 2014)

Abstrak

Pada penelitian ini dikembangkan metode klasifikasi berdasarkan pengelompokan *K-Means* dan LVQ. Metode-metode klasifikasi yang telah ada jika ada data dengan frekuensi kecil cenderung tidak digunakan dalam pengujian kelas, padahal dimungkinkan data tersebut sangat bermanfaat. Langkah untuk melakukan pengelompokan adalah: melakukan pengelompokkan dengan *K-Means*. Pengelompokan terus dilakukan sampai mencapai *threshold* (batasan tertentu). Jika *threshold* sudah dicapai dan pada satu cluster masih terdapat kelas yang berbeda maka dilakukan pembelajaran dengan menggunakan LVQ. Akurasi gabungan *K-Means* dan LVQ lebih baik daripada dengan *K-Means* murni. Untuk akurasi rata-rata tertinggi *K-Means* dan LVQ didapatkan 92%, sedang untuk *K-Means* murni 82%.

Kata kunci: *klasifikasi, pengelompokan, K-Means, LVQ*

Abstract

This research will develop methods of classification based on K-Means clustering. Grouping method used is a combination of K-Means and LVQ. Classification methods that have been there if there is a small frequency data tend to be used in the test class, but it is possible they are very useful. Steps to perform grouping is doing the K-Means clustering. Grouping is continues until it reaches the threshold. If the threshold has been reached and there are cluster of different classes then performed using LVQ learning. Accuracy combined K-Means and LVQ is better than with pure K-Means. For the highest average accuracy of K-Means and LVQ gained 92%, while for the K-Means only 82%.

Keywords: *classification, grouping, K-Means, LVQ*

1. PENDAHULUAN

Klasifikasi merupakan bagian dari data mining. Terdapat beberapa metode klasifikasi antara lain : SLIQ, ID3,C4.5, KNN. Secara umum, metode klasifikasi yang melakukan pembentukan pohon (*decision tree*), sering menghilangkan informasi data yang frekuensinya kecil. Hal ini diterapkan agar memperoleh jumlah aturan yang tidak terlalu banyak. Pada kenyataannya, frekuensi data yang kecil belum tentu data tersebut tidak berguna, bisa jadi kecilnya data disebabkan kesulitan untuk mendapatkannya dibandingkan dengan data yang lain.

Oleh karena itu pada penelitian ini akan dibangun sebuah model klasifikasi yang tetap memperhitungkan semua data, baik data yang frekuensinya kecil maupun data yang frekuensinya besar. Untuk bisa melakukan klasifikasi yang bisa mempertimbangkan kemunculan data dengan frekuensi kecil tersebut, pada penelitian ini akan digabungkan dua metode, yaitu K-Means dan LVQ (Learning Vector Quantization). Penggabungan metode tersebut diharapkan akan dapat saling melengkapi.

Ada beberapa pertimbangan dipergunakannya kedua metode tersebut. K-Means dipakai karena kemudahan dan kemampuannya untuk mengklaster data besar dan data outlier dengan sangat cepat. Tetapi K-Means mengelompokkan datanya tanpa ada pelatihan dahulu, sehingga hasil dari pengelompokkan dengan menggunakan K –Means tidak diketahui kelas dari data. Biasanya dari penelitian–penelitian sebelumnya, untuk pelabelan kelas dari hasil pengelompokan dengan menggunakan K-Means dilakukan secara manual, hal ini menyebabkan akurasi yang rendah(Agusta 2007, Marji 2012). Sehingga pada penelitian ini untuk pelabelan kelas akan dipergunakan Learning Vector Quantization(LVQ).

LVQ adalah suatu metode neural network untuk melakukan pembelajaran pada lapisan kompetitif yang terawasi. Pada LVQ masing-masing unit keluaran mewakili kategori atau kelas tertentu. (Fausett 1994).

Dengan menggabungkan kedua karakteristik yang berbeda tersebut diharapkan baik data kecil atau besar bisa terwakili kelasnya, begitu juga jika datanya sudah *tercluster* dengan *K-Means*, maka pencarian kelas dari data yang diujikan akan lebih cepat, karena hanya akan mencari kelasnya dengan LVQ pada *cluster* tersebut.

2. LANDASAN TEORI

2.1. Klasifikasi

Klasifikasi merupakan proses untuk menyatakan suatu objek ke dalam salah satu kategori yang sudah didefinisikan sebelumnya. Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang labelnya tidak diketahui. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi (Kusnawi 2007).

2.2. Learning Vector Quantization (LVQ)

LVQ merupakan salah satu metode klasifikasi yang *supervised*. Karena LVQ juga merupakan bagian dari *neural network*, maka pembelajarannya dilakukan pada layernya(lapisan kompetitif)

Algoritma LVQ adalah sebagai berikut (Fausett 1994):

0. Tetapkan:
   a. Bobot awal variabel input ke-j menuju ke kelas (cluster) ke-i: Wij, dengan i=1,2,…, K; dan j=1,2,…, m.
   b. Maksimum epoh: MaxEpoh.
   c. Parameter *learning rate*: α.
   d. Pengurangan *learning rate*: Decα.
   e. Minimal *learning rate* yang diperbolehkan: Minα.
1. Masukkan:
   Data input: Xij; dengan i=1,2,…,n; dan j=1,2,…,m.
   Target berupa kelas: Tk;dengan k=1,2,…,n.
2. Tetapkan kondisi awal: epoh=0;
3. Kerjakan jika: (epoh ≤ MaxEpoh) dan (α ≥ Minα)
   a. epoh = epoh+1;
   b. Kerjakan untuk i=1 sampai n
      i. Tentukan J sedemikian hingga ‖Xi-Wj‖ minimum; dengan j=1,2,…,K.
      ii. Perbaiki Wj dengan ketentuan:
      - Jika T = Cj maka:
        Wj = Wj + α (Xi - Wj) (21)
      - Jika T ≠ Cj maka:
        Wj = Wj - α (Xi - Wj) (22)
   c. Kurangi nilai α.(pengurangan α bisa dilakukan dengan: α = α - Decα; atau dengan cara: α =α - α*Decα)

2.3 K-means clustering

K-Means sangat terkenal karena kemudahan dan kemampuannya untuk mengklaster data besar dan data outlier dengan sangat cepat. Kelemahan metode ini memungkinkan bagi setiap data yang termasuk cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster yang lain (Larose 2005).

Algoritma K-Means adalah seperti berikut (Larose 2005):

1. Tentukan k (jumlah cluster) yang ingin dibentuk
2. Bangkitkan k *centroid* (titik pusat cluster) awal secara random
3. Untuk setiap record , temukan pusat cluster terdekat
4. Untuk setiap k cluster , temukan pusat cluster ,dan update lokasi dari setiap pusat cluster dengan nilai centroid yang baru. Pusat cluster diperoleh dengan cara menghitung nilai rata-rata dari data-data yang berada pada cluster yang sama.
5. Kembali ke langkah 3 – 5 sampai konvergen.

2.4. Perhitungan Akurasi dan Error Rate

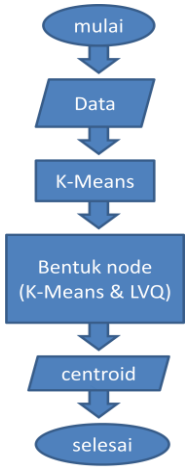
Uji coba dilakukan untuk mencari kelas dari data uji. Pada penelitian ini kebenaran sistem dilakukan dengan menghitung akurasi dan error rate (Lab Data Mining).

$$Akurasi = \frac{\sum \text{prediksi benar}}{\sum \text{total prediksi}} \quad (1)$$

$$error\ rate = \frac{\sum \text{prediksi salah}}{\sum \text{total prediksi}} \quad (2)$$

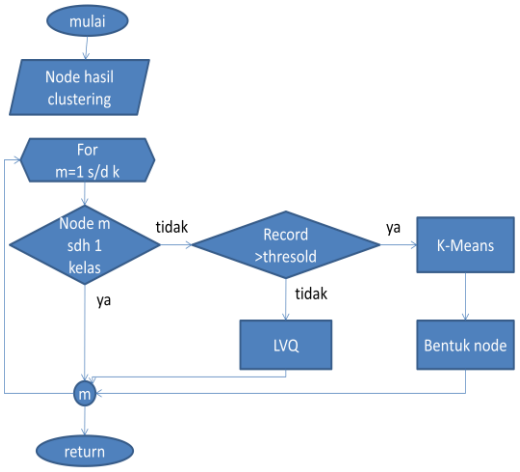
3. PERANCANGAN SISTEM

Pada penelitian ini, secara umum tahapan yang dilakukan adalah sebagai berikut :



Gambar 1. Flowchart Sistem Keseluruhan

Pada Gambar 2 ini adalah flowchart setelah dilakukan klustering dengan K-Means, yang merupakan proses dari bentuk node.



Gambar 2. Flowchart setelah dilakukan clustering dengan *K-Means*

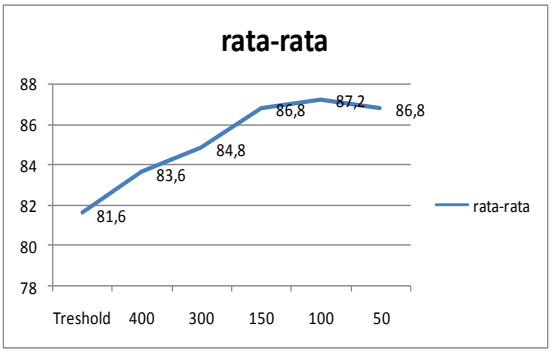
Data Penelitian

Data untuk penelitian ini diambil dari website dengan alamat :http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data. Jumlah data 699 record. Dari semua data tersebut dibagi 2 yakni 630 record sebagai data latih dan 50 record sebagai data uji.

4. HASIL UJI COBA DAN ANALISA

4.1. Uji coba terhadap nilai treshold

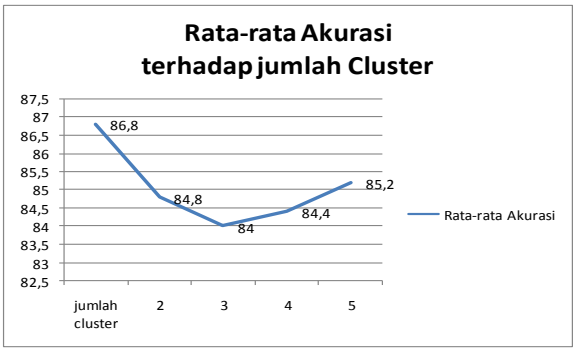
Ujicoba ini dilakukan untuk melihat seberapa besar pengaruh nilai threshold terhadap akurasi dari sistem. Parameter yang digunakan untuk ujicoba ini adalah: Jumlah cluster: 2, Error minimum = 0.000001, Max epoh = 100, Learning rate = 0,05 dan Faktor pengurang = 0,1. Untuk setiap nilai threshold ujicoba dilakukan sebanyak 5 kali, hasil ujicoba bisa dilihat pada Gambar 3.



Gambar 3. Grafik Rata-rata akurasi terhadap nilai threshold

4.2. Uji coba jumlah cluster

Ujicoba ini dilakukan untuk mengetahui jumlah cluster yang sesuai untuk data breast-cancer-wisconsin. Parameter yang digunakan untuk ujicoba ini adalah:Threshold: 50,Error minimum= 0.000001, Max epoh = 100, Learning rate = 0,05 dan Faktor pengurang = 0,1. Hasil ujicoba bisa dilihat pada Gambar 4.

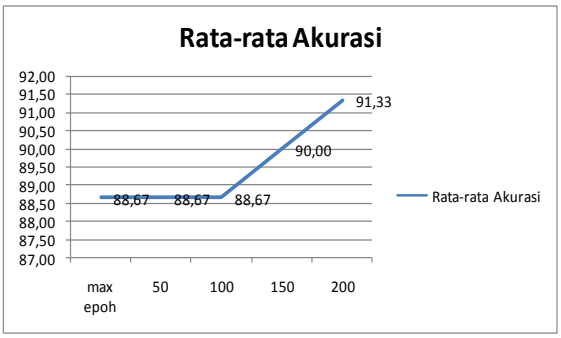


Gambar 4. Grafik akurasi terhadap jumlah cluster

Dari jumlah cluster yang diujikan ,jumlah cluster yang paling baik adalah 2. Hal ini karena jumlah kelas yang ada pada data breast-cancer-wisconsin yang dipergunakan untuk penelitian ini adalah 2 kelas.

4.3. Uji coba max Epoh

Max epoh adalah jumlah iterasi maksimum yang boleh dilakukan selama pelatihan. Iterasi akan dihentikan jika nilai epoh melebihi epoch maksimum. Hasil ujicoba bisa dilihat pada Gambar 5.



Gambar 5. Grafik akurasi terhadap Max Epoh

Dari hasil ujicoba tersebut diketahui bahwa semakin tinggi iterasi , maka akan didapatkan akurasi yang tinggi pula. Hal ini karena semakin banyak iterasi maka bobot yang didapatkan akan semakin baik, bobot semakin menyesuaikan dengan data yang dilatihkan, karena pada setiap iterasi dilakukan update bobot.

4.4. Uji coba terhadap K Means murni

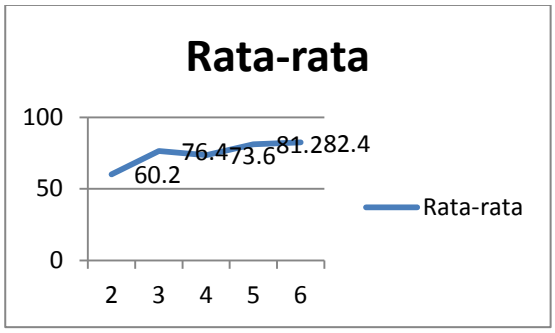
Dari Tabel 1 dan Gambar 6 didapatkan bahwa dengan pengklusteran menggunakan K-Means murni, untuk yang jumlah cluster 2 didapatkan rata-rata akurasi 60.2%, padahal untuk gabungan K-Means dan LVQ dengan jumlah cluster 2 didapatkan hasil yang paling baik.

Pada ujicoba ini, semakin banyak cluster didapatkan akurasi yang semakin baik. Hasil akurasi

yang tinggi pada metode ini 82,4%, sedangkan dengan menggunakan gabungan K-Means dan LVQ didapatkan rata-rata akurasi terbaik 92%. Jadi gabungan K-Means dan LVQ mempunyai akurasi yang lebih tinggi daripada dengan menggunakan K-Means murni.

Tabel 1. Hasil uji coba *K-Means*

cluster	uji 1	uji 2	uji 3	uji 4	uji 5	rata-rata
2	30	80	55	76	60	60,2
3	84	74	72	72	80	76,4
4	72	82	70	66	78	73,6
5	82	78	80	84	82	81,2
6	88	80	76	84	84	82,4



Gambar 6.Grafik rata-rata Akurasi K-Means

5. Daftar Pustaka

AGUSTA, Y. 2007. K-means – penerapan, permasalahan dan metode terkait. *Jurnal Sistem dan Informatika,* 3, 47-60. http://yudiagusta.files.wordpress.com/2008/03/k-means.pdf, , tanggal akses 20 Pebruari 2012

BARAKBHAH, A. R. 2006. Optimasi titik pusat k-means dengan algoritma genetik. *Workshop on Soft Computing, PENS-ITS*, http://lecturer.eepis-its.edu/.../Optimasi%20Titik%20Pusat%20K-..., tanggal Akses 25 Maret 2012

FAYYAD, U. dkk., 1996. *From Data Mining to Knowledge Discovery in Databases*. American Association for Artificial Intelligence.

FAUSETT, L. 1994, *Fundamentals of Neural Network, Architecture, Algorithms and Applications*. Prentice Hall, New Jersey.

IRAWAN, M. I.; SATRIYANTO, E. 2008. Virtual pointer untuk identifikasi isyarat tangan sebagai pengendali gerakan robot secara real-time. *Jurnal Informatika,* 9(1) Mei, 78 – 85.

KANTARDZIC, M. 2003. *Data Mining: Concepts, Models, Methods and Algorithm*. John Wiley & Sons, New York.

KUSNAWI. 2007. *Pengantar Solusi Data Mining.* http://p3m.amikom.ac.id/p3m/56-PENGANTAR-SOLUSI-DATAMINING.pdf, tanggal akses : 18 Maret 2012.

LAROSE, D. T . 2005. *Discovering Knowledge in Data : An Introduction to Data mining*. Wiley-Interscience A John Wiley & Sons,Inc Publication.

LABORATORIUM DATA MINING. *Modul Klasifikasi Decsion Tree*. Jurusan teknik industri Fakultas Teknologi Industri Universitas Islam Indonesia, www.trigunadharma.ac.id/.../Modul%20Klasifikasi%20Decission%20, tanggal akses 25 Maret 2012.

MARJI. 2012. Optimasi anggota kelas dengan menggunakan metode Clustering K Means. *Laporan Penelitian DPP SPP*, Universitas Brawijaya, Malang.

MARTINEZ-CABEZA-DE-VACA-ALAJARIN, J., TOMAS-BALIBRA, L. M. 1999. Marble Slabs Quality Classification System using Texture Recognitiona nd Neural Network Methodology, *ESANN Proceeding*.

MOERTINI, V. S. 2002. *Data Mining Sebagai Solusi Bisnis.* http://home.unpar.ac.id/~integral/Volume7/Integral7No1/idataminingok.pdf, tanggal akses: 19 Maret 2012.