PERBANDINGAN QUANTUM CLUSTERING DAN SUPPORT VECTOR CLUSTERING UNTUK DATA MICROARRAY EXPRESSION YEAST CELL DALAM RUANG SINGULAR VALUE DECOMPOSITION (SVD)

Riwinoto
Program Studi Teknik Informatika, Jurusan Teknik Informatika, Politeknik Negeri Batam, Indonesia
Park Way St, Batam Centre, 29461
E-mail: riwi@polibatam.ac.id

ABSTRAK
Sekarang ini, metode clustering telah diimplementasikan dalam riset DNA. Data dari DNA didapat melalui teknik microarray. Dengan menggunakan metode teknik SVD, dimensi data dikurangi sehingga mempermudah proses komputasi. Dalam paper ini, ditampilkan hasil clustering tanpa pengarahan terhadap gen-gen dari data bakteri ragi dengan menggunakan metode quantum clustering. Sebagai pembanding, dilakukan juga clustering menggunakan metoda Support Vector Clustering. Selain itu juga ditampilkan data hasil clustering menggunakan metode quantum clustering dengan reduksi dimensi pada dimensi 4. Data menunjukkan skor Jackard untuk clustering menggunakan metoda quantum clustering mencapai 0.49625 dengan 4 kluster , SVC menghasilkan 0.24462 dengan 2 kluster. Jadi metoda quantum clustering dengan reduksi dimensi menjadi 4 menghasilkan perfomansi clustering yang lebih baik dibandingkan dengan metoda SVC.

Kata kunci: microarray,quantum clustering, support vector clustering,singular value decomposition,jakcard score

1. PENDAHULUAN

Beberapa penelitian belakangan ini telah membuktikan bahwa penggunaan SVD (*Singular Value Decomposition*) dapat mengekstraksi hasil pemetaan gen pada makhluk hidup dengan cukup menarik. Data gen dari beberapa sampel didapat dari pemetaan matrik sampel dan gen. Dengan menggunakan SVD, data gen tersebut diekstraksi menjadi data gen (*gen space*) dan data sampel (*sample space*). Dalam proses *clustering* data gen, data gen (*gen space*) dipotong sedemikian rupa sehingga data hasil perpotongan (*truncated data*) cukup representatif. Horn dan Axel (2003) menggunakan *quantum clustering* untuk mengidentifikasikan beberapa kluster data gen dari beberapa sampel.

Quantum clustering merupakan model pencarian titik pusat dengan mencari *local minima* dari fungsi gelombang dan potensial schrodinger (Horn dan Axel, 2003). *Quantum clustering* menentukan kluster dengan cara menentukan pusat kluster terlebih dahulu (Kumar dan Behera , 2004). Dengan menentukan pusat kluster, *Quantum Clustering* melakukan *assignment* sebuah titik masuk ke sebuah kluster dengan cara membandingkan titik tersebut dengan pusat kluster. Jika masih dibawah nilai tertentu maka titik tersebut masuk sebagai bagian dari kluster dimana titik kluster berada. Dengan metoda *assignment* titik tersebut, bentuk kluster tidak terlihat. Penggunaan kombinasi dengan metoda *quantum walk* oleh Li et al (2011) juga belum membahas masalah bentuk kluster.

Paper ini merupakan penelitian awal dari pencarian bentuk kluster dari *quantum clustering*. Paper ini membahas bentuk kluster berdasarkan metoda lain yang mampu mengenali bentuk kluster yaitu *Support Vector Clustering* (SVC) oleh Ben-hur et al (2001). Stapor (2006) meneliti penggunaan SVC dalam identifikasi glukomadengan hasil yang menjanjikan. Penelitian ini bertujuan untuk membandingkan perfomansi hasil kluster dari dua metode tersebut. Penelitian diujicobakan terhadap data microarray dalam ruang SVD (Press et al. , 2007) sebagaimana penelitian dari Horn dan Axel (2003).

2. TEORI

2.1 *Singular Value Decomposition* (SVD)

Misalkan terdapat matrik X sampel gen berukuran m x n. m menyatakan jumlah gen dalam 1 sampel dan n menyatakan jumlah sampel yang digunakan. SVD mengekstraksi matrik X menjadi tiga buah matrik yaitu U, $\sum$ dan VT. dimana matrik $\sum$ adalah matrik diagonal (*non square*), dan U,V adalah matrik orthogonal. Dengan mengurutkan elemen *non zero* menurun pada matrik $\sum$ didapatkan aproksimasi rank r yang lebih rendah dengan mengambil nilai element $\sum_j$=0. Oleh karena itu didapatkan matrik Y=U $\sum_r$ VT.

Aproksimasi rank r ke X terbaik dengan menggunakan rumus :

$$S = \sum_i^m \sum_j^n (X_{ij} - Y_{ij})^2 \quad (1)$$

Sehingga jika mengaplikasikan SVD ke matrik X maka didapatkan dua ruang yaitu U dan V. Matrik U mempunyai kolom ortogonal untuk merepresentasikan seluruh gen. Satu gen didefinisikan sebagai satu baris. Matrik V

mempunyai kolom ortogonal untuk merepresentasikan seluruh sampel kasus dengan memotong U atau V. Dengan pemotongan dimensi tersebut memungkinkan komputasi lebih cepat.

2.2 *Quantum Clustering*

Quantum Clustering merupakan metode *clustering* dengan menggunakan mekanisme fisika kuantum. Untuk sebuah titik dicari nilai fungsi gelombang dan potensial schrodinger. Persamaan schrodinger adalah sebagai berikut :

$$H\psi = \left(-\frac{\sigma^2}{2}\nabla^2 + V(x)\right)\psi = E\psi \qquad (2)$$

Dimana H adalah fungsi schrodinger, ψ merupakan probabilitas kepadatan sebuah titik dan V adalah fungsi potensial titik.

Rumus ψ adalah sebagai berikut:

$$\psi(x) = \sum_i e^{\frac{-(x-x_i)^2}{2\sigma^2}} \qquad (3)$$

Rumus V adalah sebagai berikut:

$$V(x) = E + \frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} \qquad (4)$$

Dimana

$$E = -\min\frac{\frac{\sigma^2}{2}\nabla^2\psi}{\psi} \qquad (5)$$

Rumus E didapat untuk mendapatkan nilai V minimal dengan cara mendapatkan turunan pertama sama dengan nol.

Hal ini dilakukan untuk mendapatkan puncak-puncak pada kurva V(x). Puncak-puncak pada kurva V tersebut diidentifikasikan sebagai pusat dari kluster.

Setelah pusat kluster ditemukan, maka dapat ditentukan point mana saja yang termasuk dari kluster tersebut. Algoritma *gradient descent* digunakan untuk mengidentifikasikan kluster dari point-point.

$$y_i(t+\Delta t) = y_i(t) - \eta(t) \mid \nabla V(y_i(t)) \qquad (6)$$

Titik yi(t+Δt) didapatkan melakukan penurunan kurva melawan arah gradient dengan lompatan sejumlah konstanta η. Dengan itu diharapkan ditemukan *local minima* yang dikenali sebagai puncak-puncak kurva.

2.3 *Support Vector Clustering* (SVC)

Support vector clustering merupakan metode *clustering* dengan menggunakan probabilitas kepadatan titik menggunakan kernel jarak pada dimensi tinggi (Ben-hur et al, 2001). Dua tahapan dari SVC adalah pelatihan data untuk menentukan jarak dan pelabelan kluster.

Pada metode ini, data dipetakan ke dalam dimensi yang lebih tinggi dengan kernel jarak. Pada ruang dimensi yang baru, dilakukan kluster data terlihat sebagai bentuk bola. Untuk mendapatkan kluster data yang sesuai, dilakukan pencarian bentuk bola yang minimal (*minimal sphere*) .

Misalkan terdapat {xi} merupakan himpunan bagian dari X sebagai data dari N titik. Pada pemetaan ke dimensi yang lebih tinggi, bola minimal didapat dengan rumus sebagai berikut:

Dimana Φ merupakan fungsi transformasi non linear Xj dari dimensi rendah ke dimensi tinggi.

Sehingga persamaan diatas dapat diubah menjadi

$$\|\Phi(X_j) - a\|^2 \leq R^2 + \zeta_j \qquad (8)$$

Dimana

a merupakan titik tengah bola minimal

R merupakan radius bola minimal

Variabel slack ζ untuk *pinalty term* bentuk bola yang tidak selalu ideal, dimana ζj >= 0.

Untuk dapat menyelesaikan permasalah bola minimal, diperkenalkan Langrangian

$$L = R^2 - \sum_j \left(R^2 + \zeta_j - \left\|\Phi(X_j) - a\right\|^2\right)\beta_j - \sum \zeta_j \mu_j + C\sum\zeta_j \qquad (9)$$

Untuk setiap titik xj dengan ζj =0 merupakan titik yang berada di permukaan atau di dalam bola.

Dimana βj >=0 dan μj>=0 merupakan Langrangian Multiplier yang bisa didapatkan dengan mengubah ke bentuk Dual problem (W):

$$W = \sum_j \Phi(X_j)^2\beta_j - \sum \beta_i\beta_j\Phi(X_i).\Phi(X_j) \qquad (10)$$

Dengan konstrain

$$0 \leq \beta_j \leq C, j = 1, \ldots, N$$

Titik yang berada dipermukaan bola disebut dengan support vector. Syarat titik menjadi support vector adalah 0< βj < C .

Sedangkan titik yang berada di βj=C berada diluar dari boundary *(bounded support vector*, BSV), sedangkan titik lain berada di dalam bola.

Fungsi transormasi Φ ke dimensi tinggi dapat digantikan dengan kernel dalam kasus ini adalah kernel Gaussian sehingga Dual Wolfe menjadi bentuk sebagai berikut:

$$W = \sum_j K(X_j, X_j)\beta_j - \sum_{ij}\beta_i\beta_j K(X_j, X_j)\beta_j \qquad (11)$$

Dengan mengeset turunan dari Langrarian menghasilkan a=$\sum_j \beta_j \Phi(X_j)$

Bola minimal yang telah didapat kemudian dipetakan kembali ke dimensi awal (rendah) dengan menjadi kontur yang secara eksplisit memperlihatkan bentuk kluster. Seluruh titik yang berada pada kontur tersebut diasosiasikan sebagai anggota kluster tersebut.
Ciri titik berada di dalam kontur adalah jarak titik tersebut dengan pusat kluster lebih kecil atau sama dengan radius bola.

$$R^2(X) = \| \Phi(X) - a \|^2 \quad (12)$$

Dengan aturan Wolfe rumus diatas menjadi:

$$R^2(X) = K(X,X) - 2\sum_j \beta_j K(X_j, X) + \sum_{ij} \beta_i \beta_j K(X_j, X_j) \quad (13)$$

Sehingga bentuk kluster dapat dilihat dengan melihat titik –titik support vector dari kluster tersebut.

Untuk menentukan titik masuk ke kluster mana diperlukan pengujian jarak titik tersebut dengan titik yang lain. Misal terdapat titik i dan j maka i dan j termasuk dalam kluster yang sama jika jarak seluruh titik-titik antara i dan j dalam garis lurus lebih kecil atau sama dengan radius bola minima.

Cara diatas mengharuskan dibuatnya matrik ketetanggan antar titik dimana A_{ij}=1 jika titik i dan j terletak dalam 1 kluster dan Aij=0 jika i dan j tidak terletak dalam 1 kluster.

2.4 Algoritma clustering

2.4.1 Algoritma Quantum Clustering

1. Lakukan inisialisasi data
2. Get nilai V,P,E dan dV data
3. Pindah titik menuju local minima (*Gradient Descent*)
4. Jika belum konvergen set nilai konstanta rate pergerakan kembali ke langkah 2

2.4.2 Algoritma Support Vector Clustering

1. Lakukan inisialisasi data
2. Lakukan pencarian nilai beta melalui optimasi persamaan linear dual wolfe dengan konstrain 0< βj < C dan ∑βj=1
3. Lakukan pembuatan matrik ketetanggaan dengan menentukan 3 titik pada garis lurus antara 2 titik yang dicek keterhubungan clusternya.

2.5 *Experimental Setup*

2.5.1 Profil Data

Data yang digunakan adalah data bakteri ragi (yeast cell) dari penelitian Spellman et all (1998). Data tersebut berisi 800 gen yang level mRNA diatur oleh siklus sel. Data gen tersebut merupakan representasi dari 5 kelas. Berdasarkan penelitian Horn dan Axel (2003), ukuran data yang digunakan adalah 798 data gen dengan setiap gen terdiri dari 72 informasi gen.

2.5.2 Proses *clustering*

Matrik 798 x 72 diekstrak dengan menggunakan teknik SVD sehingga menghasilkan matrik U, S dan V. Pengujian *clustering* dilakukan dengan pemotongan pada 4 dimensi dengan memotong kolom matrik U menjadi matrik berukuran 798 x 4. Hal ini dilakukan berdasarkan Horn dan Axel (2003)

Nilai q untuk *Quantum Clustering* dan SVC masing-masing 2.46. Sedangkan untuk nilai C pada SVC adalah 1 untuk memastikan seluruh titik masuk ke kluster.

Perfomansi *clustering* dihitung berdasarkan skor jackard dan jumlah kluster yang dihasilkan apakah sesuai dengan jumlah kluster yang sebenarnya yaitu 5. Skor jackard mempunyai rumus sebagai berikut (Horn dan Axel, 2003):

$$J = \frac{n_{11}}{n_{11} + n_{10} + n_{01}} \quad (14)$$

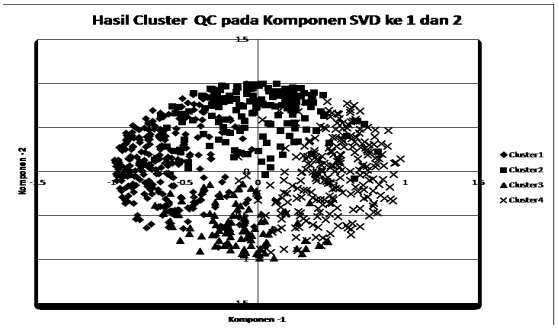
Dimana n_{11} menyatakan jumlah pasangan sampel yang terlihat sama antara hasil *clustering* dengan kluster yang sebenarnya.

$n_{10}+n_{01}$ menyatakan jumlah pasangan sampel yang muncul bersama di satu klasifikasi tapi tidak muncul di klasifikasi yang lain.

2.6 Hasil dan pembahasan

2.6.1 Hasil Quantum Clustering

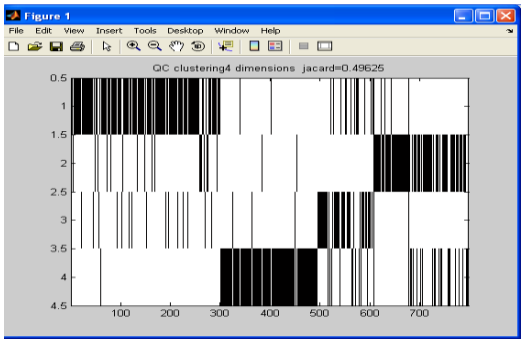
Pengujian dengan menggunakan *Quantum Clustering* dengan memotong dimensi menjadi 4. Pengujian ini menghasilkan kluster sebanyak 4 buah. Dengan parameter q=2.46 dan dimensi=4. Berikut adalah gambah hasil kluster dengan penyajian komponen 1 dan 2 dari 4 dimensi hasil SVD.



Gambar 1. Hasil kluster QC

Skor Jakcard untuk *Quantum Clustering* dimensi 4 adalah 0.49625.

 ISSN: 1907-5022

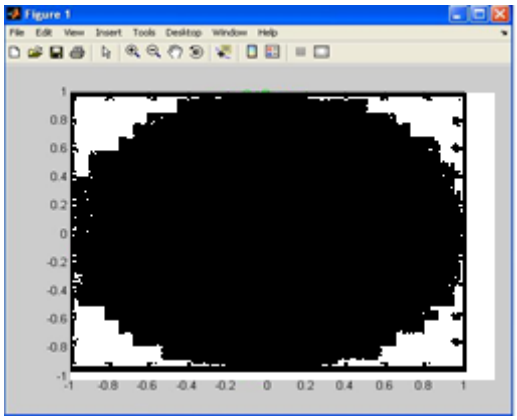


Gambar 2. Skor jackard QC dimensi 4

2.6.2 Hasil Support Vector Clustering

Pengujian dengan Support Vector Clustering dengan dimensi data sampel gen=4 menghasilkan cluster 2 buah

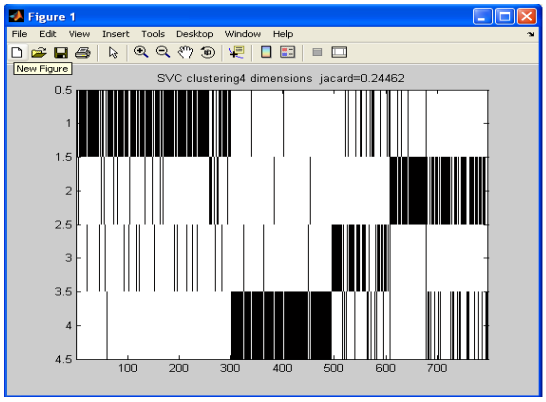
Berikut adalah kontur dari kluster dari SVC



Gambar 3. Hasil Cluster SVC dimensi 4

Sumbu X menyatakan titik pada dimensi 1 matrik U dan sumbu Y menyatakan titik pada dimensi 2 matrik U.

Skor Jackard untuk SVC adalah 0.2246.



Gambar 4. Skor Jackard SVC

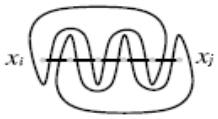
2.6.3 Analisa Perbandingan antar metode clustering

Pada reduksi dimensi 4, Perfomansi *Quantum Clustering* menunjukkan hasil yang lebih baik dibandingkan dengan SVC.

Gambar matrik ketetanggaan dari SVC memperlihatkan gambar yang sangat rapat. Hal ini disebabkan banyak data yang terlibat yaitu 798 gen. Dua kluster yang terbentuk tidak cukup sulit dikenali pada dimensi 2 karena jumlah titik yang masuk cluster sangat dominan terhadap jumlah titik yang masuk pada kluster yang lain. Perfomansi *Suport Vector Clustering* yang menjadi terendah dibandingkan metode yang lain dikarenakan banyak titik-titik berada dalam posisi ambigu pada saat pembentuk matrik ketetanggaan. Keambiguan titik tersebut disebabkan problem pengecekan 2 titik masuk 1 kluster atau tidak.

Misalkan pada saat pengecekan apakah titik A dan B masuk dalam 1 kluster, pengecekan 3 titik yaitu pada posisi 0.25, 0.5 dan 0.75 yang terletak pada garis lurus A ke B ternyata tidak cukup untuk memastikan keterhubungan A dan B terletak dalam 1 kluster. Problem ini mencakup dua jenis:

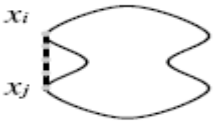
1. Titik-titik sampel ($x_i x_j$) antara dua titik A dan B berada dalam satu kontur padahal A dan B bukan satu kontur. Ilustrasi keadaan ini bisa dilihat pada gambar berikut:



Gambar 5. Problem SVC 1

2. Seluruh titik sampel ($x_i x_j$) antara A dan B berada di luar kontur walapun sebenarnya A dan B berada dalam satu kontur.

Ilustrasi keadaan tersebut bisa dilihat pada gambar berikut:



Gambar 6. Problem SVC 2

Pembentukan matrik ketetanggan pada SVC juga menimbulkan masalah komputasi karena orde transaksinya adalah $O(n^2)$ untuk setiap titik. Jika dihitung seluruh titik maka menjadi $O(n^3)$.

3. KESIMPULAN

Kesimpulan yang bisa ditarik dari pembahasan adalah:

1. Reduksi dimensi hasil SVD sampai menjadi dimensi 4 pada quantum clustering pada data gen menghasilkan clustering yang terbaik mendekati hasil kluster sebenarnya. Hal ini dilihat skor jackard dan jumlah kluster yang dihasilkan
2. Perfomansi yang buruk dari SVC disebabkan problem dalam penentukan titik-titik sampel pada *assignment* titik ke kluster
3. Komputasi SVC sangat tinggi dengan orde transaksi $O(n^3)$.

PUSTAKA

D.Horn, I.Axel(2003). *Novel clustering algorithm for microarray expression data in a truncated svd space*. Bioinformatics 19, pp. 1110–5 Journal Article England

Spellman et *al*., (1998**).** Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.

N.KUMAR, L.BEHERA (2004) , *Visual – Motor Coordination using a Quantum Clustering based Neural Control Scheme*, Neural Processing Letters , Kluwer Academic Publishers

Qiang Li, Yan He, Jing-ping Jiang (2011), *A hybrid classical-quantum clustering algorithm based on quantum walks*, Quantum Inf Process 10:13–26, DOI 10.1007/s11128-010-0169-y

A. Ben-Hur, D.Horn, H.Siegelmann, and V.Vapnik (2001). *Support Vector Clustering* . Journal of Machine Learning Research 2 , p 125-137

Press, WH, Teukolsky, SA, Vetterling, WT, Flannery, BP (2007), "Section 2.6", *Numerical Recipes: The Art of Scientific Computing* (3rd ed.), New York: Cambridge University Press, ISBN 978-0-521-88068-8

K.STAPOR (2006), *Support vector clustering algorithm for identification of glaucomain ophthalmology*, Bulletin of The Polish Academy of Science Technical Science vol.54 No 1