ISSN 2548-9712

PENGKLASTERAN DOKUMEN DENGAN MENGGUNAKAN ALGORITMA SUPPORT VECTOR CLUSTERING

Suhada[1)], A M H Pardede[2)]
[1)]AMIK Tunas Bangsa, Pematang Siantar
Jl. Sudirman, Proklamasi, Siantar Barat, Kota Pematang Siantar, Sumatera Utara
Email : suhada.atb@gmail.com
[2)]STMIK Kaputama
Jl. Veteran No. 4A-9A, Binjai, Sumatera Utara
E-mail : akimmhp@live.com

ABSTRACT

Data processing documents also become an important issue at this time. Along with the increasing amount of data collected and stored in a database increases drastically. This data can come from a variety of sources such as financial applications, Enterprise Resource Management (ERM), Customer Relationship Management (CRM), and others. These data if can be used to support the decision-making process. This paper shows the result of clustering data derived from students' final assignment AMIK Tunas Bangsa Pematangsiantar using Support Vector Clustering method but it also displayed the data using the software RapidMiner clustering results with pengklasteringan time for 11:21 minutes with a gain that took the title in 1708 and classified database 311 who took the title of the classified web programming.

Keywords: Clustering Algorithm, Support Vector Clustering, RapidMiner

ABSTRAK

Pengolahan data dokumen juga menjadi isu penting pada saat ini. Seiring dengan meningkatkan jumlah data yang dikumpulkan dan disimpan dalam suatu database meningkat secara drastis. Data ini dapat berasal dari berbagai macam sumber seperti aplikasi financial, Enterprise Resource Management (ERM), Customer Relationship Management (CRM), dan lain-lain. Data-data tersebut jika di olah dapat digunakan untuk menunjang proses pengambilan keputusan. Dalam paper ini ditampilkan hasil klastering dari data yang diambil dari tugas akhir mahasiswa AMIK Tunas Bangsa Pematangsiantar menggunakan metode Support Vector Clustering selain itu juga ditampilkan data hasil klastering menggunakan software Rapidminer dengan waktu pengklasteringan selama 11:21 menit dengan mendapatkan 1708 yang mengambil judul berklasifikasi database dan 311 yang mengambil judul yang berklasifikasi web programming.

Kata kunci : Algoritma Clustering, Support Vector Clustering, Rapidminer

1. PENDAHULUAN

Sekarang ini begitu banyak data yang terdapat suatu organisasi sehingga menimbulkan kesulitan dalam hal pengklasteran data. Tetapi adanya teknologi komputer dengan jaringan

ISSN 2548-9712

internetnya telah mampu membawa perubahan yang besar dalam kehidupan manusia. Berbagai bidang kehidupan manusia seperti politik, ekonomi, sosial dan budaya telah mengalami perubahan seiring dengan hadirnya teknologi tersebut. Salah satu segi yang secara nyata dapat dirasakan adalah dampak dari teknologi internet dalam penyebaran informasi. Kini berbagai informasi dari berbagai belahan dunia dapat diperoleh secara cepat melalui sebuah PC yang terkoneksi dengan internet. Melaluinya kita mengakses dan mendownload berbagai berita dan informasi baik dalam bentuk artikel teks, gambar maupun suara secara mudah. Dalam waktu yang relatif cepat kita dapat memperoleh berbagai macam informasi dalam bentuk digital yang nantinya dapat disimpan dalam media penyimpanan komputer.. Hal inilah yang menimbulkan penimbunan informasi, baik yang penting ataupun yang tidak penting bercampur menjadi satu sehingga sulit untuk dipisahkan. Karenanya diperlukan pengelompokkan atau penyeleksian terhadap informasi yang diterima. Penyeleksian informasi ini bias dilakukan baik secara manual ataupun secara digital (otomatis) oleh Komputer. Dalam pengklasteringan data, yang sumber berasal dari bidang pendidikan di AMIK Tunas Bangsa Pematangsiantar dipergunakan software rapidminer sehingga hasilnya dapat dilihat pada hasil dan pembahasan..

2. TINJAUAN PUSTAKA

2.1 Support Vector Clustering (SVC)

Support vector clustering merupakan metode clustering dengan menggunakan probabilitas kepadatan titik memakai kernel jarak pada dimensi tinggi [1], dua tahapan dari SVC adalah pelatihan data untuk menentukan jarak dan pelabelan klaster.

Pada metode ini, data dipetakan ke dalam dimensi yang lebih tinggi dengan kernel jarak. Pada ruang dimensi yang baru, dilakukan klaster data terlihat sebagai bentuk bola. Untuk mendapatkan klaster data yang sesuai, dilakukan pencarian bentuk bola yang minimal (minimal sphere). Misalkan terdapat {xi} yang merupakan himpunan bagian dari X sebagai data dari N titik. Pada pemetaan ke dimensi yang lebih tinggi, bola minimal didapat dengan rumus sebagai berikut :

$$||\Phi(\mathbf{x}_j) - \mathbf{a}||^2 \leq R^2 \quad \forall j \ , \quad ...(1)$$

Dimana merupakan fungsi transformasi non linear Xj dari dimensi rendah ke dimensi tinggi. Sehingga persamaan diatas dapat diubah menjadi

$$||\Phi(X_j) - a||^2 \leq R^2 + \xi_j \quad .....(2)$$

Dimana :
a merupakan titik tengah bola minimal
R merupakan radius bola minimal
Variabel slack untuk pinalty term bentuk bola yang tidak selalu ideal, dimana
j >= 0.
Untuk dapat menyelesaikan permasalah bola minimal, diperkenalkan Langrangian

$$L = R^2 - \sum_j \left(R^2 + \zeta_j - \left\|\Phi(X_j) - a\right\|^2\right)\beta_j - \sum \zeta_j \mu_j + C \sum \zeta_j \quad ....(3)$$

Untuk setiap titik xj dengan $j =0 merupakan titik yang berada di permukaan atau di dalam bola. Dimana $j >=0 dan μj>=0 merupakan Langrangian Multiplier yang bisa didapatkan dengan mengubah ke bentuk Dual problem (W):

$$W = \sum_j \Phi(X_j)^2 \beta_j - \sum \beta_i \beta_j \Phi(X_i).\Phi(X_j) \quad .....(4)$$

Dengan konstrain :

$$0 \leq \beta_j \leq C, j = 1, \ldots, N \quad .....(5)$$

ISSN 2548-9712

Titik yang berada dipermukaan bola disebut dengan support vector. Syarat titik menjadi support vector adalah 0< $j < C . Sedangkan titik yang berada di $j=C berada diluar dari boundar (bounded support vector, BSV), sedangkan titik lain berada di dalam bola. Fungsi transormasi $ ke dimensi tinggi dapat digantikan dengan kerne dalam kasus ini adalah kernel Gaussian sehingga Dual Wolfe menjadi bentuk sebagai berikut:

$$L = R^2 - \sum_j \left(R^2 + \zeta_j - \left\|\Phi(X_j) - a\right\|^2\right)\beta_j - \sum \zeta_j \mu_j + C\sum \zeta_j \quad ..(6)$$

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

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

Dengan aturan Wolfe rumus diatas menjadi:

$$R^2(X) = K(X,X) - 2\sum_j \beta_j K(X_j,X) + \sum_{ij} \beta_i \beta_j K(X_j,X_i) \quad ...(7)$$

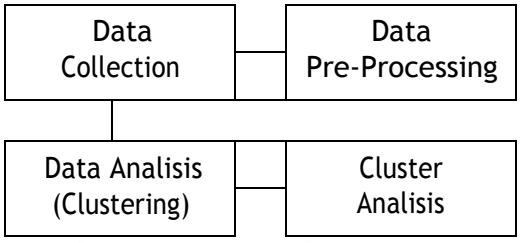
Sehingga bentuk klaster dapat dilihat dengan melihat titik –titik support vector dari klaster tersebut. Untuk menentukan titik masuk ke klaster mana diperlukan pengujian jarak titik tersebut dengan titik yang lain. Misal terdapat titik i dan j maka i dan j termasuk dalam klaster yang sama jika jarak 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 Aij=1 jika titik i dan j terletak dalam 1 klaster dan Aij=0 jika i dan j tidak terletak dalam 1 klaster.

2.2 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].

3. METODE PENELITIAN

Secara detail metodologi penelitianini dirancang seperti diagram block berikut ini :



Gambar 1. Diagram Blog Sistem

3.1. Data Colection

Data tugas akhir mahasiswa diperoleh dari database AMIK Tunas Bangsa Pematangsiantar berjumlah 2020 dengan atribut sebagai berikut :

1. Nim
2. Nama
3. Tahun Akademik
4. Judul TA
5. Klasifikasi TA

3.2. Data Preprocessing

Data yang telah dikumpulkan dilakukan proses untuk mempersiapkan data inputan pada aplikasi SVC untuk proses klastering. Langkah- langkah dalam preprocessing data dilakukan sebagai berikut [3], [4] :

1. Data Cleaning
   Proses data cleaning dilakukan untuk memastikan dalam tabel data tidak

ISSN 2548-9712

terdapat missing value, tidak konsisten dan atribut yang hilang.

2. Data Transformasi
   Format data dibuat dalam bentuk Excel, akan dikonversi menjadi format XML sebagai inputan untuk aplikasi SVC.

3.3. Data Analisis (Clustering)

Data yang sudah dipersiapkan sebagai output dari preprocessing akan dianalisis oleh 2 metode klastering yakni : SVC.

3.4. Cluster Analisis

Hasil klaster yang diperoleh dari proses SVC diinterpretasikan untuk mendapatkan performa dari masing-masing klaster yang dihasilkan oleh SVC, kedua hasil klaster akan dibandingkan untuk melihat model klaster yang terbaik. Model klaster yang terbaik tersebut diperoleh dengan parameter yakni jumlah data pada setiap klaster, atribut data yang mudah dilihat dari setiap klaster yang terbentuk.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Proses SVC

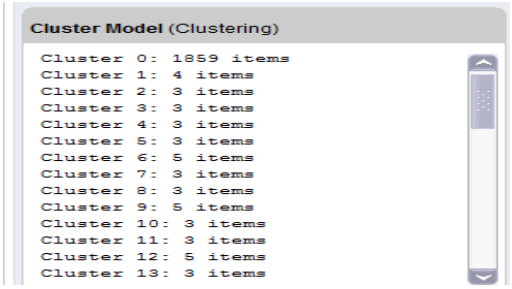
Pada gambar 2 menunjukkan bahwa setelah dilakukan pengklasteran dengan menggunakan metode SVC didapat waktu berkisar 11:21 menit.

Meta data View	
Processing time	suhada_svc (3 results, Process results) Completed: May 17, 2013 8:47:54 AM (execution time: 10:49)
Performance Vector	Performance Vector (Performance) PerformanceVector: Data Based Performance: 2035.000

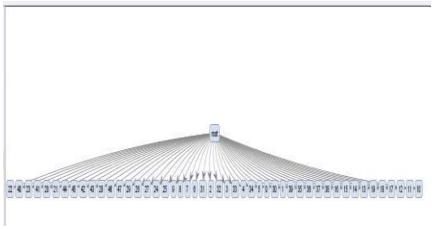
Gambar 2. Hasil klastering dengan proses SVC

Pada proses dengan Support Vector Clustering (SVC) menunjukkan kecepatan pengklasteran data yakni judul tugas akhir mahasiswa AMIK Tunas Bangsa Pematangsiantar hasil klasteringnya didapat waktu 11:21 menit, dengan klaster yang lebih baik. Ini dapat dilihat pada gambar 3 berikut ini :

Gambar 3. Proses dengan Support Vector



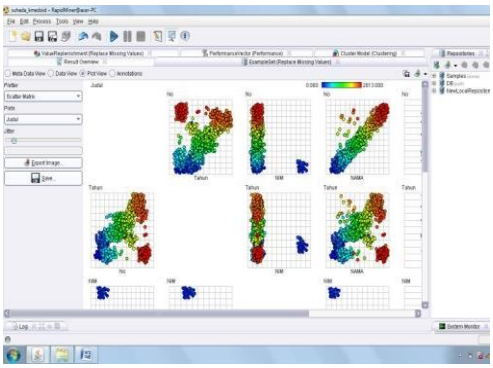
Clustering



Gambar 3. Model Klaster dari ALgoritma SVC

4.2 Komunitas Rapidminer

Hasil pengujian dengan rapidminer dengan mengambil dari judul tugas akhir mahasiswa didapat pengklasteran dari SVC dengan waktu 11:21 menit dengan hasil klastering yang dapat dilihat pada gambar 4 sebagai berikut



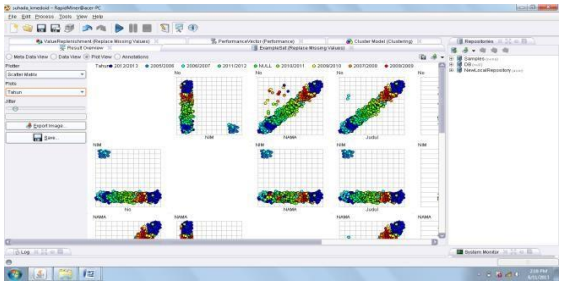
Gambar 4. Hasil klastering SVC berdasarkan judul dengan Scatter Matrix

Pada gambar 4 dapat dilihat bahwa hasil klastering dari SVC menggunakan Scatter

ISSN 2548-9712

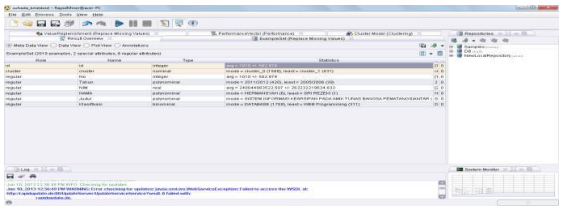
Matrix dengan mengambil judul sebagai sumbu-X dan tahun,nim dan nama sebagai sumbu-Y.

Berikutnya adalah hasil klastering SVC berdasarkan klasifikasi dapat dilihat pada gambar 5 berikut ini :



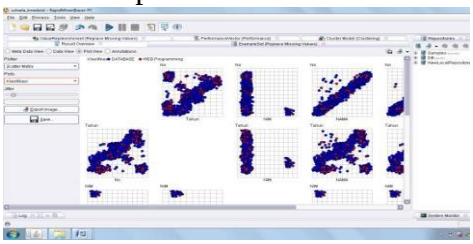
Gambar 5. Hasil klastering SVC berdasarkan klasifikasi dengan Scatter Matrix

Selanjutnya hasil klastering dari SVC berdasarkan Nim dapat dilihat pada gambar 6 sebagai berikut :



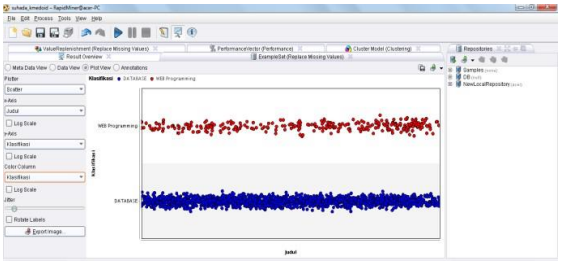
Gambar 6. Hasil klastering SVC berdasarkan Nim dengan Scatter Matrix

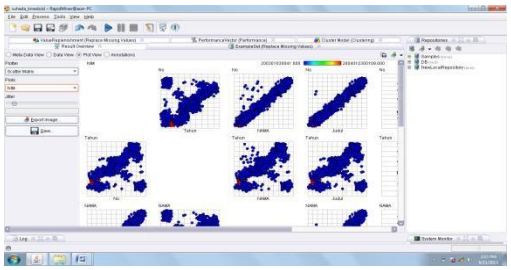
Dengan pengklasteringan kan tahun akademik dapat berikut ini :



Gambar 7. Hasil klastering SVC berdasarkan Tahun Akademik dengan Scatter Matrix

Untuk melihat hasil klastering baik proses SVC maupun proses K-Medoid ditemukan pengklasteran yang dapat dilihat pada tabel berikut :





Gambar 7. Hasil klastering SVC didapat berdasarkan Klasifikasi

Pada gambar 5 menunjukan hasil pengklasteran dengan SVC judul Tugas Akhir mahasiswa dengan klasifikasi sebagai sumbu-Y dan sumbu-X sebagai Judul, dengan klasifikasi judul database berwarna biru dan web programming berwarna merah.

5. KESIMPULAN

Berdasarkan analisis dan implementasi sistem maka diperoleh kesimpulan sebagai berikut :

1. Dengan menggunakan algoritma SVC dan software Rapidminer maka sistem telah mampu menemukan klaster-klaster dalam judul tugas akhir mahasiswa.
2. Hasil yang diperoleh pengklasteran dengan SVC mengambil waktu berkisar 11:21 Menit.
3. Jumlah klaster secara klasifikasi untuk database adalah 1708 judul dan untuk web programming berjumlah 311 judul.

ISSN 2548-9712

6. SARAN

Diharapkan para peneliti selanjutnya untuk menggunakan software dan algoritma klastering yang lainnya.

DAFTAR PUSTAKA

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

[2]. K.STAPOR (2006), Support Vector Clustering Algorithm for Identification of Glaucomain Opthalmology, Bulletin of The Polish Academy of Science Technical Science vol.54 No1.

[3]. Mochammad Juniarto. , (2009), ”Implementasi Metode Support Vector Clustering untuk Pengklasteran Produk”, Institut Teknologi Sepuluh November.

[4]. Riwinoto, (2012), ”Perbandingan Quantum Clustering dan Support Vector Clustering untuk Data Microarray Expression Yeast Cell dalam Ruang Singular Value Decomposition (SVD), Seminar Nasional Aplikasi Teknologi Informasi 2012(SNASTI 2012).

[5]. Relita Buaton, Yeni Sundari, Yani Maulita, (2016), “Clustering Tindak Kekerasan Pada Anak Menggunakan Algoritma K-Means Dengan Perbandingan Jarak Kedekatan Manhattan City Dan Euclidean”, MEANS (Media Informasi Analisa dan Sistem).