

ANALISIS DISKRIMINAN LINEAR DALAM KLASIFIKASI DATA PADA TEORI INFORMASI DENGAN METODE CROSS VALIDATION

Arini Panjaitan¹, Bambang Hidayat², Kris Sujatmoko³

¹Teknik Telekomunikasi, Fakultas Teknik Elektro, Universitas Telkom

Abstrak

Analisis diskriminan adalah salah satu teknik statistik yang bisa digunakan pada hubungan dependensi atau hubungan antarvariabel dimana sudah bisa dibedakan mana variable respond dan mana variable penjelas. Lebih spesifik lagi, analisis diskriminan digunakan pada kasus dimana variable respon berupa data kualitatif dan variabel penjelas berupa data kuantitatif. Analisis diskriminan bertujuan untuk mengklasifikasikan suatu individu atau observasi ke dalam kelompok yang saling bebas (mutually exclusive) dan menyeluruh (exhaustive) berdasarkan sejumlah variabel penjelas.

Dalam tugas akhir ini akan dilakukan suatu analisis yang bertujuan untuk merealisasikan metode Linear Discriminant Analysis pada penggunaan dalam pengklasifikasian suatu keakurasian dengan perangkat lunak MATLAB, sehingga dapat diperoleh keunggulan daripada metode Linear Discriminant Analysis ini dalam mengolah pengklasifikasian data secara linear serta mengaplikasikan sistem dalam metode Cross Validation.

Analisis Diskriminan merupakan dasar utama pada sistem pengenalan pola dengan memaksimalkan keragaman kelas, dimana sifat dari Linear Discriminant Analysis ialah distribusi normal dan homoskedastis, yaitu suatu keadaan variansi dari error bersifat konstan atau dapat dikatakan identik. Dan dari penelitian yang telah dilaksanakan, nilai analisis dataset Wine yang diperoleh LDA sebesar 1.9 yang berarti lebih baik dibandingkan dengan PCA sebesar 5.9.

Kata Kunci : Linear Discriminant Analysis, data, klasifikasi.

Abstract

Discriminant analysis is one of statistical technique which can be used in dependency or variables correlation where can distinguish the respond and the explanatory one. For more specifically, discriminant analysis is used where the respond cvariable is qualitative data and the explanatory variables is quantitative data . Discriminant analysis aims to classify an individual or group observations into mutually independent and thorough (exhaustive) based on number of explanatory variables.

In the final assignment will be carried out an analysis aiming to realize the use of Linear Discriminant Analysis in classifying data and seeing accuracy of the data classification using MATLAB software, to obtain the advantages of Discriminant Linear Analysis method to process the data linearly and to apply the system in Cross Validation method.

Discriminant analysis is the main basic on pattern recognition system to maximize the diversity of the class, where the typical of Linear Discriminant Analysis is homoscedastic and normal distribution, which is a state of variance is constant or used to be identical. From this research, the value of LDA is 1.9 which is better than PCA's value, 5.9.

Keywords : Linear Discriminant Analysis, data classification.

BAB I

PENDAHULUAN

1.1. Latar Belakang

Analisis diskriminan merupakan teknik menganalisis data, dimana variabel dependen merupakan data kategori (nominal dan ordinal) sedangkan variabel independen berupa data interval atau rasio. Analisis diskriminan bertujuan untuk mengklasifikasikan suatu individu atau observasi ke dalam kelompok yang saling bebas (*mutually exclusive*) atau dikenal juga dengan metode dependensi yaitu variabel-variabelnya tidak saling bergantung satu dengan yang lain. Berbeda dengan metode interdependensi yaitu metode yang antarvariabelnya ada saling ketergantungan. Sistem tersebut juga dapat menganalisa dan mengidentifikasi bentuk, tekstur, atau bahkan warna dari suatu objek.

Dalam industri telekomunikasi, penerapan analisis diskriminan pun sangat diperlukan. Berbagai macam data yang penting mengenai infrastruktur pembangunan jaringan telekomunikasi maupun data yang berkaitan dengan pelanggan akan dikumpulkan dalam suatu database yang dikumpulkan oleh manusia atau pekerja industri telekomunikasi tersebut. Hal ini kerap kali menimbulkan kesalahan atau *human error* secara tidak sengaja. Meskipun jumlah error sedikit, tetapi cukup berdampak pada infrastruktur telekomunikasi lain yang masih berkaitan. Maka analisis perlu dilakukan agar dapat diketahui letak data yang error atau bersifat sebagai *outlier*.

Sebagai suatu metode yang alternative dibandingkan metode-metode yang sudah ada digunakan metode yang memanfaatkan hubungan antara fitur-fitur pada suatu kelas untuk mengklasifikasikan suatu sampel pada kelas tertentu dengan data yang berjumlah besar yang tidak dapat dengan mudah diselesaikan oleh metode-metode klasifikasi yang lain. Maka akan dilakukan suatu analisis yang bertujuan untuk merealisasikan penggunaan metode Linear Discriminant Analysis dalam keakurasian klasifikasi data termasuk menentukan munculnya *data error* dalam bentuk outlier dengan menggunakan perangkat lunak MATLAB. Dari keunggulan metode Linear Discriminant Analysis ini kemudian akan diaplikasikan Cross Validation. Untuk membuktikan kemampuan klasifikasi, akan dilakukan pengujian pada metode ini dengan menggunakan data set *Wine*. Penggunaan dataset tersebut dikarenakan

inter-relation fitur pada data set tersebut lebih kuat daripada contoh data set lain. Hal ini semakin menunjang untuk menampilkan keunggulan penggunaan metode klasifikasi.

Pembangunan klasifikasi dataset dengan menggunakan metode Linear Discriminant Analysis dan *Principal Component Analysis* dengan *Cross Validation* ini dibagi menjadi dua bagian, yaitu tahap pelatihan dan tahap pengujian. Tahap pelatihan bertujuan untuk mendapatkan fitur-fitur penting yang akan menjadi masukan untuk MATLAB, dimana pada tahap pelatihan proses klasifikasi ini bertujuan untuk mendapatkan *classifier* yang dapat memisahkan kelas-kelas yang ada pada dataset. Kemudian tahap pengujian bertujuan untuk melihat bagaimana sistem berjalan dari awal sampai akhir dengan beberapa parameter pengujian. Diharapkan dengan adanya penelitian ini dapat menghasilkan sistem klasifikasi data dengan akurasi yang baik dan penggunaan metode yang tepat.

1.2. Tujuan

Adapun tujuan dari penelitian dan analisis dari Tugas Akhir ini adalah:

1. Melakukan pengklasifikasian data menggunakan Linear Discriminant Analysis yang bersifat distribusi normal dan homoskedastik yang dapat memaksimalkan sistem pengenalan pola dalam mengklasifikasikan data.
2. Mengklasifikasikan data dari dataset ke dalam bentuk yang ternormalisasi, baik pada Linear Discriminant Analysis maupun pada Principal Component Analysis.
3. Mengaplikasikan teknik klasifikasi yang telah ternormalisasi dalam metode *Cross Validation*.

1.3. Rumusan Masalah

Permasalahan yang dijadikan obyek penelitian dan pengembangan pada Tugas Akhir ini adalah:

1. Bagaimana mengklasifikasikan data dengan menggunakan metode Linear Discriminant Analysis dan Principal Component Analysis?
2. Bagaimana membentuk sebuah dataset ke dalam bentuk data yang telah ternormalisasi?

3. Bagaimana merancang algoritma metode klasifikasi pada perangkat lunak MATLAB?

1.4. Batasan Masalah

Untuk menghindari meluasnya materi dalam pembahasan Tugas Akhir ini, maka penulis membatasi permasalahan dalam Tugas Akhir ini dengan hanya mencakup hal-hal berikut:

- a. Data yang digunakan dalam penelitian berasal dari UCI Repository.
- b. Penelitian ini menggunakan metode Linear Discriminant Analysis.
- c. Klasifikasi yang telah dihasilkan akan dinormalisasi dan diiterasi dengan teknik Cross Validation.

1.5. Metode Penelitian

Metode penelitian yang digunakan untuk menyelesaikan tugas akhir ini adalah metode ekperimental, sedangkan prosedur penelitian untuk menyelesaikan Tugas Akhir ini adalah:

- a. Studi Literatur
Prosedur ini dilakukan untuk mempelajari konsep dan teori-teori yang mendukung proses perancangan sistem
- b. Survey lapangan
Analisis permasalahan dan kebutuhan
- c. Perancangan dan realisasi
Merancang suatu analisis dengan menggunakan algoritma pada *software* MATLAB
- d. Analisis dan pengujian
Berdasarkan teori yang sudah diperoleh sebelumnya, kemudian melakukan pengujian terhadap rancangan analisis pada *software*.
- e. Konsultasi
Dilakukan dengan dosen pembimbing menyangkut petunjuk dan pertimbangan-pertimbangan praktis mengenai perancangan analisis tersebut.

1.6. Sistematika Penulisan

Penulisan proposal ini akan dibagi menjadi beberapa bagian sebagai berikut:

BAB I Pendahuluan

Berisi latar belakang, tujuan, rumusan masalah, batasan masalah, metode penelitian, dan sistematika penulisan

BAB II Dasar Teori

Mengenai dasar-dasar teori yang diperlukan serta literature-literatur yang mendukung dalam penelitian dan analisis penggunaan metode Linear Discriminan Analysis pada sistem *error rate classifier*.

BAB III Perancangan dan Implementasi

Mengenai pembahasan perancangan awal system yang kemudian diimplementasikan dengan menggunakan algoritma pada perangkat lunak MATLAB.

BAB IV Pengujian dan Analisis Hasil Implementasi

Membahas mengenai pengujian dan analisis hasil implementasi yang telah dilakukan sebelumnya. Hal ini bertujuan untuk mengetahui kecocokan antara analisis yang dilakukan secara teori dengan menggunakan perangkat lunak MATLAB.

BAB V Kesimpulan dan Saran

Berisi kesimpulan dari penelitian Tugas Akhir ini serta saran yang diperlukan untuk pengembangan sistem lebih lanjut.

BAB V

KESIMPULAN DAN PENUTUP

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis yang telah dilakukan, dapat disimpulkan beberapa hal berikut ini:

1. Penerapan *cross validation* pada metode klasifikasi *Linear Discriminant Analysis* berhasil dilakukan. Dan dari data pengujian yang diperoleh, hasil penerapan *cross validation* pada teknik klasifikasi meningkatkan tingkat akurasi hasil klasifikasi dibandingkan tanpa menerapkan teknik *cross validation* pada *Linear Discriminant Analysis*, begitu pula pada *Principal Component Analysis*.
2. Tingkat akurasi penyebaran dengan metode PCA dapat dikatakan tidak baik, dikarenakan sifat distribusi PCA dimana rata-rata dan kovariansinya tidak resisten pada distribusi data. Seperti pada Tabel 4.5 dapat diketahui bahwa nilai terkecil **1.9** yang diperoleh LDA dengan normalisasi menunjukkan bahwa metode LDA lebih baik dibanding PCA yaitu sebesar **5.9**.
3. Sebelum dilakukan Cross validation dan klasifikasi, harus dilakukan proses normalisasi data terlebih dulu pada dataset agar persebaran data merata dengan menggunakan algoritma *zscore*, dimana cakupan nilai rata-rata penyebaran antara 0 hingga 1.
4. Penerapan Cross validation menggunakan toolbox *crossvalind* pada data yang sudah distribusi secara LDA.

5.2 Saran

Berikut ini merupakan beberapa hal yang disarankan penulis untuk dilakukan pada penelitian yang sehubungan dengan penelitian ini di masa mendatang:

1. Dengan penerapan Cross validasi ini, penelitian dapat dilanjutkan pada data yang bersifat non-linear.

2. Untuk menambah tingkat akurasi sistem pada klasifikasi data, sebaiknya penelitian dilakukan pada jenis dataset yang lain agar hasil yang diperoleh pun lebih banyak dan beragam, serta menambah tingkat kepastian dalam penggunaan teknik *cross validation*.
3. Untuk penelitian lebih lanjut dapat digunakan metode lain selain *Linear Discriminant Analysis* maupun *Principal Component Anlysis* sebagai pembanding terhadap hasil yang telah diperoleh dari penelitian saat ini.



DAFTAR PUSTAKA

- [1] Hart, Duda. *Pattern Classification*. Diperoleh 28 Januari 2013, dari <http://syridink.com/files/159Download%20pattern%20classification%20.pdf>
- [2] Haihong ZHANG, Cuntai GUAN, and Kai Keng Ang. 2010, *An Information Theoretic Linear Discriminant Analysis Method*. Diperoleh 29 Januari 2013, dari <http://www.ieeexplore.ieee.org:80?xpl/articleDetails.jsp?reload=true&arnumber=5597750>.
- [3] *Analisis Diskriminan*. Diperoleh pada 3 Februari 2013. Dari http://daps.bps.go.id/file_artikel/65/ANALISIS%20DISKRIMINAN.pdf
- [5] Godam64. (2010, 21 Juni), *Klasifikasi Jenis, dan Macam Data – Pembagian Data Dalam Ilmu Eksak*. Diperoleh 13 Maret 2013, dari http://organisasi.org/klasifikasi_jenis_dan_macam_data_pembagian_data_dalam_ilmu_eksak_sains_statistik_statistika
- [7] *Machine Learnig Repositroy: Wine Data Set*. Diperoleh pada 14 Maret 2013. Diperoleh dari <http://archive.ics.uci.edu/ml/datasets/Wine>
- [8] *Validation*. Diperoleh pada 15 April 2013. Diperoleh dari http://research.cs.tamu.edu/prism/lectures/iss/iss_113.pdf
- [9] Natalius, Samuel. *Metoda Naive Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen*, Makalah II2092 Probabilitas dan Statistik – Sem. I Tahun 2010/2011, Bandung. Diperoleh pada 16 April 2013.

- [10] Scholdz, Matthias. *Principal Component Analysis*. Diperoleh pada 29 April 2013. Dari http://www.nlpca.org/pca_principal_component_analysis.html
- [12] *Principal Component Analysis and Linear Discriminant Analysis with GNU Octave*, Diperoleh 27Mei 2013 dari http://www.bytefish.de/blog/pca_lda_with_gnu_octave/
- [13] Shlenz, Jonathan. *Principal Component in Matla*. Diperoleh pada 4 Juni 2013. Dari http://www.cs.stevens.edu/~mordohai/classes/cs559_s09/PCA_in_MATLAB.pdf
- [14] *DataSet Object Examples*. Diperoleh pada 4 Juni 2013. Dari wiki.eigenvector.com/index.php?title=DataSet_Object_Examples
- [15] *Principal Component Analysis of Raw Data*. Diperoleh pada 10 juni 2013. Dari <http://www.mathworks.com/help/stats/pca.html#bti6n7k-2>
- [16] *Linear Discriminan Analysis for Dataset*. Diperoleh pada 10 Juni 2013. Dari <http://www.mathworks.com/help/stats/lda-for-dataset.html#ok4f7n-8>
- [17] *Cross Validation*. Diperoleh pada 8 September 2013. Dari http://www.cse.iitb.ac.in/~tarung/smt/papers_ppt/ency-cross-validation.pdf

Telkom
University