Forum Statistika dan Komputasi,September 2004, p:101-106
ISSN 0853-8115

ANALISIS DISKRIMINAN KERNEL UNTUK PENGELOMPOKKAN WARNA

Anik Djuraidah[1] dan Aunuddin[2]

[1]Jurusan Statistika, FMIPA ITS
[2]Departemen Statistika, FMIPA IPB

Ringkasan

Analisis diskriminan kernel merupakan salah satu metode nonparametrik untuk mengklasifikasikan obyek dan reduksi data. Pada metode ini fungsi kepadatan peluang posterior diduga dengan menggunakan metode kernel. Pendugaan fungsi kepadatan peluang untuk data peubah ganda lebih kompleks dibanding peubah tunggal karena adanya korelasi antar peubah. Performansi dari penduga fungsi kepekatan peluang dengan metode kernel sangat ditentukan oleh pemilihan parameter penghalus, dan sedikit dipengaruhi fungsi kernel. Penerapan metode diskriminan kernel untuk mengelompokkan empat warna cat memberikan persentase salah kelas 0 %. Sedangkan analisis diskriminan linear Fisher memberikan persentase salah kelas sebesar 17.92 %.

Kata Kunci : kernel, analisis diskriminan, parameter pemulus

PENDAHULUAN

Analisis diskriminan adalah salah satu metode pada peubah ganda yang digunakan untuk mengklasifikasikan obyek dan reduksi data. Pendekatan analisis ini sangat beragam, mulai dari model parametrik sampai model nonparametrik. Pada pendekatan parametrik diperlukan asumsi bahwa data menyebar normal ganda. Metode diskriminan parametrik yang terkenal adalah analisis diskriminan linear (Fisher) dan kuadratik. Sedangkan salah satu metode pendekatan nonparametrik yang bersifat fleksibel adalah analisis diskriminan dengan kernal.

Metode kernel yang dikemukakan oleh Rosenblatt pada tahun 1956 banyak digunakan untuk menduga fungsi kepadatan peluang secara nonparametrik. Kelebihan penduga kernel adalah bentuknya lebih fleksibel dan bentuk matematis dari penduganya lebih mudah disesuaikan. Akan tetapi kesulitan metode ini terutama untuk data peubah ganda adalah menentukan penduga parameter penghalus. Penduga fungsi kepadatan peluang nonparametrik digunakan pada berbagai aplikasi, salah satunya adalah pada analisis diskriminan.

Prediksi dan klasifikasi dengan teknik nonparametrik mempunyai sejarah sukses yang panjang. Pada awalnya penerapan penduga fungsi kepekatan peluang digunakan sebagai bagian dari analisis diskriminan nonparametrik dikemukakan oleh Fix dan Hodges pada tahun 1951 (Silverman, 1986). Kemudian diikuti oleh Aitchison dan Aitken (1976) menggunakan metode kernel pada analisis diskriminan untuk data biner. Habbema, Herman, dan Van der Broek pada tahun 1974 mencatat hasil yang yang sangat bagus tentang penggunaan kernel normal ganda untuk klasifikasi. Titterington menggunakan kernel untuk klasifikasi pada data dengan peubah campuran (Scott, 1992). Holmström dan Sain (1997) menggunakan analisis diskriminan parametrik dan nonparametrik untuk mengelompokkan gejala gempa bumi. Analisis diskriminan kernel memberikan hasil yang bagus untuk pengenalan pola, seperti Li *et al* (2003) melakukan identifikasi wajah dengan analisis diskriminan kernel. Di bidang kesehatan Kim *et al* (2003) menggunakannya untuk mengklasifikasikan subtipe kanker.

TINJAUAN PUSTAKA

Penduga Fungsi Kepadatan Kernel

Fungsi kepadatan peluang adalah suatu karakteristik dasar yang menunjukkan sifat dari peubah acak. Penduga fungsi kepadatan peluang dapat memberikan informasi yang berharga mengenai data yang diamati seperti kemiringan, keragaman dan modus. Misalkan $X = (X_1, X_2, \ldots, X_d)'$ adalah vektor acak berdimensi d dengan fungsi kepadatan peluang $f(x)$ didefinisikan pada R^d, dan misalkan $\{ x_1, x_2, \ldots, x_n \}$ adalah contoh acak yang dipilih dari $f(x)$. Bentuk umum peduga kernel dari $f(x)$ adalah (Zhang *et al*, 2004) :

$$\hat{f}_H(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i) \qquad (1)$$

dengan $K_H(x) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}}x)$, $K(.)$ adalah fungsi kernel peubah ganda, dan H adalah matriks parameter penghalus berukuran $d \times d$ yang definit positif. Fungsi kernel $K(.)$ diasumsikan memenuhi tiga momen yaitu (Scott, 1992) :

1. $\int_{R^d} K(w) = 1$
2. $\int_{R^d} wK(w) = 0$
3. $\int_{R^d} ww' K(w) = \mathbf{I}_d$

Matriks H dapat digabung ke dalam fungsi kernel, misalnya menggunakan $K = N_d(0, \Sigma)$ dan $H = I$ ekilalen dengan memilih $K = N_d(0, \mathbf{I})$ dan $H = \Sigma$. (Scott, 1992).

Bila matriks H adalah matriks diagonal yang definit positif maka fungsi kernel padanannya dikenal sebagai perkalian kernel (*product kernel*). Misalkan nilai parameter penghalus dari setiap peubah acak berbeda sebesar h_j maka penduga fungsi kepadatan peluangnya adalah:

$$\hat{f}(x) = \frac{1}{nh_1 \ldots h_d}\sum_{i=1}^{n}\prod_{j=1}^{d} K\left(\frac{x_i - x_{ij}}{h_j}\right) \qquad (2)$$

dengan fungsi kernel $K(.)$ adalah fungsi kernel pada peubah tunggal (Scoot, 1992).

Performansi dari penduga fungsi kepekatan peluang dengan metode kernel sangat ditentukan oleh pemilihan parameter penghalus, dan sedikit dipengaruhi fungsi kernel. Banyak litelatur membahas pemilihan parameter penghalus untuk data peubah tunggal, akan tetapi untuk data peubah ganda masih sangat terbatas. Menurut Zhang (2004) pemilihan parameter penghalus baik dengan menggunakan metode bootsrap, validasi silang, maupun MCMC (*Markov chain Monte Carlo*) semakin kompleks sejalan meningkatnya dimensi data.

Untuk mendapatkan parameter penghalus h_j yang ideal ditentukan dengan minimisasi AMISE. Bila diasumsikan fungsi densitas yang diduga menyebar Normal ganda dan kernel yang digunakan adalah perkalian kernel normal peubah tunggal (persamaan 2) maka penduga parameter penghalus h_j optimal adalah :

$$h_j^{opt} = \sigma_i \left[A(K)/n_i\right]^{\frac{1}{d+1}} \qquad (3)$$

Nilai σ_i diduga dari contoh, sedangkan nilai $A(K)$ untuk beberapa kernel sebagai berikut (Khattree, 2000, Silvermann, 1986, Scott, 1992) :

- Kernel Seragam : $A(K) = \dfrac{2^{d+1}(d+2)\Gamma(d/2)}{d}$
- Kernel Normal : $A(K) = \dfrac{4}{2d+1}$
- Kernel Epanechinikov : $A(K) = \dfrac{2^{d+2}d^2(d+2)(d+4)\Gamma(d/2)}{2d+1}$

Nilai A(K) untuk kernel Normal berada pada selang (0.924 , 1.059) dengan nilai limit 1. Pada d = 2 nilai $h_i = 1$ dan terkecil pada d = 11 (Scoot, 1992).

Bila matriks H bukan matriks diagonal, maka data lebih dahulu ditransformasi. Bentuk transformasi yang disarankan adalah transformasi sperik atau transformasi skala (Hernández and Velilla, 2001; Zhang *et al*, 2004), yaitu :

$$x_i^* = S^{-\frac{1}{2}} x_i \text{ atau } x_i^* = S_d^{-\frac{1}{2}} x_i \quad \ldots (4)$$

dengan S adalah matriks ragam-peragam contoh, dan S_d adalah matriks diagonal dengan elemen diagonal ke-i adalah s_i^2. Misalkan matriks parameter penghalus H optimal dari data yang ditransformasi dinyatakan sebagai $\hat{H}^*$, matriks H dari data asli dapat dihitung melalui transformasi

$$\hat{H} = S^{\frac{1}{2}} \hat{H}^* (S^{\frac{1}{2}})' \quad \text{atau}$$

$$\hat{H} = S_d^{\frac{1}{2}} \hat{H}^* (S_d^{\frac{1}{2}})'$$

Analisis Diskriminan dengan Metode Kernel

Pada analisis diskriminan nonparametrik, fungsi kepadatan peluang dari grup diduga secara nonparametrik. Salah satu metode nonparametrik untuk menduga fungsi kepadatan adalah metode kernel. Umumnya fungsi kernel K adalah fungsi kepadatan peluang yang simetrik dan unimodal. Beberapa kernel peubah ganda yang sering digunakan pada pendugaan fungsi kepadatan peluang $f(x)$adalah (Khattree, 2000) :

- Kernel Seragam :

$$K_t(z) = \begin{cases} \dfrac{1}{v_{h(t)}} & \text{jika} \quad z' \, \mathbf{V}_t^{-1} z \le h^2 \\ 0 & \text{untuk lainnya} \end{cases}$$

$$\text{dengan } v_{h(t)} = \frac{\pi^{\frac{d}{2}} h^d}{\Gamma(\frac{d}{2} + 1)} |\mathbf{V}_t|^{\frac{1}{2}}$$

- Kernel Normal :

$$K_t(z) = \frac{1}{c_{0(t)}} \exp(-0.5 \, z' \, \mathbf{V}_t^{-1} z / h^2)$$

$$\text{dengan } c_{0(t)} = (2\pi)^{\frac{d}{2}} h^d |\mathbf{V}_t|^{\frac{1}{2}}$$

- Kernel Epanechinikov :

$$K_t(z) = \begin{cases} c_{1(t)}(1 - z' \mathbf{V}_t^{-1} z / h^2) & \text{jika} \quad z' \mathbf{V}_t^{-1} z \le h^2 \\ 0 & \text{untuk lainnya} \end{cases}$$

$$\text{dengan } c_{1(t)} = (1 + \frac{d}{2}) v_{h(t)}$$

- Kernel *Biweight* :

$$K_t(z) = \begin{cases} c_{2(t)}(1 - z' \mathbf{V}_t^{-1} z / h^2)^2 & \text{jika} \quad z' \mathbf{V}_t^{-1} z \le h^2 \\ 0 & \text{untuk lainnya} \end{cases}$$

$$\text{dengan } c_{2(t)} = (1 + \frac{d}{4}) c_{1(t)}$$

- Kernel *Triweight* :

$$K_t(z) = \begin{cases} c_{3(t)}(1 - z'\mathbf{V}_t^{-1}z/h^2)^3 & \text{jika} \quad z'\mathbf{V}_t^{-1}z \le h^2 \\ 0 & \text{untuk lainnya} \end{cases}$$

$$c_{3(t)} = (1 + \frac{d}{6})c_{2(t)}$$

dimana $t = 1, \ldots, g$ menyatakan grup ke-t, h menyatakan nilai parameter penghalus, dan matriks $\mathbf{V}_k$ adalah matriks ragam-peragam berukuran d x d yang mempunyai salah satu bentuk di bawah ini :

$\mathbf{V}_t = \mathbf{S}$	Matriks ragam-peragam gabungan.
$\mathbf{V}_t = \text{diag } \mathbf{S}$	Matriks diagonal dari ragam-peragam gabungan
$\mathbf{V}_t = \mathbf{S}_t$	Matriks ragam-peragam grup t
$\mathbf{V}_t = \text{diag } \mathbf{S}_t$	Matriks diagonal dari ragam-peragam pada grup t
$\mathbf{V}_t = \mathbf{I}$	Matriks Identitas

Misalkan $\hat{f}_t(\boldsymbol{x})$ adalah penduga fungsi kepadatan kernel dari grup π_t, dan p_t adalah peluang awal dari grup π_t untuk $t = 1, 2, \ldots, g$. Peluang posterior suatu pengamatan x berasal dari grup π_t adalah

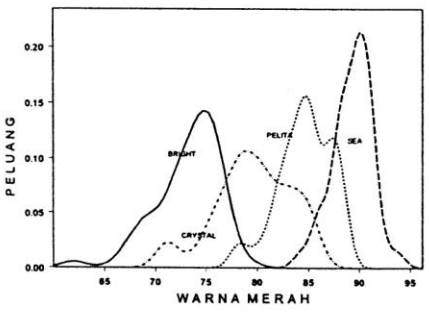
$$P(x_i \in \pi_t | x_i = x) = \frac{p_t \hat{f}_t(x)}{\sum_{j=1}^{g} p_j \hat{f}_j(x)}$$

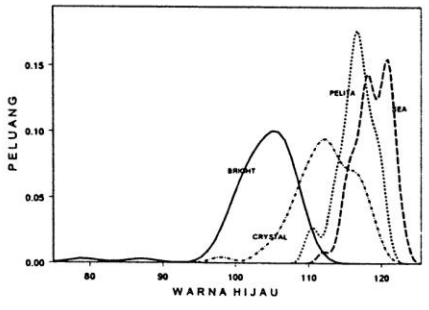
Aturan pengklasifikasian pada analisis diskriminan nonparametrik sama dengan analisis diskriminan parametrik yaitu pengamatan x_0 dialokasikan ke π_t jika peluang posterior $P(\pi_t | x_0) > P(\pi_i | x_0)$ untuk semua $t \neq i$.

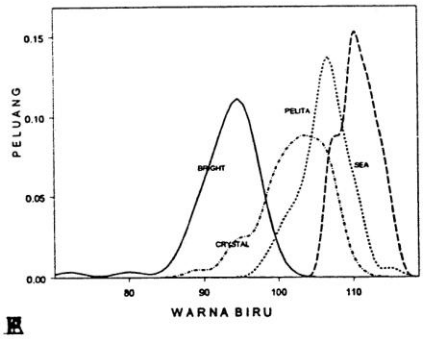
PENERAPAN PADA PENGELOMPOKKAN WARNA CAT

Data yang digunakan adalah nilai komposisi pixel dari warna cat yang mempunyai tiga peubah yaitu warna merah, warna hijau, dan warna biru. Obyek terdiri dari empat warna cat yang memiliki warna yang mirip, yaitu *bright green, pelita green, crystal green,* dan *sea green*. Pengambilan data menggunakan kamera yang dihubungkan ke komputer dengan *Card Video Baster SE-100 Creative* (Satriyanto,1999) menggunakan bahasa pemrograman C. Ukuran contoh dari masing-masing warna cat sebesar 60 pixel.

Penduga fungsi kepadatan dari keempat warna cat untuk masing-masing peubah warna merah, hijau, dan biru disajikan pada Gambar 1. Kurva mulus pada Gambar 1 diperoleh dengan menggunakan kernel normal dan nilai parameter penghalus h optimal. Dari gambar ini terlihat bentuk sebaran data yang tidak simetris, keragaman data pada setiap warna cat tidak homogen, antar warna cat terlihat tumpang tindih, dan ada cat yang berpuncak dua.



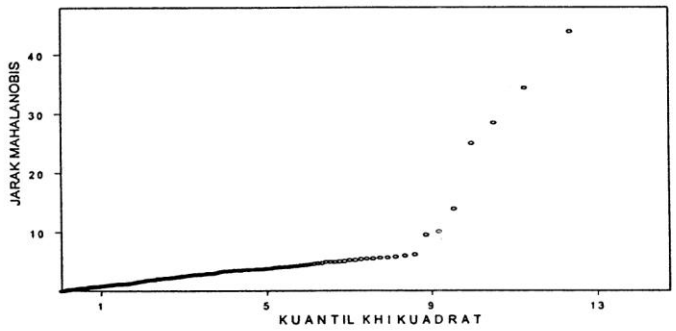




Gambar 1. Kurva Mulus dari Empat Warna Cat pada masing-masing Peubah Warna

Untuk mengetahui apakah data dapat didekati dengan sebaran normal ganda dilakukan dengan membuat plot antara jarak Mahalanobis dengan kuantil χ^2. Dari plot yang disajikan pada Gambar 2, tampak tebaran data tidak berpola lurus. Sehingga secara eksplorasi dapat dikatakan data tidak mempunyai sebaran normal ganda. Pengujian kehomogenan matriks ragam-peragam menggunakan statistik uji Box M, dengan nilai statistik uji F sebesar 8.868 yang mempunyai nilai-p sebesar 0.00. Dengan demikian matriks ragam-peraham antar warna cat tidak homogen.

Berdasarkan hasil pengujian pada sebaran data ini, maka pendekatan yang paling baik untuk mengelompokkan warna cat menggunakan analisis diskriminan nonparametrik. Di samping itu karena dari hasil pengujian matriks ragam-peragam yang menunjukkan tidak homogen maka untuk mendapatkan nilai parameter penghalus yang sederhana data ditranformasi sperik (persamaan 4). Nilai h optimal pada data yang ditransformasi dihitung dengan persamaan (3). Bila kernel yang digunakan kernel Normal nilai h optimalnya sebesar 0.3129. Hasil analisis diskriminan dengan kernel normal diolah dengan paket program SAS V8.2 memberikan kesalahan klasifikasi 0 %. Sedangkan hasil analisis diskriminan linear Fisher memberikan kesalahan klasifikasi 17.92 %.



Gambar 2. Plot antara Jarak Mahalanobis dengan Kuantil Sebaran χ^2

KESIMPULAN

Analisis diskriminan kernel mempunyai cakupan yang luas yaitu dapat digunakan untuk sembarang fungsi kepadatan peluang atau mengikuti perilaku data. Penerapan metode ini pada data yang tidak menyebar normal ganda dan mempunyai matriks ragam-peragam tidak homogen memberikan kesalahan klasifikasi yang lebih kecil dibanding dengan analisis diskriminan Fisher.

DAFTAR PUSTAKA

Aitchison, J. and C. G. G. Aitken. 1976. *Multivariate Binary Discrimination by Kernel Method.* Biometrika 63(3) : 413-420.

Hernández, A. and S. Velilla. 2001. *Dimension Reduction in Nonparametric Discriminant Analysis.* Working Paper 01-39 Statistical and Econometrics Series 25. Departement de Estadística y Econometría. University Carlos III de Madrid. Madrid

Holmström, L. and S. R. Sain. 1997. *Multivariate Discriminant Methods for Top Quark Analysis.* Technometrics 39 : 91-99.

Khattree, R. and Naik, D. N. 2000. *Multivariate Data Reduction and Discriminatio. With SAS Software.* SAS Institute. North Caroline.

Kim, D. H. and J.Cho, and D. Lee, and I. Lee. 2003. *Cancer Subtype Classification Using a New Gene Selection Method and Kernel Fisher's Discriminant Analysis.* Departement of Chemical Engineering POSTECH. http://cbit.snu.ac.kr/bac2003/workshop03/CHIP1.pdf

Li, Y., S. Gong and H. Liddell.2003. *Recognising Trajectories of Facial Indentities Using Kernel Discriminant Analysis.* Image and Vision Computing, 21(13-14) : 1077-1086. http://www.brunel.ac.uk/~csstyyl/papers/bmvc2001.pdf

Satriyanto, E. 1999. *Analisis Pengelompokkan pada Kasus Identifikasi Warna Obyek menggunakan Kamera.* Tugas Akhir S1 Institut Teknologi 10 Nopember. (tidak dipublikasikan)

Scott D. W. 1992. *Multivariate Density Estimation. Theory, Practice and Visualization.* John Wiley and Sons. New York.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall. London.

Zhang, X, M. L. King, and R. J. Hyndman. 2004. *Bandwith Selection for Multivariate Density Estimation Using MCMC.* Working Paper. Department Econometric of Econometrics and Business Statistics Monash University. http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/