Statistika, Vol. 4, No. 2, November 2016

KETEPATAN PENGKLASIFIKASIAN FUNGSI DISKRIMINAN LINIER ROBUST DUA KELOMPOK DENGAN METODE *FAST MINIMUM COVARIATE DETERMINANT* (*FAST* –MCD)

Budyanra
Jurusan Statistika, Sekolah Tinggi Ilmu Statistik, Jakarta
Email: budy@stis.ac.id

ABSTRAK

Penaksiran fungsi diskriminan linier dua kelompok dengan metode klasik/*MLE* tidakakan optimal pada saat data mengandung *outlier*. Agar analisis diskriminan tetap optimal maka diperlukan suatu metode penaksir yang *robust* terhadap *outlier*. Salah satu penaksir *robust* adalah metode *fast-MCD*. Makalah ini mengkaji metode penaksir*robustfast-MCD* dalam analisis diskriminan linier dua kelompok dan mengukur ketepatan pengklasifikasian dari fungsi diskriminan metode *fast-MCD* jika dibandingkan fungsi diskriminan metode klasik/*MLE*. Untuk menguji ketepatan penaksir fungsi diskriminan linier *robust* dua kelompok dengan metode *fast-MCD* tersebut digunakan data hasil simulasi.Data simulasi diperoleh dengan membangkitkan dua kelompok data normal multivariat dengan $n_1 = n_2 = 100$, $n_1 = n_2 = 200$, $n_1 = n_2 = 500$ dan $n_1 = n_2 = 1000$ serta dengan variasi *outlier* mulai dari 5 persen, 10 persen, 15 persen sampai 20 persen. Hasil pengolahan pada data simulasi dengan kontaminasi outlier sebesar 5 sampai 20 persen, terlihat bahwa metode *fast-MCD* menghasilkan rata-rata salah pengklasifikasian sebesar 11 persen, yang masih jauh lebih rendah jika dibandingkan metode *MLE* dengan rata-rata sebesar 22 persen. Untuk data yang banyak mengandung *outlier*, ternyata fungsi diskriminan linier dengan penaksir *robust fast-MCD* sangat efektif digunakan untuk mengurangi kesalahan dalam pengklasifikasian dari fungsi diskriminan linier tersebut.

Kata Kunci : Fungsi Diskriminan Linier, *Robust*, Fast-MCD

PENDAHULUAN

Analisis diskriminan merupakan metode statistika yang berkaitan dengan penentuan fungsi pembeda yang bertujuan untuk mengelompokkan objek pengamatan ke dalam kelompok yang telah didefinisikan berdasarkan sejumlah variabel pembeda [2]. Tujuan lainnya adalah untuk menentukan variabel pembeda diantara populasi dengan populasi lainnya. Analisis diskriminan dapat menggambarkan perbedaan antar kelompok populasi melalui suatu persamaan matematis yang disebut fungsi diskriminan. Menurut [6], fungsi diskriminan berfungsi sebagai aturan yang dapat membedakan objek pengamatan sekaligus mengelompokkannya ke dalam salah satu kelompok secara optimal. Aturan pengelompokan melalui fungsi diskriminan menurut Fisher dalam [11], yaitu dengan menggunakan kombinasi linier dari pengamatan dan memilih koefisien, sehingga rasio selisih rata-rata dari kombinasi linier dua kelompok terhadap variansinya maksimum.

Asumsi dasar dari analisis diskriminan adalah asumsi bahwa variabel pembeda harus mengikuti distribusi normal multivariat [12]. Ketika asumsi distribusi normal multivariat telah terpenuhi maka untuk menaksir parameternya bisa dengan menggunakan metode *maximum likelihood estimator* (*MLE*). Penaksiran dengan *MLE* berdasarkan pada rata-rata empirik dan matriks kovariansi dari data, tetapi karena metode ini sangat dipengaruhi oleh pengamatan *outlier* sehingga penaksirnya menjadi kurang tepat pada saat data telah terkontaminasi oleh *outlier* [3].

Agar analisis diskriminan tetap optimal dalam pengklasifikasian meskipun dalam kondisi data yang mengandung *outlier* maka diperlukan suatu penaksir *robust* terhadap data yang mengandung *outlier*. Analisis diskriminan yang menggunakan penaksir *robust* selanjutnya akan disebut sebagai analisis diskriminan *robust*. Konsep dasar dari analisis diskriminan *robust* adalah mengganti vektor rata-rata dan matriks kovariansi dari data dengan vektor rata-rata dan matriks kovariansi yang *robust* dalam menaksir parameter model [8].

Tujuan penulisan makalah ini adalah mendapatkan penaksir yang *robust* dalam kaitannya dengan data yang mengandung *outlier* dan membandingkanketepatan pengklasifikasian fungsi diskriminan *robust* metode *fast-MCD*dengan fungsi diskriminan metode klasik/*MLE* pada beberapa data hasil simulasi. Untuk mengevaluasi ketepatan fungsi diskriminan dalam pengklasifikasian, salah satu metode yang dapat dipakai yaitu *apparent error rate* (*APER*). Nilai *APER* menurut [6] adalah banyaknya persentase yang salah dalam pengelompokannya oleh fungsi klasifikasi.

METODE PENELITIAN

Sumber Data dan Variabel Penelitian

Penelitian ini menggunakan data simulasi yang merupakan data berdistribusi normal multivariat yang telah dikontaminasi dengan berbagai variasi *outlier* dan menggunakan matriks kovariansi yang homogen yang di-*generate* menggunakan *syntax* pembangkitan data simulasi melalui *software Matlab* 7.8.0. Data simulasi yang digunakan, dibatasi hanya pada jumlah pengamatan $n_1 = n_2 = 100$, $n_1 = n_2 = 200$, $n_1 = n_2 = 500$ dan $n_1 = n_2 = 1000$ dengan masing-masing tiga variabel bebas.

Data dibangkitkan secara *random* mengikuti distribusi normal multivariat dan akan dikontaminasi dengan data *outlier* dengan proporsi 5 persen, 10 persen, 15 persen dan 20 persen.

Metode Analisis

Adapun langkah-langkah dalam membangkitkan data simulasi untuk jumlah pengamatan $n_1 = n_2 = 1000$ dengan proporsi *outlier* sebesar 5 persen, adalah sebagai berikut :

1. Bangkitkan 950 data berdistribusi normal multivariat dengan $\mu1=(1,0,0)$ dan $\Sigma01=\text{diag}(0.4,0.4,0.4)$ dan 50 data berdistribusi normal multivariat dengan $\mu2=(0,0,6)$ dan $\Sigma11=\text{diag}(0.4,0.4,0.4)$ untuk matriks data kelompok-1. Data yang berjumlah 50 pengamatan merupakan *outlier* dengan proporsi 5 persen. $\mu1$ dan $\Sigma01$ merupakan parameter rata-rata dan matriks kovariansi untuk data normal multivariat yang dikondisikan tidak mengandung *outlier*, sedangkan $\mu2$ dan $\Sigma11$ merupakan parameter rata-rata dan matriks kovariansi untuk data normal multivariat yang dikondisikan sebagai data *outlier*. Data *outlier* yang dibangkitkan

berdistribusi normal multivariat dengan vektor rata-rata yang berbeda jauh dengan data dasar yang tidak mengandung *outlier*, sedangkan matriks kovariansinya diasumsikan sama.

2. Bangkitkan 950 data berditribusi normal multivariat dengan μ1=(0,1,0) dan Σ02=diag(0.4,0.4,0.4) dan 50 data berdistribusi normal multivariat dengan μ2=(6,0,0) dan Σ12=diag(0.4,0.4,0.4) untuk matriks data kelompok-2. Data yang berjumlah 50 pengamatan merupakan *outlier* dengan proporsi 5 persen. μ1 dan Σ02 merupakan parameter rata-rata dan matriks kovariansi untuk data normal multivariat yang dikondisikan tidak mengandung *outlier*, sedangkan μ2 dan Σ12 merupakan parameter rata-rata dan matriks kovariansi untuk data normal multivariat yang dikondisikan sebagai data *outlier*.
3. Gabungkan matriks data kelompok-1 dan matriks data kelompok-2 kedalam matriks data x.
4. Untuk meyakinkan bahwa data normal multivariat yang sudah dibangkitkan tidak *overlap* dan sesuai dengan proporsi *outlier* yang kehendaki, maka dilakukan pendeteksian besarnya proporsi *outlier* yang terkontaminasi kedalam data simulasi tersebut.
5. Bangkitkan masing-masing 1000 data variabel respon kelompok-1 dan 1000 data variabel respon kelompok-2, kemudian gabungkan data variabel respon kelompok-1 dan data variabel respon kelompok-2 menjadi vektor data y.
6. Tempatkan masing-masing matriks data x (hasil bangkitan) kedalam satu file *excel*, dan vektor data y kedalam satu file *excel* tersendiri.

Untuk membangkitkan data dengan jumlah pengamatan dan variasi *outlier* yang lain, dilakukan dengan mengganti jumlah *n* data dan proporsi *outlier*-nya.

HASIL PENELITIAN

Pengolahan data pada penelitian ini menggunakan bantuan *software* statistik *Matlab* 7.8.0 dengan *syntax* (*cda.m*, *rda.m* dan *fmcd*.m) yang harus sudah tersimpan terlebih dahulu di *folder toolbox Matlab*. Hasil pengolahan data menunjukkan pada data dengan $n_1 = n_2 =$ 100 yang terkontaminasi pengamatan *outlier* dari 5 persen sampai 20 persen, kesalahan pengklasifikasian fungsi diskriminan dengan penaksir *Fast-MCD* selalu jauh lebih kecil jika dibandingkan dengan kesalahan pengklasifikasian dengan penaksir klasik/MLE

Tabel 1. Kesalahan Pengklasifikasian Analisis Diskriminan Linier pada Data Simulasi dengan $n_1 = n_2 = 100$

Persentase Outlier	Kesalahan pengklasifikasian (%)	
	MLE	*Fast-MCD*
5%	19,00	8,06
10%	23,50	10,92
15%	20,50	11,52
20%	22,50	12,10

Berikutnya jumlah data simulasi ditambah menjadi $n_1 = n_2 =$ 200 dengan menyertakan pengamatan outlier sebanyak 5%, 10%, 15% dan 20%. Kesalahan pengklasifikasian fungsi diskriminan dengan penaksir*fast-MCD* ternyata juga masih jauh lebih kecil jika dibandingkan dengan kesalahan pengklasifikasian dengan menggunakan penaksirklasik/MLE.

Tabel 2. Kesalahan Pengklasifikasian Analisis Diskriminan Linier pada Data Simulasi dengan $n_1 = n_2 = 200$

Persentase Outlier	Kesalahan pengklasifikasian (%)	
	MLE	*Fast-MCD*
5%	22,25	18,33
10%	24,50	11,68
15%	23,50	12,31
20%	26,00	14,51

Selanjutnya dengan jumlah data simulasi ditambah lagi menjadi $n_1 = n_2 = 500$ dengan menyertakan pengamatan outlier sebanyak 5%, 10%, 15% dan 20%. Kesalahan pengklasifikasian fungsi diskriminan dengan penaksir*fast-MCD* ternyata juga masih jauh lebih kecil jika dibandingkan dengan kesalahan pengklasifikasian dengan menggunakan penaksirklasik/MLE.

Tabel 3. Kesalahan Pengklasifikasian Analisis Diskriminan Linier pada Data Simulasi dengan $n_1 = n_2 = 500$

Persentase Outlier	Kesalahan pengklasifikasian (%)	
	MLE	*Fast-MCD*
5%	21,10	11,52
10%	21,80	11,16
15%	22,90	12,82
20%	23,10	13,09

Terakhir, dengan jumlah data simulasi ditambah menjadi $n_1 = n_2 = 1000$ dengan menyertakan pengamatan outlier sebanyak 5%, 10%, 15% dan 20%. Kesalahan pengklasifikasian fungsi diskriminan dengan penaksir*fast-MCD* ternyata juga masih jauh lebih kecil jika dibandingkan dengan kesalahan pengklasifikasian dengan menggunakan penaksirklasik/MLE.

Tabel 4. Kesalahan Pengklasifikasian analisis diskriminan linier pada data simulasi dengan $n_1 = n_2 = 1000$

Persentase Outlier	Kesalahan pengklasifikasian (%)	
	MLE	*Fast-MCD*
5%	19,20	11,96
10%	23,85	11,27
15%	23,65	12,27
20%	23,55	11,13

Dari keempat variasi jumlah data simulasi untuk $n_1 = n_2 = 100$; $n_1 = n_2 = 200$; $n_1 = n_2 = 500$; $n_1 = n_2 = 1000$ dan dengan kondisi data terkontaminasi pengamatan *outlier* dari 5 sampai 20 persen, kesalahan pengklasifikasian fungsi diskriminan menggunakan penaksir dengan metode *fast-MCD* jauh lebih kecil jika dibandingkan kesalahan pengklasifikasian fungsi diskriminan menggunakan penaksir dengan metode klasik/*MLE*.

Hasil tersebut menunjukkan bahwa kesalahan pengklasifikasian dengan metode penaksir *fast-MCD* jauh lebih kecil jika dibandingkan dengan metode penaksir klasik/*MLE*. Atau dengan kata lain, ketepatan pengklasifikasian fungsi diskriminan linier dua kelompok dengan menggunakan penaksir *robust* fats-*MCD* jauh lebih baik dibandingkan fungsi diskriminan dengan penaksir klasik/*MLE*.

KESIMPULAN

1. Untuk mendapatkan taksiran fungsi diskriminan linier *robust* dua kelompok dengan metode *fast-MCD* diperlukan vektor rata-rata dan matriks kovariansi yang *robust* yang didapat dari penghitungan subsampel dari data dengan prosedur *c-steps.*
2. Analisis diskriminan linier *robust* dua kelompok dengan metode *fast-MCD* pada data simulasidengan beberapa tingkat *outlier*, memberikan tingkat ketepatan pengklasifikasian yang lebih baik jika dibandingkan dengan metode klasik/*MLE*

DAFTAR PUSTAKA

[1] Davies,L., 1992, The Asymptotics of Rousseuw's Minimum Volume

Ellipsoid Estimator. *The Annals of Statistics,20,1828-1843*

[2] Dillon,W.R dan Goldstein,M., 1984 *Multivariate Analysis Methods and Application,* John Wiley&Sons.Inc

[3] Hubert,M dan Driessen, 2002, Fast and Robust Discriminant Analysis¸ *Computational Statistics anda Data Analysis,45,301-320*

[4] Hubert,M dan Driessen, 2004, Syntax and Toolbox for Robust Discriminant Analysis, *http://wis.kuleuven.be/stat/robust.html*

[5] Hubert, M., Rousseuw,P.J, dan Aelst,S,. 2008, High Breakdown Robust Multivariate Method. *Statistical Science. vol 23 no.1, 92-119*

[6] Johnson, R.A dan Wichern,D.W., 2002, *.Applied Multivariate Statistical Analysis, fifth edition,*. Prantice Hall, New Jersey.

[7] Joossens, K., 2006, *Robust Discriminant Analysis,* Disertation Ph.D, Katholieke Universiteit Leuven, Leuven Nederland

[8] Pires,A.M,. 2002, Robust Linear Disciminant Analysis and the Projection Pursuit Approach*, Dept of Mathematics, Technical University of Lisbon,Portugal*

[9] Rousseuw,P.J., 1984, Least Median of Squares Regression, *Journal of The American Statistical Association,79,871880*

[10] Rousseuw,P.J.,Driessen., 1999, A Fast Algorithm for the Minimum Covariance Determinant Estimator , *Technometrics Vol.46, no.3, 293-305*

[11] Sharma, S., 1996, *.Applied Multivariate Techniques,* John Wiley&Sons.Inc

[12] Timm, N.H., 2002, *Applied Multivariate Analysis,* Springer-Verlag New York.Inc