S1

ISBN : 978-979-16353-5-6

PROSIDING
SEMINAR NASIONAL
MATEMATIKA DAN PENDIDIKAN MATEMATIKA

"Peningkatan Kontribusi Penelitian dan Pembelajaran Matematika dalam Upaya Pembentukan Karakter Bangsa"

Yogyakarta, 27 November 2010

Penyelenggara :
Jurusan Pendidikan Matematika FMIPA UNY
Kerjasama dengan
Himpunan Matematika Indonesia (Indo-MS)
wilayah Jateng dan DIY

Jurusan Pendidikan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Yogyakarta
2010

PROSIDING SEMINAR NASIONAL MATEMATIKA DAN PENDIDIKAN MATEMATIKA

27 November 2010 FMIPA Universitas Negeri Yogyakarta

Artikel-artikel dalam prosiding ini telah dipresentasikan pada
Seminar Nasional Matematika dan Pendidikan Matematika
pada tanggal ***27 November 2010***
di Jurusan Pendidikan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Yogyakarta

Jurusan Pendidikan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Negeri Yogyakarta
2010

Imputasi Berganda K-Medoid *General Regression Neural Network* Untuk Menangani *Missing Data*

Oleh :
Achmad Syahrul Choir[1], Brodjol Sutijo S.U[2]
[1]Mahasiswa Magister Jurusan Statistika ITS [2]Dosen Jurusan Statistika ITS
email : [1]madsyair@gmail.com[2] brodjol_su@statistika.its.ac.id

ABSTRAK

Missing data sering terjadi pada sebagian besar hasil survei. Analisis statistik yang didasarkan pada data lengkap tidak bisa diterapkan pada data survei yang di dalamnya terdapat missing data. Langkah ideal untuk menanganinya dengan imputasi. Metode imputasi berganda General Regression Neural Network (GRNN) adalah metode imputasi yang tidak berdasarkan asumsi distribusi tertentu. Dalam GRNN, data training proporsional dengan jumlah unit dalam hidden layer. Data training yang besar akan menghasilkan arsitektur jaringan yang besar sehingga menyebabkan kerugian dalam penerapannya. Penelitian ini mengaplikasikan metode pengelompokan K-Medoid untuk mengefisiensikan GRNN dan diterapkan dalam imputasi berganda pada kasus missing data tipe Missing at Random (MAR) dan Missing Completely at Random (MCAR) di Survei Tahunan Industri Besar Sedang. Hasilnya Imputasi Berganda K-medoid GRNN memerlukan waktu lebih singkat dibandingkan Imputasi Berganda GRNN dengan keakuratan yang lebih baik.

Kata Kunci: missing data, Imputasi Berganda, GRNN, K-medoid

1. Pendahuluan

Pada sebagian besar survei terdapat kejadian nonrespon, meskipun survei tersebut sudah direncanakan sebaik-baiknya. Nonrespon ini mengakibatkan adanya *missing data*. *Missing data* ini menjadi masalah ketika tiba saatnya untuk menganalisis data survei tersebut (Cohen, 1996). Analisis statistik yang didasarkan pada data lengkap tidak bisa diterapkan pada data survei yang di dalamnya terdapat *missing data* (Rubin, 1988).

Little dan Rubin (1987) mengklasifikasikan mekanisme *missing data* dalam tiga kategori: Pertama, *Missing Completely at Random* (MCAR) yang berarti bahwa terjadinya *missing data* tidak berkaitan dengan nilai semua variabel, apakah itu variabel dengan *missing value* atau variabel pengamatan. Ini berarti *missing data* terjadi secara acak. Kedua, *Missing at Random* (MAR). MAR berarti terjadinya *missing data* hanya berkaitan dengan variabel respon/pengamatan. Ketiga, *Non-Ignorable* (NI) berarti terjadinya *missing data* pada suatu variabel berkaitan dengan variabel itu sendiri, sehingga ini tidak bisa diprediksi dari variabel lain pada data set. Nilai pada *non ignorable* paling sulit diperkirakan daripada kedua metode lainnya. (Rubin, 1978).

Little dan Rubin (1987) menyatakan bahwa salah satu cara mengatasi *missing data* adalah dengan imputasi. Metode imputasi digolongkan menjadi imputasi tunggal dan imputasi berganda. Imputasi tunggal berarti setiap *missing value* digantikan dengan satu nilai. Sedangkan imputasi berganda adalah metode imputasi dengan mengganti setiap *missing value* dengan dua atau lebih nilai yang menggambarkan ketidakpastian nilai yang benar untuk diimputasikan.

Menurut Gheyas dan Smith (2009), metode imputasi yang paling bagus adalah metode yang berdasarkan model implisit. Keuntungan utamanya adalah metode ini memperhatikan keterkaitan di antara variable dan metode ini berdasarkan sedikit atau tidak ada asumsi yang berkaitan dengan data. Metode implisit yang saat ini banyak digunakan oleh peneliti adalah jaringan syaraf tiruan (*Neural Network*), yakni Jaringan *Multilayer Perceptron* (MLP) dan Jaringan *Radial Basis Function* (RBF). Namun

Makalah dipresentasikan dalam Seminar Nasional Matematika dan Pendidikan Matematika dengan tema "Peningkatan Kontribusi Penelitian dan Pembelajaran Matematika dalam Upaya Pembentukan Karakter Bangsa " pada tanggal 27 November 2010 di Jurusan Pendidikan Matematika FMIPA UNY

arsitektur dari MLP dan RBF cukup kompleks dan bobot yang dihitung banyak. Berbeda dengan MLP dan RBF, *General Regression Neural Network* (GRNN) memiliki arsitektur yang lebih sederhana.

GRNN adalah pengembangan dari jaringan syaraf tiruan, khususnya *Radial Basis Function* (RBF) (Specht,1991). Penelitian yang dilakukan Gheyas dan Smith (2009) menunjukkan bahwa imputasi dengan GRNN yang dimodifikasi lebih baik dibandingkan dengan *Markov Chain Monte Carlo* (MCMC). Yang dimaksud GRNN yang dimodifikasi pada penelitian Gheyas dan Smith (2009) adalah mengganti jarak Euclidean dengan jarak Mahalanobis.

Dalam GRNN, data *training* proporsional dengan jumlah unit dalam *hidden layer*. Data *training* yang besar akan menghasilkan arsitektur jaringan yang besar sehingga menyebabkan kerugian dalam penerapannya. Disarankan untuk dilakukan pengelompokan, misalnya K-mean dan fuzzy c-mean, sehingga jumlah unit adalah sebesar jumlah kelompok (Specht, 1991).

Metode pengelompokan yang berdasarkan sentroid (mean), seperti *fuzzy c-mean* dan K-Mean sensitif terhadap adanya outlier sehingga hasil pengelompokannya kurang tepat jika terdapat outlier (Binu Thomas dan Raju G (2009), Park dan Jun (2007)),. Menurut Park dan Jun (2007), metode K-Medoid digunakan untuk menangani hal ini karena K-Medoid *robust* terhadap keberadaan outlier. Hasil penelitian Park dan Jun (2007) memperlihatkan kinerja dari metode K-Medoid ini lebih baik dibandingkan metode K-Mean.

Berdasarkan uraian di atas, maka masalah dalam penelitian ini adalah bagaimana melakukan imputasi dengan menggunakan metode imputasi berganda dengan metode K-Medoid GRNN pada *missing data* tipe MCAR dan MAR ?

Berdasarkan permasalahan yang sudah dirumuskan di atas, maka penelitian ini bertujuan untuk mengetahui kinerja metode imputasi berganda K-Medoid GRNN untuk menangani *missing data* tipe MAR dan MCAR pada data yang mengandung outlier. Sedangkan manfaat yang ingin dicapai dari hasil penelitian ini adalah diperoleh metode imputasi berganda dengan pendekatan K-Mdoid GRNN untuk data yang besar dan terdapat outlier.

2. Tinjauan Pustaka

2.1 Pengertian *Missing Data*

Missing data berarti bahwa kita kehilangan beberapa tipe informasi tentang fenomena yang kita teliti. Keberadaan data hilang menghalangi kemampuan kita untuk menjelaskan dan memahami fenomena yang kita teliti *(McKnight,2007).* Menurut Little dan Rubin (1987), suatu data disebut hilang jika sebetulnya data bisa diperoleh apabila teknik penelitian didesain lebih baik.

2.2 *General Regression Neural Network*

General Regression Neural Network (GRNN) diusulkan dan dikembangkan oleh Specht (1991). GRNN didasarkan pada estimasi sebuah fungsi densitas peluang. Ini memanfaatkan model probabilistik di antara vektor random variabel independen $\mathbf{x}$ berdimensi p dan variabel random (skalar) dependen y. Jika $f(\mathbf{x}, y)$ mewakili fungsi densitas peluang bersama, nilai harapan y dengan syarat $\mathbf{x}$ (juga disebut regresi y pada $\mathbf{x}$) dapat diestimasi sebagai:

$$E(y|\mathbf{x}) = \frac{\int_{-\infty}^{\infty} y\mathrm{f}(\mathbf{x}, y)dy}{\int_{-\infty}^{\infty} \mathrm{f}(\mathbf{x}, y)dy} \quad (1)$$

Ketika densitas f(**x**,*y*) tidak diketahui, biasanya diestimasi dari pengamatan sampel x dan *y*. Untuk estimasi nonparametrik f(x,*y*), kita akan menggunakan estimator densitas kernel yang diperkenalkan Parzen (1962) dan oleh Cacoullos (1966) untuk kasus multidimensi. Estimator ini merupakan pilihan yang baik untuk mengestimasi fungsi densitas probabilitas, f, jika diasumsikan bahwa densitas adalah kontinyu dan turunan parsial pertama dari fungsi yang dievaluasi pada setiap **x** adalah kecil.
Pendekatan estimator densitas kernel terhadap persamaan 2.1 menghasilkan formula

$$\hat{Y}(\mathrm{x}) = \frac{\sum_{i=1}^{n} y_i exp\left[-\frac{{D_i}^2}{2\sigma^2}\right]}{\sum_{i=1}^{n} exp\left[-\frac{{D_i}^2}{2\sigma^2}\right]} \quad (2)$$

dimana

$$D_i = \sqrt{(\mathbf{x} - \mathbf{x}_i)'(\mathbf{x} - \mathbf{x}_i)} \quad (3)$$

adalah jarak Euclidean.

Dalam penelitiannya, Gheyas dan Smith (2009) mengganti jarak *Euclidean* dengan jarak *Mahalanobis*. Formulanya adalah

$$D_i = \sqrt{(\mathbf{x} - \mathbf{x}_i)'\mathrm{V}^{-1}(\mathbf{x} - \mathbf{x}_i)} \quad (4)$$

dimana V^{-1} adalah invers matrik kovarian. GRNN dengan fungsi jarak Mahalanobis disebut sebagai *Modified General Regression Neural Network (MGRNN).*

Permasalahan dari GRNN adalah jika jumlah pengamatan banyak sehingga tidak praktis menempatkan *neuron* tersendiri untuk setiap pengamatan. Metode pengelompokan dapat diterapkan sehingga setiap kelompok dapat diwakili satu *neuron* (Specht, 1991). Melalui pengelompokan, $\mathbf{x}_i$ diwakili nilai pusat/medoid kernel, c_i, i=1,2,..,k, dimana k adalah jumlah kelompok.

Metode K-mean telah diketahui sebagai teknik yang baik untuk pengelompokan. Namun, metode K-mean ini sensitif terhadap adanya *outlier*. Alternatifnya, metode K-medoid biasanya digunakan (Park dan Jun, 2009). Menurut penelitiannya algoritma ini menghasilkan kinerja yang baik dibandingkan K-mean dan dengan waktu yang lebih cepat.

Algoritma K-Medoid adalah sebagai berikut:

1. Tahap pertama (Pilih inisial medoid)

1-1 Hitung d_{ij}, jarak di antara setiap pasangan objek berdasarkan ukuran ketidaksamaan tertentu (misalnya Mahalanobis)

$$d_{ij} = \sqrt{\sum_{a=1}^{p} (x_{ia} - x_{ja})'V^{-1}(x_{ia} - x_{ja})} \quad (5)$$

Dimana i=1,…,n;j=1,…,n dan p adalah jumlah variable, dan V adalah matrik varian-kovarian

1-2 Hitung v_j untuk setiap objek j dimana

$$v_j = \sum_{i=1}^{n} \frac{d_{ij}}{\sum_{l=1}^{n} d_{il}} \quad (6)$$

1-3 Urutkan v_j dari terkecil ke terbesar. Pilih objek terkecil sebesar k sebagai inisial medoid.

1-4 Hitung jarak setiap objek dengan inisial medoid dan kelompokkan objek dalam k kelompok berdasarkan jarak minimal terhadap setiap medoid.

1-5 Hitung jumlah jarak dari semua objek ke medoid kelompoknya.

2. Tahap kedua (Update medoid)

Cari medoid baru pada setiap kelompok dimana jarak antar objek minimal . Update medoid setiap kelompok yang ada dengan medoid yang baru.

3. Tahap ketiga (Menghubungkan objek pada medoid)

3-1 Hitung jarak semua objek ke setiap medoid dan dihasilkan kelompok baru berdasarkan jarak minimal.

3-2 Hitung jarak semua objek ke medoid kelompoknya. Jika jumlahnya sama dengan jumlah sebelumnya, hentikan algoritma. Jika tidak kembali ke tahap 2.

Permasalahan lain GRNN adalah penentuan *smoothing factor.* Menurut Zhong (2007), *smoothing factor* sebaiknya menurut dimensi. Untuk mendapatkan nilai *smoothing factor* tersebut, Zhong, dkk (2007) mengusulkan metode estimasi *Gap-Based*. Hasil penelitian Zhong dkk ini menunjukkan estimasi *smooting factor* dengan metode *Gap-Based* lebih cepat, akurasi yang stabil dan baik. Rumus dari metode ini adalah:

$$\sigma_i = \min(4d, 0.5) \, . \, STD(\mathbf{x}_i) \tag{7}$$

dimana d adalah rata-rata nilai jarak minimum di antara dua titik input setelah distandarisasi dan $\mathrm{STD}(\mathbf{x}_\mathrm{i})$ adalah standar deviasi pada dimensi ke-i sebelum distandarisasi.

Dengan adanya pengelompokan dan *smoothing factor* yang berdasarkan dimensi, persamaan 2.2 menjadi

$$\hat{Y}(\mathbf{x}) = \frac{\sum_{i=1}^{k} y_i \frac{n_i}{\prod_{j=1}^{p} \sigma_{ij}} exp\left[-\frac{{C_i}^2}{2}\right]}{\sum_{i=1}^{k} \frac{n_i}{\prod_{j=1}^{p} \sigma_{ij}} exp\left[-\frac{{C_i}^2}{2}\right]} \tag{8}$$

Dimana

$$C_i = \sqrt{(\mathbf{x} - \mathbf{c}_i)' \mathrm{V}^{-1} {\Sigma_\mathrm{i}}^{-1} (\mathbf{x} - \mathbf{c}_i)} \tag{9}$$

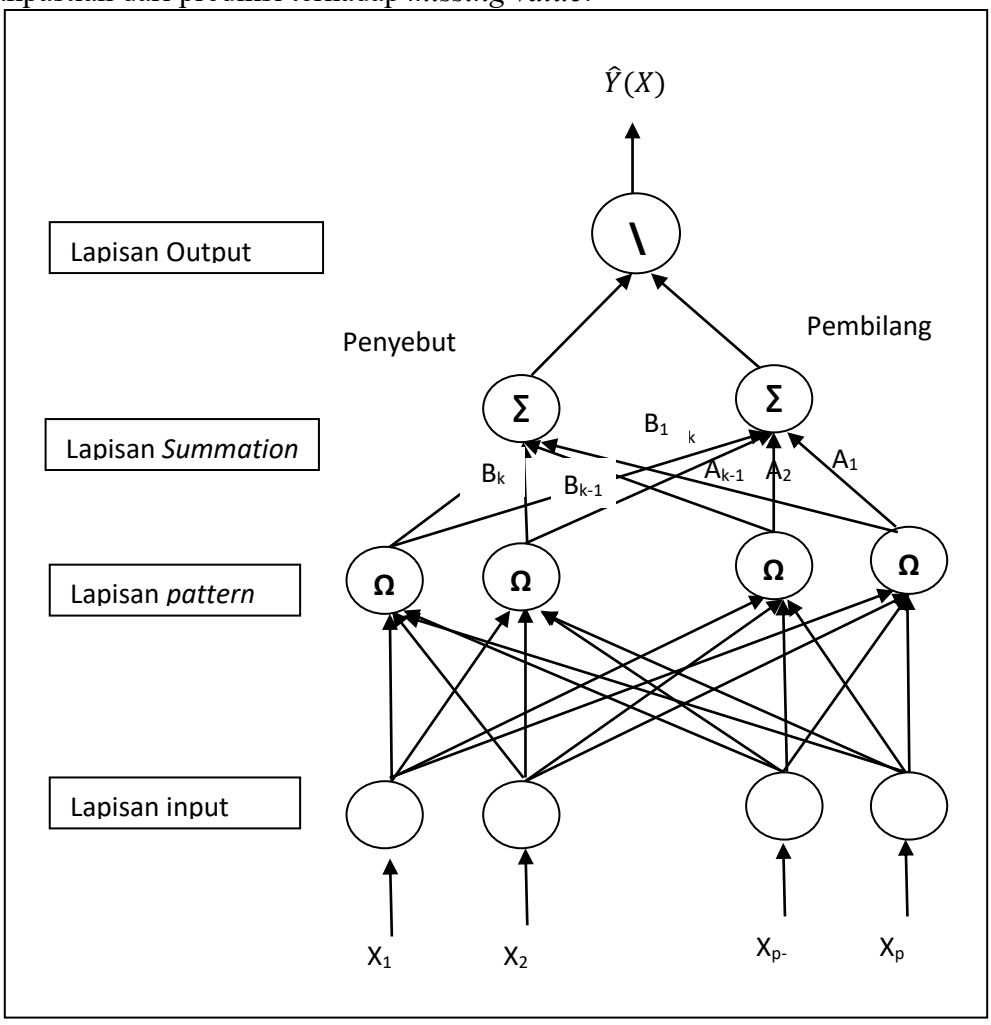
dan $\Sigma_\mathrm{i} = \begin{bmatrix} \sigma_{1\mathrm{i}} & 0 & \ldots & 0 \\ 0 & \sigma_{2\mathrm{i}} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \sigma_{\mathrm{pi}} \end{bmatrix}$

Berdasarkan persamaan 2.5 dibentuk arsitektur jaringan K-Medoid GRNN seperti ditunjukan pada gambar 1. Arsitektur K-Medoid GRNN tersebut memiliki empat lapisan, yaitu: (1) Lapisan input dengan jumlah unit sebanyak dimensi dari input. Lapisan ini yang terhubung langsung dengan lapisan *pattern* dengan bobotnya adalah nilai pusat dari kelompok (c_{ij}) dan *smoothing factor* (σ_{ij}),(2) Lapisan p*attern* yang memiliki unit sebanyak kelompok dengan fungsi aktivasi $exp\left[-\frac{{C_i}^2}{2}\right]$, (3) Lapisan *Summation* , memiliki dua unit, yaitu Pembilang dan Penyebut. Pembilang mempunyai bobot sebesar $\mathrm{A_i} = y_i \frac{n_i}{\prod_{j=1}^{p} \sigma_{ij}}$ dan dihitung dengan menjumlahkan perkalian antara

$exp\left[-\frac{{C_i}^2}{2}\right]$ dengan $A_{\text{i.}}$ Penyebut mempunyai bobot $B_i = \frac{n_i}{\prod_{j=1}^{p}\sigma_{ij}}$, kemudian dihitung dengan menjumlahkan perkalian antara $exp\left[-\frac{{C_i}^2}{2}\right]$ dengan $B_{\text{i.}}$ dan terakhir adalah lapisan *output*, merupakan pembagian pembilang oleh penyebut yang akan menghasilkan nilai prediksi.

2.3 Imputasi Berganda GRNN

Imputasi berganda GRNN dikembangkan oleh Gheyas dan Smith (2009) didasarkan pada imputasi berganda yang dikembangkan oleh Rubin (1987). Pada imputasi berganda, setiap *missing value* digantikan dengan satu set m>1 nilai yang dibangkitkan dari distribusi datanya. Variasi di antara m imputasi mencerminkan ketidakpastian dari prediksi terhadap *missing value*.



Gambar 1 Arsitektur K-Medoid GRNN

Ada tiga tahapan dalam imputasi berganda (Gheyas, 2010):

1. Imputasi
   Pada tahap ini, setiap *missing value* diimputasi beberapa (M) kali, sehingga menghasilkan M data yang dilengkapi. Pada imputasi berganda GRNN, M *training set* yang baru dibentuk secara acak dari training set asli. *Training set* yang berbeda menyebabkan model imputasi yang berbeda.

2. Analisis
   Analisis dilakukan pada setiap M data yang sudah dilengkapi. Hasil tahapan ini adalah mean dan varian sebanyak M. Misalnya $\hat{Q}_i$ dan $\hat{U}_i$ adalah estimasi mean dan varian dari data yang diimputasi ke-i, i=1…M.
3. *Pooling*
   Hasil M analisis digabungkan ke dalam hasil akhir. Hasilnya adalah
   a. Estimasi mean
   $$\bar{Q} = \frac{1}{M}\sum_{i=1}^{M}\hat{Q}_i \quad (10)$$
   b. Estimasi varian di dalam imputasi
   $$\bar{U} = \frac{1}{M}\sum_{i=1}^{M}\hat{U}_i \quad (11)$$
   c. Estimasi varian antar imputasi
   $$\bar{B} = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{Q}_i - \bar{Q}\right)^2 \quad (12)$$
   d. Total varian
   $$T = \bar{U} + \left(1+\frac{1}{M}\right)B \quad (13)$$

3. Metodologi

Penelitian ini menggunakan data sekunder dari Survei Tahunan Perusahaan Industri Besar dan Sedang Propinsi Jawa Timur Tahun 2008 sebagai data untuk studi kasus dari metode imputasi berganda K-Medoid GRNN. Variabel yang terdapat missing data adalah variabel produksi, dan yang menjadi variable input adalah Bahan bakar dan pelumas (X1) Banyaknya bahan baku dan bahan penolong (*X*3) serta julah tenaga kerja. Data survey ini digunakan karena terdapat outlier pada data tersebut (Basuki,2009).

Sesuai dengan tujuan penelitian, langkah-langkah dalam penelitian ini adalah sebagai berikut:

1. Melakukan imputasi berganda K-Medoid GRNN dan GRNN standar pada data Survei Tahunan Industri Besar dan Sedang Jawa Timur .Tahapannya adalah sebagai berikut
   a. Membuat simulasi *missing data*:
      Simulasi mekanisme MCAR dan MAR dibuat dengan menghapus beberapa nilai variabel produksi (Y) pada data Survei Tahunan Industri Besar dan Sedang yang lengkap. Berdasarkan algoritma Gheyas dan Smith (2009), prosedur simulasi pola data hilang MCAR dan MAR adalah:
      i. Pola data hilang MCAR, bilangan acak berdistribusi uniform (0,1) dibangkitkan pada setiap pengamatan dan tentukan persentase data yang akan dihilangkan. Nilai pengamatan (y_i) dihapus jika bilangan acak kurang dari persentase data yang dihilangkan.
      ii. Pola data hilang *MAR,* data dihilangkan sedemikian rupa sehingga nilai yang dihilangkan pada variabel Y tergantung pada variabel X_1, X_2, X_3 dan X_4. Untuk membuat simulasi data MAR, kita definisikan model non renponsif:
      $$p(y_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\beta_3 x_{3i})}} \quad (14)$$
      Dimana $p(y_i)$ adalah peluang menghapus y pada pengamatan ke-i. Data disusun berurutan sesuai peluang bahwa elemen data ini seharusnya hilang

pada variabel Y. Data berurutan ini dibagi menjadi dua bagian dengan ukuran yang sama. Jika persentase total data hilang adalah p, hilangkan persentase data yang berbeda dari masing-masing bagian. *Missing data* dari langkah 1 diimputasi dengan metode imputasi berganda K-medoid GRNN dan GRNN standar dengan fungsi jarak Mahalanobis dan *smoothing factor* yang digunakan metode *Gap-Based.*

b. Melakukan imputasi Berganda K-Medoid GRNN.

Imputasi missing data hasil simulasi dengan menggunakan algoritma imputasi berganda K-medoid GRNN (diadopsi dari algoritma imputasi berganda GRNN yang diusulkan Gheyas dan Smith (2009)). Algoritmanya adalah sebagai berikut:

i. Normalisasikan setiap variabel pada data sehingga setiap nilai berada pada interval 0 sampai dengan 1.
ii. Kodekan *missing value*, y_i, sebagai 'NaN'.
iii. Bagi data, D, menjadi dua bagian, yakni 1) D_{tes} = input tes = nilai semua variabel (X_1, X_2, X_3 dan X_4) pada pengamatan ke-i dan 2) D_{baru} = nilai semua variabel (X_1, X_2, X_3 dan X_4, Y) selain pengamatan ke-i. Variabel Y menjadi variabel output dan variabel X_1, X_2, X_3 dan X_4 menjadi variabel input.
iv. Tentukan M adalah jumlah imputasi, dalam hal ini 5
v. Buat M jaringan K-medoid GRNN $G^{(mean)}$ dan jaringan K-Medoid GRNN $G^{(var)}$ untuk mengestimasi mean dan varian y_i
   1) Buat data *training* D_p^{mean} secara acak sebanyak 70% dari D_{baru}.
   2) Lakukan *training* jaringan K-Medoid GRNN, G^{mean}, pada data *training* D_p^{mean}
   3) Evaluasi kinerja jaringan *training* G^{mean} pada data *training* D_p^{mean} dan estimasi barisan residual kuadrat **r**
   4) Buat data *training* D_p^{var} untuk jaringan G^{var} dengan menggunakan input pola D_p^{mean} *dan* **r** sebagai output
   5) Lakukan *training* jaringan G^{var} pada D_p^{var}
   6) Lakukan *testing* D_{tes} pada jaringan *training* G^{mean} untuk mem-prediksi mean $\hat{Q}^i$
   7) Lakukan testing D_{tes} pada jaringan *training* G^{var} untuk mem-prediksi varian $\hat{U}^i$
   8) Ulangi 1) sampai dengan 7) sebanyak M
   9) Hitung *mean* dari *missing value* (persamaan (2.18))
   10) Hitung estimasi varian *within-imputation* (persamaan (2.19))
   11) Hitung estimasi varian antar imputasi (persamaan (2.20))
   12) Estimasi varian total T (persamaan (2.21))
vi. Imputasi y_i dengan nilai mean
vii. Ulangi langkah vi dan vii untuk setiap nilai y_i

2. Membandingkan kinerja dari setiap metode dengan melihat waktu dan besaran *Mean Absolute Persentage Error* (MAPE). Formula MAPE adalah sebagai berikut:

$$MAPE = \left(\sum_{i=1}^{n_m} \left| \frac{(y_{asli\,i} - y_{imp\,i})}{y_{asli\,i}} \right| \right) \Big/ n_m \qquad (15)$$

Dimana:

n_m = banyaknya *missing value*

$y_{asli\ i}$ = nilai Y asal sebelum disimulasi *missing value* ke-i

$y_{imp\ i}$ = nilai Y hasil imputasi

4. Hasil dan Pembahasan

Langkah awal dari penelitian ini adalah membuat simulasi missing data tipe MCAR dan MAR dengan jumlah missing data sebesar 10 persen. Selanjutnya missing data tersebut diimputasi dengan imputasi berganda Kmedoid GRNN dan imputasi berganda GRNN dengan jumlah imputasi adalah 5. Hasilnya bisa dilihat pada tabel 1.

Tabel 1 MAPE dan waktu imputasi berganda pendekatan K-Medoid GRNN dan GRNN

Jenis Imputasi Berganda	MCAR		MAR	
	MAPE	Waktu (detik)	MAPE	Waktu (detik)
Kmedoid GRNN	7,36	37,49	8,09	38,64
GRNN	7,56	69,9	7,76	69,52

Hasil yang tercantum pada table 4.1 menunjukkan pada tipe MCAR, nilai MAPE untuk imputasi berganda K-Medoid GRNN sebesar 7,36 dan imputasi berganda GRNN sebesar 7,56. Waktu yang diperlukan untuk melakuka proses imputasi berganda K-medoid GRNN adalah 37,49 detik dan imputasi berganda GRNN sebesar 69,9 detik. Sementara itu, pada tipe MAR, nilai MAPE imputasi berganda K-Medoid GRNN sebesar 8,09 dengan waktu yang diperlukan 38,64 detik dan nilai MAPE imputasi berganda GRNN adalah 7,76 dengan waktu yang diperlukan 69,52 detik.

5. Kesimpulan dan Saran

Kesimpulan yang diperoleh dari simulasi missing data tipe MCAR pada data Survei Industri Besar dan Sedang Propinsi Jawa Timur adalah metode imputasi K-medoid GRNN memerlukan waktu yang lebih cepat dengan ketepatan (MAPE) yang lebih baik. Sedangkan pada tipe MAR metode imputasi K-medoid GRNN memerlukan waktu yang lebih cepat namun ketepatannya (MAPE) masih berada dibawah GRRN. Mengingat MAPE yang masih besar (lebih dari 1), perlu dikaji lebih lanjut mengenai GRNN ini yakni mengenai *smoothing factor* dan perlu dikaji metode pengelompokan yang lainnya.

6. Penghargaan

Penghargaan yang setinggi-tingginya penulis berikan Badan Pusat Statistik yang membiaya penulis untuk melanjutkan studi dan kepada dosen pembimbing yang telah banyak memberikan arahan dan masukan demi kesempurnaan penulisan ini serta berbagai pihak yang telah berjasa yang tidak dapat penulis rinci satu persatu.

7. Daftar Pustaka

Basuki, R. (2009), Imputasi Berganda Menggunakan Metode Regresi Dan Metode *Predictive Mean Matching* Untuk Menangani *Missing Data., T*hesis, Jurusan

Statistika, Fakultas MIPA Institut Teknologi Sepuluh Nopember Surabaya

Cacoullos, T., (1966). "Estimation of Multivariate Density", *Ann.Inst. Statistics.Math*, Vol.18 No. 2 , Hal 179-189.

Cohen, M.P, (1996), "A new approach to imputation". *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Hal 293-298.

Gheyas, I.A, (2009), *Novel Computationally Intelligent Machine Learning Algorithm for Data Mining and Knowledge Discovery*, Thesis Ph.D, Deparment of Computing Science and Mathematics. University Of Stirling, Scotland.

Gheyas, I.A dan Smith, L.S, (2009). "A Novel Nonparametric Multiple Imputation Algorithm for Estimating Missing Data". *Proceedings of the World Congress on Engineering* 2009 Vol II WCE 2009, 1-3 Jul 2009., London.

Little, RJA dan Rubin, DB, (1987), *Statistical Analysis with Missing Data*, John Wiley and Sons, New York.

McKnight,PE, McKnight, K.M., Sidani,S., Figueredo.,A.J, (2007), *Missing Data. A Gentle Introduction*, Guilford Press, NewYork.

Park,H.S, Jun,C.H, (2009). "A Simple and Fast Algorithm for K-Medoid Clustering", *Expert System With Application* 36. Hal 3336-3341.

Parzen, E, (1962), "On Estimation of Probability Density Function and Mode". *Ann. Math. Statist*, Vol.33, hal 1065-1076.

Rubin, D. B, (1978), "Multiple Imputations In Sample Surveys -A Phenomenological Bayesian Approach to Nonresponse", *The Proceedings of the Survey Research Methods Section of the American Statistical Association*, Hal 20–34.

Rubin, D.B, (1988), "An Overview of Multiple Imputation", *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79-8.

Specht, D.F, (1991), "A General Regression Neural Network", *IEEE Transactions on Neural Networks*, 2(6),hal . 568-576.

Thomas,B dan G,Raju, (2009), "A Novel Fuzzy Clustering Method for Outlier Detection in Data Mining", *International Journal of Recent Trends in Engineering*, Vol. 1, No. 2, Mei 2009, Academic Publisher, Hal 161-165