Jurnal As-Salam, 1(1) Januari - April 2017
(Print ISSN 2528-1402, Online ISSN 2549-5593)

Optimisasi Dengan Adanya *Big data Problem*

Muhammad Huda Firadaus

Mahasiswa Pasca Sarjana Matematika FMIPA USU
Email: m.huda.firdaus86@gmail.com

Abstrak. Penelitian ini bertujuan menangani jumlah data yang banyak didalam permasalahan identifikasi keanggotaan. Sistem identifikasi digunakan dalam statistik untuk membentuk model matematika sistem dinamis dari suatu data. Akan didefinisikan istilah *big data* secara khusus yaitu *big data* akan ditunjukkan melalui banyaknya jumlah dari *input-output* data, kemudian data *input-output* dimodelkan, dari model ini kemudian diselesaikan untuk memperoleh estimasi parameter yang optimal. Model diselesaikan dengan teori *semidefinit programming* (SDP). Contoh perhitungan disajikan.

Kata kunci: *big data*, *set-membership* identifikasi.

Pendahuluan

Sistem identifikasi/estimasi digunakan dalam Teknik kontrol dengan permasalahan membuat pemodelan matematika dari sistem dinamis, sekumpulan data dan beberapa informasi tersebut yang akan diidentifikasi untuk selanjutnya dimodelkan. Suatu data berisi rata-rata prosedur pengukuran dari data yang tidak ditentukan. Karena data tidak ditentukan maka yang menjadi persoalan adalah menaksir data tersebut agar diperoleh model matematika, cara yang dapat dilakukan untuk data yang tidak ditentukan adalah menentukan data tersebut dan membuatnya menjadi data yang jumlahnya besar.

Walaupun diketahui bahwa mengumpulkan data/informasi merupakan persoalan dasar untuk memodelkan. Dengan estimasi/perkiraan, suatu kumpulan data yang banyak masih tidak dapat diselesaikan melalui cara komputer. Kata kunci *big data* muncul menjadi suatu permasalahan terbaru, yang terdiri dari sekumpulan data yang besar dan komplek dan sulit untuk di proses dan di identifikasi.

Tujuan penelitian ini adalah mendeskripsikan *set-membership* identifikasi untuk permasalahan *big data* agar dapat di modelkan.

Metode Penelitian

Istilah *big data* telah digunakan sejak tahun 1990-an, dengan memberikan kredit kepada Jhon Mashey untuk *coining* atau setidaknya membuatnya populer. *Big data* isinya mencakup set data dengan ukuran diluar kemampuan perangkat lunak yang biasa digunakan untuk menangkap, pendeta, mengelola, dan memproses data dalam waktu yang telah berlalu toleransi. *Big data* (ukuran) adalah target terus bergerak, seperti 2012 mulai dari beberapa lusin terabyte banyak *petabyte* data. Data besar membutuhkan seperangkat teknik dan teknologi dengan bentuk-bentuk baru dari integrasi untuk mengungkapkan wawasan dari data set yang beragam, kompleks, dan skala besar.

Dalam laporan penelitian di tahun 2001 oleh META Group/Gartner analis Doug Laney di defenisikan tantangan pertumbuhan data dan peluang sebagai tiga dimensi, yaitu meningkatkan volume (jumlah data), kecepatan (kecepatan data masuk dan keluar), dan jenis (berbagai

jenis data dan sumber). Sekarag banyak industri menggunakan 3 V model untuk menggambarkan data besar . Istilah 3 V telah diperluas untuk karekter pelengkap lainnya dari data besar. karakter-karakter *big data* antara lain:

- Volume: Data besar bukan sampel yang hanya melacak dan mengamati apa yang terjadi
- *Velocy*: data besar sering tersedia secara real time
- *Variety*: data besar menarik dari teks, gambar, audio, video ditambah bahagian yg hilang melalui fusi data

Pertumbuhan *big data* menggambarkan konsep lebih luas dengan melukiskan perbedaan antara *big data* real dan *big data* dalam *business intelligence*. *Big data* yang di bicarakan pada *business intelligence* terbagi dua yaitu statistik deskriptif dan statistik induktif.

Big data menggunakan statistik deskriptif yaitu data yang kepadatan tinggi informasi untuk mengukur hal, mendeteksi tren (sampel). *Big data* menggunakan statistik induktif dan konsep sistem non linier identifikasi untuk menyimpulkan hukum (regresi, hubungan nonlinier, dan efek kausal) dari set data yang besar dengan kepadatan informasi rendah, untuk mengungkapkan hubungan dan ketergantungan, atau untuk melakukan pediksi hasil dan perilaku.

Jadi *big data* yang di bicarakan pada penelitian ini adalah pada busines intelegensi dengan menggunakan statistik induktif. Penelitian ini bersifat studi literature ataupun studi kepustakaan dengan mengacu pada jurnal-jurnal internasional yang berhubungan dengan sistem identifikasi/estimasi.

Hasil dan pembahasan

Unsur pokok dari *set-membership* estimation dibuat dengan referensi dari permasalan yang disebut permasalahn *error-in-variables* (EIV), sederhana tetapi komplek yang menjadi fokus pembahasan yaitu masalah linier EIV dan untuk waktu yang sama, konsep dasar ini sedererhana tetapi kompleks, dari sudut pandang kom- pleksitas pada komputer, yang menjadi konsep utama yang akan di bahas.

Dalam sistem kontrol ada input dan ada output, anggap sistem kontrol yang di bahas yaitu sistem *single-input* dan single-output (SISO) *linier-time-invariant* (LTI) (Cerone 1993) dimana x_t adalah masukan barisan *noise-free* dan *linier block* dimodelkan dengan sistem waktu diskrit yang merubah x_t menjadi keluaran *noise-free* w_t menurut persamaan beda/*difference*.

$$A(q^{-1})\ w_t = B(q^{-1})\ x_t$$

dimana A(.) dan B (.) adalah polinomial didalam operator terbalik q^{-1} ($q^{-1}w_t = w_{t-1}$) dari bentuk

$A(q^{-1}) = 1 + a_1q^{-1} + \ldots + a_{na}q^{-na}$

$B(q^{-1}) = b_0 + b_1q^{-1} + \ldots + b_{nb}q^{-nb}$

Input dan output dari barisan data dirusak oleh penambahan noise ξ_t dan η_t di tulis

$$u_t = x_t + \xi_t$$

$$y_t = w_t + \eta_t$$

Dimana ξ_t dan η_t dianggap interval yang di berikan batasan $\Delta\xi_k$ dan $\Delta\eta_k$ di tulis

$$|\xi_t| \leq \Delta\xi_k$$

$$|\eta_t| \leq \Delta\eta_k$$

Parameter vector θ tidak diketahui dan $\theta \in \mathbb{R}^p$ untuk diidentifikasi dan didefenisikan sebagai

$$\theta = [\ a_1 \ldots a_{na}\ \ b_0\ \ b_1 \ldots\ b_{nb}]^T$$

Dimana $na + nb + 1$ = p ketika feasible parameter set (FPS) $\mathfrak{D}_\theta$ adalah

$\mathfrak{D}_\theta = \{ \ \theta \in \mathbb{R}^p$: A(q^{-1}) $(y_t - \eta_t) =$ B(q^{-1}) $(u_t - \xi_t)$,

$|\xi_t| \leq \Delta\xi_k$, $|\eta_t| \leq \Delta\eta_k$; t = 1,...,N$\}$

Dimana N = panjang barisan data

Model di atas menggambarkan himpunan semua nilai yang mungkin dari parameter θ tetap yang tidak diketahui dari ukuran data, batasan nilai error dan dianggap suatu model.

$\mathfrak{D}_\theta = \{ \ (\theta, \xi, \eta) \in \mathbb{R}^p$ x $\mathbb{R}^N$ x $\mathbb{R}^N$: A(q^{-1}) $(y_t - \eta_t) =$ B(q^{-1}) $(u_t - \xi_t)$, $|\xi_t| \leq \Delta\xi_k$, $|\eta_t| \leq \Delta\eta_k$; t = 1,...,N$\}$

$\mathfrak{D}_\theta = \{ \ (\theta, \xi, \eta) \in \mathbb{R}^p$ x $\mathbb{R}^N$ x $\mathbb{R}^N$: $\sum_{i=1}^{na} a_i (y_{t-i} - \eta_{t-i}) = \sum_{j=0}^{nb} b_j (u_{t-j} - \xi_{t-j})$,

$|\xi_t| \leq \Delta\xi_k$, $|\eta_t| \leq \Delta\eta_k$; t = 1,...,N$\}$

Dimana $\xi = [\xi_1, \xi_2, \dots, \xi_N]^T \in \mathbb{R}^N$ dan

$\eta = [\eta_1, \eta_2, \dots, \eta_N]^T \in \mathbb{R}^N$

Jadi model yang akan dislesaikan adalah untuk menhitung fungsi tujuan yang disebut parameter uncertainty intervals PUI_j di defenisikan sebagai berikut:

$PUI_j = [\ \underline{\theta}_j , \overline{\theta}_j\ ]$ untuk j = 1, ..., p

Dimana

$$\underline{\theta}_j = \min_{\theta, \xi, \eta \ \in \ \mathfrak{D}_{\theta,\xi,\eta}} \theta_j$$

$$\overline{\theta}_j = \max_{\theta, \xi, \eta \ \in \ \mathfrak{D}_{\theta,\xi,\eta}} \theta_j$$

Dari model diatas fungsi tujuan nya adalah PUI_j dan kendalanya adalah $\underline{\theta}_j$ dan $\overline{\theta}_j$ Karena $\mathfrak{D}_{\theta,\xi,\eta}$ sebuah himpunan non konvex

$(y_t - \eta_t) + \sum_{i=1}^{na} a_i (y_{t-i} - \eta_{t-i}) = \sum_{j=0}^{nb} b_j (u_{t-j} - \xi_{t-j})$

Contoh penanganan *big data* dengan menggunakan formula / model diatas:

Anggap system order 2 dengan nilai parameter sebenarnya θ^T = [a_1 a_2 b_1 b_1] = [-1.75 0.87 -1.25 1.88] batasan parameter yang akan di evaluasi untuk kumpulan data yang akan di uji dari N panjang persebaran data yaitu untuk N = 30,N = 200 dan N = 1000. Perhitungan Numerik di lakukan pada Matlab .

Sistem dijalankan dengan barisan input data random x_t menggunakan distribusi uniform antara [$-1, +1$]. Data input x_t dan data output di rusak oleh penambahan *random noise* yaitu ξ_t dan η_t, distribusi uniform antara $[-\Delta\xi_t \ , +\Delta\xi_t\ ]$ dan $[-\Delta\eta_t , +\Delta\eta_t\ ]$. Pemilihan batasan error $\Delta\xi_t$ dan $\Delta\eta_t$ agar sinyal *noise rasio* dari input SNR_x = 10 log { $\sum_{t=1}^{N} x_t^2 / \sum_{t=1}^{N} \xi_t^2$} dan output SN$R_w$ = 10 log { $\sum_{t=1}^{N} w_t^2 / \sum_{t=1}^{N} \eta_t^2$} antara 20 db dan 22 db. Taksiran utama dirumuskan θ_j^{cs} =($\overline{\theta}_j^s$ + $\underline{\theta}_j^s$) / 2 dan batas parameter tak tentu $\Delta\theta_j^{cs}$ =($\overline{\theta}_j^s - \underline{\theta}_j^s$) / 2.

Tabel 1. Perkiraan parameter utama (θ_j^{cs} , $\theta_j^{c,\delta}$) dan batasan tak tentu ($\Delta\theta_j^s$, $\Delta\theta_j^\delta$) terhadap banyak data N. Nilai dari $\theta_j^{c,\delta}$ dan $\Delta\theta_j^\delta$ di hitung untuk relaxion tingkat δ = 2

N	Nilai sebenarnya	θ_j^{cs}	$\Delta\theta_j^s$	$\theta_j^{c,\delta}$	$\Delta\theta_j^\delta$
30	-1.760	-2.790	2.016	-1.773	0.218
	0.870	1.544	1.328	0.889	0.222
	-1.250	-2.031	1.441	-1.436	0.712
	1.880	3.048	2.348	2.132	1.017
200	-1.760	-2.088	.123	1.741	.150
	0.870	1.020	0.566	0.860	0.149
	-1.250	-1.759	1.110	-1.376	0.679

	1.880	2.358	1.506	2.062	0.925
1000	-1.760	-2.011	0.965	-1.759	0.091
	0.870	0.955	0.467	0.864	0.094
	-1.250	-1.447	0.769	-1.316	0.462
	1.880	2.369	1.326	2.106	0.515

Simpulan dan Saran

1. Kesimpulan

Sistem identifikasi digunakan dalam statistik untuk membentuk model matematika sistem dinamis dari suatu data. Pembahasan terpenting yang akan di paparkan dalam tulisan ini yaitu pemodelan (model identifikasi) dan penanganan *big data*. Intinya adalah bagaimana mendapatkan model sistem/sinyal jika dihadapkan (diperlukan) jumlah data yang sangat besar.

Teknik moment/Sum of squares relax diperlukan untuk menyelesikan medel matematis yang berbentuk polinomial (konvek optimisasi) dengan teori *Semidefinite Programming* (SDP). Hasilnya dapat dilihat semakin besar N maka makin mendekati nilai interval parameter model dari nilai parameter sebenarnya. Interval batas bawah diperoleh dengan meminimumkan fungsi tujuan model Parameter *Uncertainty Intervals* (PUI_j) dan interval atas dengan memaksimumkan fungsi tujuan model *Parameter Uncertainty Intervals* (PUI_j).

2. Saran

Ada beberapa metode untuk menyelesaikan optimisasi dari *big data* yang telah di modelkan secara matematis, tapi di sarankan menggunakan metode *semidefinite programming* (SDP), karena sdp dapat menangani permasalahan yang nonkonvek. Pembahasan dilakukan dari jumlah N terkecil yaitu 30 dan N terbesar yaitu 1000 disarankan menggunakan N yang besar lagi, agar di peroleh estimasi interval yang mendekati nilai parameter sebenarnya.

Daftar Pustaka

Brockwell, P.J., dan Davis, R.A. (2002). Introduction to Time Series and Forecasting, Second Edition. Spriger.

Cerone, V., Piga, D., and Regruto, D. (2011). Improved parameters bounds for set-membership EIV problems. International Journal of Adaptive Control and Signal Processing, 57(2), 208227.

Cerone, V., Piga, D., and Regruto, D. (2012). SetMembership Error-in-Variables Identification Through Convex Relaxation Techniques.IEEE Transactions on Automatic Control, 57(2), 517522.

Cerone, V.dan Regruto, D. (2015). Handling *Big data* in set-membership identifi- cation through a sparse optimization approach. IFAC Papers On Line, 48-28, 12721278.

Hillier, F. S. dan Lieberman, G. J. (2000). Introduction To Operations Research Seventh Edition. Stanford Unversity.

Lasserre, J.B. (2006). Convergent semidefinite relaxationsin polynomial optimization with sparsity. SIAM J.Optimiz, 17(1), 822843.

Schweppe, F., (1968).Recursive State Estimation: Unknown but Bounded Errors and System Inputs. IEEE Trans. Aut. Control, AC-13, 556-558

Bisgaard, S. dan Kulahci, M. (2011). Time Series Analysis and Forecasting by Example. Wiley.

Taha, H. A., (2007). Operations Research Am Introduction Eighth Edition. University of Arkansas, Fayetteville.