Kombinasi Linier Target Data Untuk Regresi Multitarget Menggunakan Principal Component Analysis

by LPPM STT Nurul Fikri

Submission date: 20-Jun-2022 12:29PM (UTC+0900)
Submission ID: 1787358988
File name: egresi_Multitarget_Menggunakan_Principal_Component_Analysis.docx (497.03K)
Word count: 3579
Character count: 21824

Jurnal Teknologi Terpadu

https://journal.nurulfikri.ac.id/index.php/jtt
ISSN : 2477-0043 ISSN ONLINE : 2460-7908

KOMBINASI LINIER TARGET DATA UNTUK REGRESI MULTITARGET MENGGUNAKAN PRINCIPAL COMPONENT ANALYSIS

Yonathan Purbo Santosa[1]

[1]Teknik Informatika, Universitas Katolik Soegijapranata
Semarang, Jawa Tengah, Indonesia 50275
yonathansantosa@unika.ac.id

Abstract

Linear Regression is a method to predict numbers which is a dependent variables (output) based on some independent variables (inputs). The problem with regression is that some data does not fall into linear problem. Based on this problem, a method named RLC was invented to randomly finds correlation between output by projecting the data into the higher dimension. This method does not provides ways to inverse the projection, resulting in poor performance of the results. On top of that, by projecting the data into the higher dimension will increase the learning algorithm complexity. Instead, using PCA we will alleviate by projecting the data into a lower dimension while preserving the ability to inverse the projection. The results shows that PCA can achieve overall lower error compared to RLC.

Keywords: *Linear regression, multitarget regression, multidimension regression, dimension reduction, PCA*

Abstrak

Regresi linier adalah metode untuk memprediksi sebuah nilai (variabel dependen) berdasarkan beberapa input (variabel indepenen). Permasalahan pada regresi linier adalah beberapa data tidak termasuk kedalam kategori linier. Sehingga sebuah metode bernama RLC diciptakan untuk menemukan korelasi antara data output dengan cara memproyeksikan data ke dalam dimensi yang lebih tinggi. Sayangnya, metode ini tidak memberikan cara untuk melakukan proses invers dari proyeksi tersebut sehingga menyebabkan performa dari regresi linier menjadi lebih buruk. Selain itu, dengan memproyeksikan data ke dimensi yang lebih tinggi akan menambah kompleksitas dari algoritma pembelajaran. Oleh karena itu, PCA akan digunakan untuk memecahkan masalah ini dengan cara memproyeksikan data ke dimensi yang lebih rendah sembari mempertahankan kemampuan untuk melakukan invers proyeksi. Hasilnya, PCA masih dapat mempertahankan error yang lebih rendah secara keseluruhan dibandingkan dengan RLC.

Kata kunci: Regresi linier, regresi multitarget, regresi multidimensi, reduksi dimensi, PCA

1. PENDAHULUAN

Dalam analisis statistik, ada metode yang dapat digunakan untuk memprediksi suatu nilai dependen (*output)* berdasarkan beberapa data yang diberikan sebagai data *input*. Metode ini dikenal sebagai metode regresi. Regresi mencari dan menghitung korelasi antara variabel dependen (output) berdasarkan beberapa variabel independen (input) untuk mendefinisikan suatu fungsi yang akan memetakan input ke output [1]. Karena data yang dikumpulkan mungkin memiliki *noise*, kita hanya dapat melakukan fungsi aproksimasi yang akan memetakan input ke output dengan menemukan deviasi minimum antara fungsi dan variabel dependen. Dalam kasus linier, korelasi antara variabel dependen dan parameter fungsi dapat diselesaikan dengan fungsi linier seperti fungsi garis, yang biasanya akan diselesaikan dengan menggunakan estimasi *least square* [1]

Di sisi lain, untuk masalah yang lebih kompleks, yang terdiri dari input dan output yang multi-dimensi seringkali tidak termasuk dalam kategori yang dapat diselesaikan secara linier sehingga tidak dapat diselesaikan dengan menggunakan estimasi *least square*. Terlebih, metode regresi hanya mampu menghasilkan satu nilai prediksi sehingga untuk kasus variabel dependen multi-dimensi diperlukan beberapa model regresi. Padahal banyak sekali kasus dalam dunia industri yang memerlukan metode prediksi untuk variabel dependen dengan multi-dimensi. Terutama dimasa digitalisasi yang menyebabkan data

Diterima (tanggal), Direvisi (tanggal), Diterima untuk publikasi (tanggal)

dengan dimensi yang banyak adalah hal yang wajar [2]. Didorong oleh kebutuhan untuk memecahkan masalah regresi non-linier dan multi-dimensi, mendorong peneliti di seluruh dunia untuk menemukan metode penyelesaian data multi-dimensi yang non-linier dengan metode pembelajaran yang lebih kuat, seperti jaringan syaraf tiruan [3], *deep regression* [4], [5], *random linear target combination* [6], *variational autoencoder regression* [7], dan lainnya.

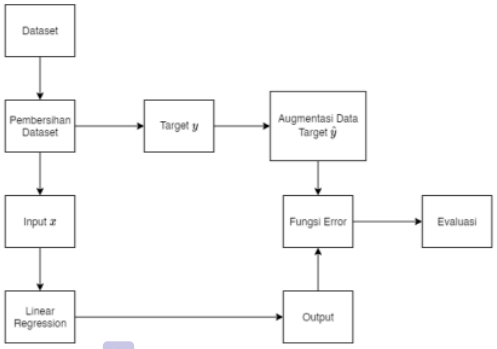
Seiring dengan bertambahnya kemampuan komputer dalam memproses data, kebutuhan untuk memecahkan permasalahan regresi multi-dimensi secara akurat dan cepat pun semakin bertambah [8]. Salah satu metode yang sering digunakan adalah dengan menggunakan regresi *single target* (ST) untuk kemudian digabung menjadi sebuah model besar menggunakan metode ensemble [8]. Metode serupa digunakan oleh Boye et al. [9] untuk melakukan estimasi harga unit rumah. Herawati et al. [10] membandingkan beberapa metode regresi kembangan untuk menyelesaikan masalah multikolinier dalam data. Hasilnya menunjukkan Principal Component Regession adalah metode dengan nilai error terrendah [10]. Dalam penelitian yang dilakukan oleh Tsoumakas et al. [6] mengusulkan sebuah metode untuk memberikan relasi yang random antar variabel dependen yang dinamakan *random linear target combination* (RLC) pada regresi ST. Metode ini merupakan turunan dari metode untuk klasifikasi multi-label RA*k*EL [6], [11]. Meskipun menggunakan kombinasi random, hasil dari penelitian tersebut menunjukkan performansi yang lebih baik (dinilai dari nilai error) dibandingkan dengan *state-of-the-art* sebelumnya. Namun pada kenyataannya, RLC mentransformasi data target ke dimensi yang lebih tinggi, sehingga dapat menyebabkan berkurangnya stabilitas dari sebuah model yang sering dikenal dengan istilah *curse of dimensionality* di dalam data mining maupun machine learning [12]. Data dengan dimensi yang lebih tinggi ini akan berpengaruh kepada kecepatan pelatihan dari model *regressor* serta dapat meningkatkan variasi sehingga mengurangi performa dari model *regressor*. Selain permasalahan diatas, proses transformasi data target dengan metode RLC menggunakan perkalian matriks yang tidak dapat dilakukan transformasi invers nya. Sehingga untuk mengembalikan hasil regresi ke nilai aslinya diperlukan metode aproksimasi dari invers matriks tersebut agar proses transformasi invers tersebut dapat berjalan. Karena metode transformasi tersebut adalah aproksimasi, keakuratan dari model *regresi* akan tergantung dari keakuratan proses aproksimasi invers matriks tersebut.

Dari kedua permasalahan diatas, penelitian ini dilakukan dalam rangka mencari metode pengganti yang bekerja dengan cara mengurangi kompleksitas dari data untuk mempercepat proses pelatihan dari model *regressor* dan mencari metode lain yang memiliki proses transformasi invers tanpa aproksimasi untuk menjaga keakuratan dari metode transformasi tersebut. Kedua hal tersebut sekaligus untuk menghasilkan suatu model *regressor* yang mampu menandingi atau melebihi kemampuan dari model *regressor* yang sudah ada untuk melakukan regresi.

2. METODE PENELITIAN

Untuk melakukan analisa terhadap efek dari perubahan dimensi pada data target, baik dari penelitian sebelumnya yaitu RLC dan usulan metode yang akan digunakan pada penelitian ini, maka penelitian ini dirancang seperti pada **Error! Reference source not found.** yang akan dijabarkan secara lebih mendalam pada bagian berikutnya.

Implementasi dari uji coba metode akan dilakukan menggunakan bahasa pemrograman *Python* dan *library machine learning scikit-learn* [13]



Gambar 1. *Diagram Alir Penelitian*

2.1 Pengumpulan Dataset

Dalam pengujian metode yang akan dilakukan, diperlukan beberapa dataset yang digunakan sebagai alat dalam memperoleh hasil pengukuran metode yang digunakan. Mengacu kepada penelitian yang dilakukan Tsoumakas dkk. [6], 12 dataset akan digunakan pada penelitian ini seperti yang dijabarkan pada **Error! Reference source not found.**. Pada **Error! Reference source not found.**, masing-masing dataset memiliki data target dengan ukuran dimensi q.

Tabel 1. Dataset dengan Deskripsi

Nama	Singkatan	$\|D\|$	q
Airline Ticket Price 1 [11]	atp1d	337	6
Airline Ticket Price 2 [11]	atp7d	296	6
Occupational Employment Survey 1 [11]	oes10	334	16
Occupational Employment Survey 2 [11]	oes97	403	16
River Flow 1 [11]	rf1	4165/5065	8
River Flow 2 [11]	rf2	4165/5065	8
Supply Chain Management 1 [11]	sf1978	8145/1658	16
Supply Chain Management 2 [11]	sf1969	7463/1503	16
Electrical Discharge Machining [14]	edm	154	2
Solar Flare 1 [15]	scm1d	323	3
Solar Flare 2 [15]	scm20d	1066	3
Water Quality [16]	wq	1060	14

Data target dari masing-masing dataset akan ditransformasikan menggunakan metode PCA dan RLC untuk menguji efektifitas kedua algoritma tersebut dalam menangani permasalahan regresi multitarget.

2.1. Pembersihan dan Pemrosesan Awal Dataset

Setelah data terkumpul, data yang didapatkan harus diproses terlebih dahulu untuk membersihkan data dari data yang tidak lengkap dan data yang tidak dapat diolah oleh regresi seperti data yang berupa text dan nominal.

Selain pembersihan data, untuk membuat variasi dari masing-masing variabel pada data, digunakan pemrosesan awal yaitu *MinMaxScaler* seperti yang dijabarkan pada persamaan (1) dan (2).

$$X_\sigma = \frac{x - min(X)}{max(X) - min(X)} \quad (1)$$

$$\hat{x} = X_\sigma * (max(X) - min(X)) + min(X) \quad (2)$$

2.2. Augmentasi Data Target menggunakan RLC

Dalam melakukan augmentasi data target dua metode akan digunakan dalam penelitian ini. Metode yang pertama adalah metode RLC yang merupakan landasan dalam penelitian ini dan metode PCA sebagai alternatif yang akan menggantikan RLC. Metode RLC sendiri berusaha menemukan korelasi linier antar data target dengan cara menginisialisasi sebuah matriks kombinasi C secara random dengan ketentuan sebagai berikut:

1. Pilih ukuran dimensi data target yang baru r. Pada penelitian sebelumnya, parameter r yang digunakan adalah 500 [6].
2. Pilih jumlah data target dari data original yang akan dikorelasikan k. Pada penelitian sebelumnya, parameter k berkisar antara 2 hingga q.
3. Buat matriks kosong C dengan ukuran $q \times r$
4. Untuk masing-masing kolom pada C, pilih k buah baris dan isikan nilai dengan nilai random yang diambil dari distribusi uniform dengan rentang $[0,1]$.

Proses kombinasi target yang dijabarkan pada dilakukan untuk semua *training data*, sehingga akan dihasilkan sebuah matriks baru $\hat{Y}$ yang akan digunakan sebagai data target dalam melakukan proses regresi seperti yang diilustrasikan dalam **Error! Reference source not found.**. Tsoumakas dkk. [6] tidak melakukan proses inverse untuk merubah $\hat{Y}$ menjadi Y kembali. Pada penelitian ini akan dilakukan proses inverse menggunakan MPPI terhadap matrix C untuk menemukan C^+. Asumsi tersebut didasarkan bahwa ada sebuah matriks C^+ yang memenuhi persamaan berikut:

Jika persamaan (3) adalah *singular value decomposition* (SVD) untuk matriks C, maka dapat dihitung pula C^+ menggunakan SVD.

$$C = Q_1 \Sigma Q_2^\top \quad (3)$$

$$\Sigma^+ = \frac{1}{\Sigma} \quad (4)$$

$$C^+ = Q_1 \Sigma^+ Q_2^\top \quad (5)$$

Dimana Q_1 dan Q_2 adalah matriks ortogonal, Σ adalah matriks diagonal yang berisikan singular value dari C, Σ^+ adalah matriks diagonal yang berisikan singular value dari C, yang masing-masing nilainya adalah nilai kebalikan dari Σ Sehingga nilai dari C^+ memenuhi sistem persamaan linier (6) dan (7), kemudian dari persamaan (5) dan (7), nilai regresi sesungguhnya dapat ditarik kembali seperti pada persamaan (8).

$$Cx = b \quad (6)$$

$$x = C^+ b \quad (7)$$

$$Y = \hat{Y} \times C^+ \quad (8)$$

Dari formula di atas, dapat diilusrasikan pada **Error! Reference source not found.**, dimana sebagai contoh terdapat 3 buah baris data dengan 2 data target. Kemudian dibuatlah sebuah matriks C yang diisi dengan nilai random. Hasil perkalian antara dataset dengan matriks C kemudian akan digunakan untuk melakukan proses pelatihan model regresi linier.

1	2
4	1
3	4

(a)

0.2	0	0	0.5	0.3	0.3
0.1	0.4	0.1	0	0.2	0

(b)

0.4	0.8	0.2	0.5	0.9	0.3
0.9	0.4	0.1	2	1.4	1.2
1	1.6	0.4	1.5	1.7	0.9

(c)

Gambar 2. (a) Data target original Y, (b) Matrix Kombinasi C, (c) Data target dari hasil kombinasi . (a) Data target original Y, (b) Matrix Kombinasi C, (c) Data target dari hasil kombinasi $\hat{Y}_{|D|\times r} = Y_{|D|\times q} \times C_{q\times r}$

2.3. Augmentasi Data Target menggunakan PCA

Sedangkan metode PCA, perhitungan matriks C akan berubah menggunakan *singular value decomposition* (SVD) yang menghasilkan *principal components* (PCs) yang dapat digunakan untuk mencari data yang dengan variasi tertinggi. Dengan melakukan limitasi terhadap beberapa data yang memiliki variasi tertinggi, diharapkan data dapat diproyeksikan ke dimensi yang lebih rendah dengan langkah-langkah sebagai berikut:

1. Pilih ukuran dimensi data target yang baru r, yang berada pada rentang antara 2 hingga q
2. Hitung menggunakan SVD untuk mendapatkan Σ
3. Pilih subset dari Σ dengan ukuran $q \times r$ mulai dari PCs yang paling signifikan sebagai matrix C

2.4. Gradient Boosting Decision Tree Regressor

Untuk melakukan uji coba metode augmentasi data, algoritma regesi linear yang melakukan regresi linier terhadap masing-masing variable target (ST singkatan dari satu target) dengan menggunakan *decision tree regressor* yang dilatih menggunakan metode ensembel *gradient boosting*. Untuk mengakomodir seluruh variabel target, maka akan diperlukan sebanyak q buah ST. Parameter dan model regresi yang digunakan mengacu pada parameter yang didefinisikan dalam penelitian sebelumnya yang menjadi landasan dalam penelitian ini [6]. Parameter tersebut dijabarkan dalam **Error! Reference source not found.**. Selain parameter yang dijabarkan di dalam **Error! Reference source not found.**, akan menggunakan parameter *default* dari implementasi *library* scikit-learn [13].

Tabel 2. Parameter *Gradient Boosting Tree Regressor*

Parameter	Nilai
Jumlah maksimum *leaf node*	4
Jumlah decision tree regressor	100

2.5. Fungsi error

Untuk mengetahui performa dari masing-masing metode, fungsi *average Relative Root Mean Squared Error* (aRRMSE) yang telah didefinisikan didalam penelitian yang menjadi landasan penelitian ini oleh Tsoumakas dkk. [6] akan digunakan. Sebelum perhitungan aRRMSE dilakukan, proses inverse dari transformasi RLC dan PCA, serta *MinMaxScaler* akan dilakukan. aRRMSE sendiri dijabarkan dalam persamaan (9) dan (10).

$$RRMSE = \sqrt{\frac{\sum_{(x,y)\in D_{test}} \left(h(x)_j - y_j\right)^2}{\sum_{(x,y)\in D_{test}} \left(\bar{y}_j - y_j\right)^2}} \qquad (9)$$

$$aRRMSE(h, D_{test}) = \frac{1}{q}\sum_{j=1}^{q} RRMSE \qquad (10)$$

Dimana $h(\mathrm{x})$ adalah output dari regresi linier untuk seluruh data input x , y_j adalah target untuk masing-masing data input x_j , $\bar{y}_j$ nilai rata-rata untuk data target ke-j ,q adalah jumlah dimensi dari data target. Dalam fungsi aRRMSE, masing-masing output dari regresi linier akan dihitung nilai deviasinya terhadap data target dengan mengacu pada nilai rata-rata masing-masing data target. Hal ini bertujuan untuk menstandarisasi performa melalui rata-rata dari data target yang tidak homogen setelah proses *training* model ensembel [6].

2.6. Evaluasi

Hasil dari perhitungan nilai error dari keduabelas dataset akan dibandingkan dengan cara mencari nilai terendah untuk masing-masing metode per dataset untuk masing-masing metode augmentasi data. Metode dengan nilai error yang lebih rendah adalah metode yang dianggap lebih berhasil dalam melakukan regresi linier. Jumlahan keberhasilan dari masing-masing metode untuk semua dataset kemudian dibandingkan untuk memperoleh kesimpulan.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Uji Coba Metode Random Linear Target Combination

Hasil uji coba metode RLC dijabarkan dalam **Error! Not a valid bookmark self-reference.Error! Reference source not found.**. Nilai k pada **Error! Not a valid bookmark self-reference.** merupakan jumlahan data target yang dikorelasikan secara linier dan random. Pada metode RLC dimensi dari data target akan menjadi konstan yaitu 500 dimensi. Hasil uji coba ini akan menjadi tolok ukur dari uji coba metode PCA.

Tabel 3. Hasil Uji Coba RLC

k	atp1d	atp7d	edm	oes10	oes97	rf1	rf2	sf1978	sf1969	wq	scm1d	scm20d
2	0.4137	0.4078	**0.6703**	**0.5649**	**0.5247**	0.1710	0.1830	1.1424	**0.5188**	0.1021	0.1140	0.0998
3	0.4016	0.3794		0.5726	0.5255	0.1687	0.1802	**1.1268**	0.5377	0.0982	**0.1139**	0.0986
4	0.4032	0.3671		0.5789	0.5351	0.1656	0.1784			0.0938	0.1168	**0.0973**
5	0.3965	**0.3659**		0.5907	0.5284	0.1665	0.1790			**0.0934**	0.1171	0.0977
6	**0.3900**	0.3660		0.5822	0.5345	**0.1638**	0.1739			0.0956	0.1194	0.0977
7				0.5876	0.5408	0.1666	0.1741			0.0966	0.1199	0.0989
8				0.5961	0.5385	0.1654	**0.1738**			0.0950	0.1215	0.0986
9				0.5871	0.5415					0.1009	0.1226	0.0988
10				0.5876	0.5413					0.1006	0.1221	0.0986
11				0.5980	0.5413					0.1032		0.1007
12				0.6029	0.5414					0.1049		0.1000
13				0.5912	0.5404					0.1109		0.1003
14				0.5957	0.5401					0.1172		0.1011
15				0.5985	0.5438							0.1012
16				0.6038	0.5424							0.1011
MIN	0.3900	0.3659	0.6703	0.5649	0.5247	0.1638	0.1738	1.1268	0.5188	0.0934	0.1139	0.0973

Jurnal Teknologi Terpadu

https://journal.nurulfikri.ac.id/index.php/jtt
ISSN : 2477-0043 ISSN ONLINE : 2460-7908

Jika dibandingkan dengan hasil uji coba yang dilakukan oleh Tsoumakas dkk. [6] pada penelitian sebelumnya dengan library yang berbeda, dari perbandingan

4. HASIL DAN PEMBAHASAN

4.1. Hasil Uji Coba Metode Random Linear Target Combination

Hasil uji coba metode RLC dijabarkan dalam **Error! Not a valid bookmark self-reference.Error! Reference source not found.**. Nilai k pada **Error! Not a valid bookmark self-reference.** merupakan jumlahan data target yang dikorelasikan secara linier dan random. Pada metode RLC dimensi dari data target akan menjadi konstan yaitu 500 dimensi. Hasil uji coba ini akan menjadi tolok ukur dari uji coba metode PCA.

Tabel 3 dan **Tabel 4** dapat terlihat bahwa hasil uji coba yang dilakukan dalam penelitian ini secara konsisten memiliki nilai error yang lebih rendah dibandingkan dengan hasil uji coba. Oleh karena itu hasil dari implementasi di dalam penelitian ini akan digunakan sebagai dasar dalam menentukan kesimpulan akhir.

4.2. Hasil Uji Coba Metode Principal Component Analysis

Tabel 5Error! Reference source not found., masing-masing dataset memiliki jumlah uji coba yang berbeda karena perbedaan jumlah data target

Tabel 4. Hasil Uji Coba RLC dari Penelitian Sebelumnya oleh Tsoumakas dkk. [6]

k	atp1d	atp7d	edm	oes10	oes97	rf1	rf2	sf1978	sf1969	wq	scm1d	scm20d
2	0.3842	**0.4614**	**0.6996**	**0.5026**	0.5593	**0.7265**	**0.7036**	1.2312	1.5746	0.9100	**0.4572**	0.7469
3	**0.3840**	0.4653		0.5084	**0.5588**	0.7878	0.7584	**1.2172**	**1.5675**	**0.9080**	0.4610	**0.7467**
4	0.3884	0.4796		0.5232	0.5730	0.8204	0.7922			0.9085	0.4663	0.7472
5	0.3952	0.4917		0.5359	0.5837	0.8584	0.8327			0.9086	0.4699	0.7477
6	0.4022	0.5029		0.5472	0.5889	0.8515	0.8257			0.9089	0.4775	0.7490
7				0.5551	0.5958	0.8446	0.8106			0.9090	0.4820	0.7513
8				0.5734	0.6076	0.8868	0.8655			0.9107	0.4855	0.7536
9				0.5911	0.6153					0.9122	0.4889	0.7548
10				0.6031	0.6229					0.9128	0.4932	0.7537
11				0.6154	0.6348					0.9150	0.4978	0.7573
12				0.6285	0.6449					0.9163	0.5020	0.7571
13				0.6354	0.6590					0.9188	0.5057	0.7619
14				0.6428	0.6682					0.9217	0.5133	0.7640
15				0.6525	0.6860						0.5155	0.7681
16				0.6652	0.6916						0.5218	0.7704
MIN	0.3840	0.4614	0.6996	0.5026	0.5588	0.7268	0.7036	1.2172	0.15675	0.9080	0.4572	0.7467

seperti yang dijabarkan pada **Error! Reference source not found.**. Tergantung dari jumlah data targetnya (q), untuk masing-masing proyeksi ke dimensi yang lebih rendah (r) dari data target akan digunakan untuk melatih *decision tree regressor* dengan *gradient boosting*. Sebagai contoh, untuk dataset atp1d yang memiliki $q = 6$, akan dilakukan uji coba dengan memproyeksikan data target dengan dimensi $r \in \{2, 3, 4, 5, 6\}$.

Dari hasil pengujian tersebut akan diambil parameter r mana yang memiliki nilai error paling rendah untuk tiap-tiap dataset.

Tabel 5Error! Reference source not found., terlihat pola dari masing-masing dataset yang memiliki nilai error yang paling rendah, yaitu pada nilai r yang cenderung lebih kecil (berkisar antara 2 hingga 5), sehingga semakin besar dimensi yang dimiliki oleh data target dari masing-masing **Tabel 5**.

dataset, semakin besar pula jumlah reduksi dimensi yang terjadi untuk mendapatkan nilai error yang lebih rendah pada pelatihan model regresi. Hal ini dapat ditunjukkan pada

Diterima (tanggal), Direvisi (tanggal), Diterima untuk publikasi (tanggal)

Jurnal Teknologi Terpadu

https://journal.nurulfikri.ac.id/index.php/jtt
ISSN : 2477-0043 ISSN ONLINE : 2460-7908

Tabel 5. Hasil Uji Coba PCA

r	atp1d	atp7d	edm	oes10	oes97	rf1	rf2	sf1978	sf1969	wq	scm1d	scm20d
2	0.3921	**0.3755**	**0.8731**	**0.3849**	0.3934	0.1801	0.1895	0.5368	**0.3372**	**0.0804**	**0.1196**	**0.0870**
3	**0.3855**	0.4681		0.4521	**0.3842**	0.1807	0.1898	**0.5362**	0.3907	0.0836	0.1197	0.0872
4	0.3870	0.5049		0.4495	0.4018	0.1741	0.1905			0.0846	0.1198	0.0876
5	0.3890	0.5110		0.4509	0.4137	**0.1722**	**0.1878**			0.0855	0.1224	0.0916
6	0.3870	0.5115		0.4486	0.4460	0.1756	0.1900			0.0888	0.1267	0.0959
7				0.4489	0.4522	0.1759	0.1903			0.0929	0.1275	0.0983
8				0.4514	0.4514	0.1759	0.1903			0.0918	0.1356	0.1191
9				0.4442	0.4504					0.0947	0.1347	0.1165
10				0.4444	0.4499					0.0949	0.1360	0.1199
11				0.4449	0.4515					0.0996	0.1393	0.1223
12				0.4451	0.4500					0.0987	0.1397	0.1217
13				0.4449	0.4504					0.0988	0.1397	0.1232
14				0.4457	0.4512					0.0999	0.1434	0.1258
15				0.4455	0.4512						0.1457	0.1275
16				0.4461	0.4512						0.1483	0.1279
MIN	0.3855	0.3755	0.8731	0.3849	0.3842	0.1722	0.1878	0.5362	0.3372	0.0804	0.1196	0.0870

Pada **Tabel 6**, dataset diurutkan mulai dari dataset yang memiliki dimensi data target yang paling kecil hingga dataset yang memiliki dimensi data target yang paling besar. Dari prosentase reduksi dimensi yang dibutuhkan untuk memperoleh nilai error yang kecil.

Tabel 6. Prosentasi Reduksi Dimensi terhadap Nilai Error Minimum

dataset	edm	sf1969	sf1978	atp1d	atp7d	rf1	rf2	wq	scm1d	oes97	scm20d	oes10
q	2	2	3	3	2	5	5	2	2	3	2	2
r	2	3	3	6	6	8	8	14	16	16	16	16
Prosentase Reduksi Dimensi	0%	33%	0%	50%	67%	38%	38%	86%	88%	81%	88%	88%

4.3. RLC dan PCA dalam Regresi Multitarget

Hasil dari uji coba RLC dan PCA kemudian dibandingkan dengan cara membandingkan tiap-tiap dataset dengan hasil error yang didapatkan, hanya hasil error terkecil yang akan dibandingkan. Rangkuman hasil error tersebut dapat dijabarkan didalam **Error! Reference source not found.** dengan nilai rata-rata keseluruhan uji coba terhadap dua belas dataset disajikan pada kolom terakhir.

Dari hasil uji coba tersebut dapat dilihat bahwa metode PCA memiliki nilai error yang lebih rendah untuk dataset atp1d, oes10, oes97, sf1978, sf1969, wq, dan scm20d. Sehingga jika dihitung jumlah dataset dimana metode tersebut memiliki error yang lebih rendah, didapatlah perbandingan RLC : PCA = 5 : 7.

Tabel 7. Perbandingan aRRMSE metode RLC dan PCA

	atp1d	atp7d	edm	oes10	oes97	rf1	rf2	sf1978	sf1969	wq	scm1d	scm20d	AVG
RLC	0.3900	**0.3659**	**0.6703**	0.5649	0.5247	**0.1638**	**0.1738**	1.1268	0.5188	0.0934	**0.1139**	0.0973	**0.4003**
PCA	**0.3855**	0.3755	0.8731	**0.3849**	**0.3842**	0.1722	0.1878	**0.5362**	**0.3372**	**0.0804**	0.1196	**0.0870**	**0.3270**

Selain perbandingan untuk masing-masing dataset, metode RLC dan PCA memiliki hasil rata-rata error masing-masing untuk semua dataset yaitu 0.4003 dan 0.3270. Dari hasil perbandingan error kedua metode tersebut, dapat ditentukan bahwa PCA mampu menghasilkan nilai error yang lebih rendah untuk sebagian besar dari dataset yang diujicobakan pada penelitian ini.

Diterima (tanggal), Direvisi (tanggal), Diterima untuk publikasi (tanggal)

5. KESIMPULAN

Dalam kasus regresi multitarget kedua metode dapat menghasilkan error yang lebih rendah dari 0.8, hal ini menunjukkan simpangan dari hasil regresi tidak akan lebih dari 1 satuan. PCA dinilai lebih unggul karena memiliki nilai error lebih rendah untuk keseluruhan dataset, yang dapat dilihat dari rata-rata nilai error.

Metode PCA mentransformasikan data ke dimensi yang lebih rendah sembari memberikan nilai error yang lebih rendah dibandingkan dengan metod RLC. Hal ini menunjukkan, reduksi dimensi tetap dapat memberikan hasil regresi dengan nilai error yang lebih kecil, sembari mengurangi kompleksitas dari data. Semakin kecil kompleksitas dari data, semakin cepat pula proses perhitungan yang dibutuhkan untuk model regresi dalam melakukan proses *training*.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih sebesar-besarnya atas kesempatan yang diberikan oleh LPPM Universitas Katolik Soegijapranata yang telah mendukung secara finansial untuk dapat melakukan penelitian dan menghasilkan tulisan ini.

DAFTAR PUSTAKA

[1] X. Yan and X. Su, *Linear regression analysis: theory and computing*. Singapore ; Hackensack, NJ: World Scientific, 2009.

[2] J. N. Hussain, "High dimensional data challenges in estimating multiple linear regression," *Journal of Physics: Conference Series*, vol. 1591, no. 1, p. 12035, 2020, doi: 10.1088/1742-6596/1591/1/012035.

[3] M. Bataineh and T. Marler, "Neural network for regression problems with reduced training sets," *Neural Networks*, vol. 95, pp. 1–9, 2017, doi: https://doi.org/10.1016/j.neunet.2017.07.018.

[4] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A Comprehensive Analysis of Deep Regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2065–2081, 2020, doi: 10.1109/TPAMI.2019.2910523.

[5] D. Rügamer *et al.*, "deepregression: a Flexible Neural Network Framework for Semi-Structured Deep Distributional Regression," *arXiv:2104.02705 [cs, stat]*, 2021, [Online]. Available: http://arxiv.org/abs/2104.02705

[6] G. Tsoumakas, E. Spyromitros-Xioufis, A. Vrekou, and I. Vlahavas, "Multi-Target Regression via Random Linear Target Combinations," *arXiv:1404.5065 [cs]*, vol. 8726, pp. 225–240, 2014, doi: 10.1007/978-3-662-44845-8_15.

[7] Q. Zhao, E. Adeli, N. Honnorat, T. Leng, and K. Pohl, "Variational AutoEncoder For Regression: Application to Brain Aging Analysis," vol. 11765, 2019, pp. 823–831. [Online]. Available: https://www.researchgate.net/publication/336393829_Variational_AutoEncoder_For_Regression_Application_to_Brain_Aging_Analysis

[8] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *WIREs Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015, doi: 10.1002/widm.1157.

[9] P. Boye, D. Mireku-Gyimah, and C. A. Okpoti, "Multiple Linear Regression Model for Estimating the Price of a Housing Unit," *Ghana Mining Journal*, vol. 17, no. 2, pp. 66–77, 2017, doi: 10.4314/gm.v17i2.9.

[10] N. Herawati, K. Nisa, E. Setiawan, N. Nusyirwan, and T. Tiryono, "Regularized multiple regression methods to deal with severe multicollinearity," *International Journal of Statistics and Applications*, vol. 8, no. 4, pp. 167–172, 2018.

[11] E. Spyromitros-Xioufis, W. Groves, G. Tsoumakas, and I. Vlahavas, *Multi-Label Classification Methods for Multi-Target Regression*. 2012. [Online]. Available: https://www.researchgate.net/publication/233780193_Multi-Label_Classification_Methods_for_Multi-Target_Regression

[12] V. Pestov, "Is the k-NN classifier in high dimensions affected by the curse of dimensionality?," *Computers & Mathematics with Applications*, vol. 65, no. 10, pp. 1427–1437, 2013, doi: 10.1016/j.camwa.2012.09.011.

[13] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] A. Karalic, "First Order Regression," *Machine Learning*, vol. 26, pp. 147–176, 1997, doi: 10.1023/A:1007365207130.

[15] A. Asuncion and D. Newman, "UCI: Machine Learning Repository : Solar Flare Dataset."

[16] S. Džeroski, D. Demsar, and J. Grbović, "Predicting Chemical Parameters of River Water Quality from Bioindicator Data," *Applied Intelligence*, vol. 13, pp. 7–17, 2000, doi: 10.1023/A:1008323212047.

Kombinasi Linier Target Data Untuk Regresi Multitarget Menggunakan Principal Component Analysis

ORIGINALITY REPORT

17%	16%	9%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to IAIN Pekalongan Student Paper	3%
2	arxiv.org Internet Source	1%
3	hal.archives-ouvertes.fr Internet Source	1%
4	www.airitilibrary.com Internet Source	1%
5	thesai.org Internet Source	1%
6	ikee.lib.auth.gr Internet Source	1%
7	article.sapub.org Internet Source	1%
8	fugumt.com Internet Source	1%
9	umat.edu.gh Internet Source	1%

10	backend.orbit.dtu.dk Internet Source	1 %
11	Hao Liu, Xin Shen, Lin Cao, Ting Yun, Zhengnan Zhang, Xiaoyao Fu, Xinxin Chen, Fangzhou Liu. "Deep learning in forest structural parameters estimation using airborne LiDAR data", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020 Publication	1 %
12	repository.its.ac.id Internet Source	<1 %
13	Jiabei Zeng, Yang Liu, Biao Leng, Zhang Xiong, Yiu-ming Cheung. "Dimensionality Reduction in Multiple Ordinal Regression", IEEE Transactions on Neural Networks and Learning Systems, 2017 Publication	<1 %
14	Saulo Martiello Mastelini, Everton Jose Santana, Victor Guilherme Turrisi da Costa, Sylvio Barbon. "Benchmarking Multi-target Regression Methods", 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), 2018 Publication	<1 %
15	Submitted to University of York Student Paper	<1 %

16	www.slds.stat.uni-muenchen.de Internet Source	<1 %
17	www.scribd.com Internet Source	<1 %
18	cercetare.ulbsibiu.ro Internet Source	<1 %
19	jtiik.ub.ac.id Internet Source	<1 %
20	Novianti Puspitasari, Andi Tejawati, Friendy Prakoso. "Estimasi Stok Penerimaan Bahan Bakar Minyak Menggunakan Metode Fuzzy Tsukamoto", JRST (Jurnal Riset Sains dan Teknologi), 2019 Publication	<1 %
21	Submitted to Universitas Brawijaya Student Paper	<1 %
22	adoc.pub Internet Source	<1 %
23	docplayer.info Internet Source	<1 %
24	dspace.cc.tut.fi Internet Source	<1 %
25	www.whanganui.govt.nz Internet Source	<1 %

26	ouci.dntb.gov.ua Internet Source	<1 %
27	www.waset.org Internet Source	<1 %
28	coek.info Internet Source	<1 %
29	etd.repository.ugm.ac.id Internet Source	<1 %
30	intanmeiwell.blogspot.com Internet Source	<1 %
31	text-id.123dok.com Internet Source	<1 %
32	www.ilo.org Internet Source	<1 %
33	123dok.com Internet Source	<1 %
34	Repository.umy.ac.id Internet Source	<1 %
35	ejurnal.its.ac.id Internet Source	<1 %
36	id.scribd.com Internet Source	<1 %
37	mycourseville-default.s3.ap-southeast-1.amazonaws.com	<1 %

	Internet Source	
38	ppj.uniska-bjm.ac.id Internet Source	<1%
39	spbeptagtk.blogspot.com Internet Source	<1%
40	www.arxiv-vanity.com Internet Source	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		