

Analisis 9 Saham Sektor Industri di Indonesia Menggunakan Metode SVR

Nur Adhi Nugroho^{1,a)}, Acep Purqon^{1,b)}

¹Laboratorium Fisika Bumi,
Kelompok Keahlian Fisika Bumi dan Sistem Kompleks,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung,
Jl. Ganesha no. 10 Bandung, Indonesia, 40132

^{a)} nur.adhi21@gmail.com

^{b)} acep@fi.itb.ac.id

Abstrak

Saham merupakan salah satu instrumen dalam pasar keuangan. Karakteristik dari data indeks harga saham yaitu sifatnya yang berubah dari waktu ke waktu. Oleh karena itu, untuk meminimumkan resiko yang dihadapi para investor maka perlu adanya sistem yang mampu memprediksi trend harga saham. Dalam makalah ini, akan dibahas Support Vector Regression (SVR) untuk menentukan prediksi closing harga saham. SVR merupakan pengembangan dari SVM untuk kasus regresi. Tujuan SVR adalah membuat suatu fungsi regresi untuk menentukan hyperplane terbaik. Tahap yang juga penting sebelum implementasi ke dalam SVR yaitu data mining dan data processing. Pengambilan data saham melalui yahoo finance kemudian pemrosesan data dilakukan menggunakan Excel. Kedua tahap ini diperlukan untuk menghindari data saham perusahaan yang kosong dan menghapus data waktu yang tidak diperlukan. Saham yang digunakan yaitu dari 9 sektor industri di Indonesia. Data indeks saham periode 2005-2014 akan digunakan sebagai data set training. Kemudian performansi tingkat keakuratan SVR akan dibandingkan dengan data periode 2015.

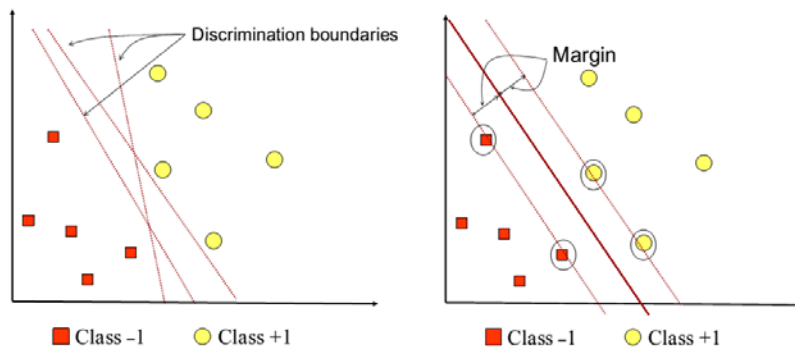
Kata-kata kunci: Data Mining, Data Processing, Regresi, Saham, Support Vector Regression (SVR)

PENDAHULUAN

Pasar modal, menurut Bursa Efek Indonesia, merupakan pasar unuk berbagai instrument keuangan jangka panjang yang bisa diperjualbelikan, baik surat utang (obligasi), ekuiti (saham), reksa dana, instrumen derivatif maupun instrumen lainnya. Adanya pasar modal memberikan kesempatan masyarakat untuk bisa berinvestasi, salah satunya yaitu saham. Indeks harga saham memberikan gambaran kondisi ekonomi saat itu serta menjadi indikator bagi para investor dalam mengambil keputusan untuk berinvestasi sehingga tidak mengalami kerugian yang besar.

Fenomena indeks harga saham merupakan salah satu peristiwa dinamika stokastik. Dinamika stokastik merupakan pergerakan dari sebuah data di mana data tersebut menunjukkan adanya efek acak. Datanya yang bersifat *time series* dan nilainya yang fluktuatif menyebabkan prediksi harga saham merupakan salah satu prediksi yang sulit untuk dilakukan. Hal ini disebabkan kondisi pasar keuangan dipengaruhi oleh banyak faktor, seperti kebijakan perusahaan, inflasi, nilai suku bunga, kebijakan ekonomi, dan gejolak politik di negara tersebut.

Support Vector Machine merupakan salah satu algoritma yang menggunakan pendekatan non linear. Konsep dasar SVM adalah mencari *hyperplane* (garis pemisah) baik untuk memisahkan antar *class*.



Gambar 1. *Hyperplane*^[1]

Pada **Gambar 1.** menunjukkan adanya persebaran data dari dua *class*. *Class* (-1) disimbolkan dengan kotak warna merah, sedangkan (+1) disimbolkan dengan lingkaran warna kuning. Gambar sebelah kiri memperlihatkan banyaknya alternatif *hyperplane* untuk memisahkan dua *class*. Sedangkan pada gambar sebelah kanan, garis merah tebal menunjukkan *hyperplane* terbaik yang memisahkan kedua *class* tepat ditengah. *Hyperplane* terbaik dapat ditemukan dengan mengukur margin dari *hyperplane* tersebut dengan mencari titik maksimalnya. Margin merupakan jarak antara *hyperplane* dengan titik terdekat dari masing-masing *class*. Titik yang terdekat ini disebut *support vector*.

Dalam makalah ini akan menggunakan *Support Vector Regression* (SVR) untuk memprediksi *trend* harga saham. SVR sendiri adalah pengembangan dari SVM untuk kasus regresi. Perbedaan dengan SVM adalah SVM digunakan untuk kasus klasifikasi sehingga output berupa diskrit. Sedangkan SVR, karena untuk kasus regresi, output bersifat kontinu (bilangan riil).

METODE SUPPORT VECTOR REGRESSION

Support Vector Regression

Misalkan kita mempunyai data *training* $\{(x_1, y_1) \dots (x_n, y_n)\} \subseteq \mathbb{R}$. Tujuan SVR adalah menemukan suatu fungsi $f(x)$ yang mempunyai deviasi ϵ paling besar (maksimal) dari target yang diobservasi y_i untuk semua data *training*. Sehingga ketika $\epsilon = 0$ maka akan dihasilkan fungsi regresi yang sempurna.

Bentuk umum fungsi linear:

$$f(\vec{x}) = \vec{w}^T \varphi(\vec{x}) + b \tag{1}$$

\vec{w} merupakan vektor pembobot, $\varphi(\vec{x})$ merupakan fungsi pemetaan di dalam *feature space*, dan b adalah konstanta. Untuk mendapatkan fungsi setipis (sedatar) mungkin maka harus meminimumkan norm dari \vec{w} (panjang vektor dari \vec{w}).

$$\min \frac{1}{2} \|\vec{w}\|^2 \tag{2}$$

dan memenuhi:

$$y_i - \vec{w}^T \varphi(\vec{x}) + b \leq \epsilon \tag{3}$$

$$\vec{w}^T \varphi(\vec{x}) - y_i + b \leq \epsilon \tag{4}$$

Pada kenyataannya tidak semua kasus datanya masuk pada range $f(\vec{x}) \pm \epsilon$. Sehingga perlu ada tambahan variabel baru yang disebut variabel *slack* (ξ, ξ^*) untuk mengatasi data yang berada diluar batas margin. Data yang berada diluar margin akan terkena pinalti. Problem optimasi di atas dapat diperoleh dengan meminimumkan:

$$\min \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^l (\xi + \xi^*) \tag{5}$$

di mana

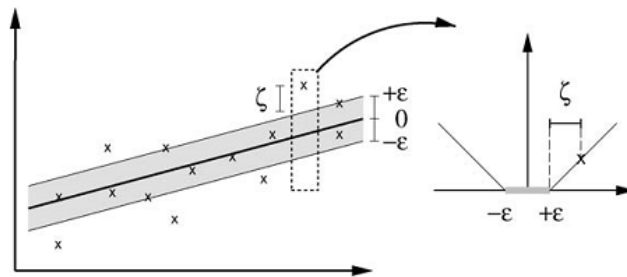
$$y_i - \vec{w}^T \varphi(\vec{x}) + b \leq \varepsilon + \xi \tag{6}$$

$$\vec{w}^T \varphi(\vec{x}) - y_i + b \leq \varepsilon + \xi^* \tag{7}$$

$$\xi, \xi^* \geq 0 \tag{8}$$

Konstanta C merupakan nilai pinalti dari data karena *error* lebih dari ε . Sedangkan data yang masih dalam jangkauan ε maka dianggap *error* bernilai nol. Korespondensi ini yang dinamakan dengan ε -insensitive loss function ($|\xi|_\varepsilon$), dapat dituliskan sebagai berikut:

$$|\xi|_\varepsilon := \begin{cases} 0, & \text{jika } |\xi| \leq \varepsilon \\ |\xi| - \varepsilon, & \text{selainnya} \end{cases} \tag{9}$$



Gambar 2. ε -insensitive loss function^[2]

Pada SVR, *support vector* merupakan data training yang terletak pada margin (ε) dan diluar margin. Sehingga jumlah *support vector* akan menurun karena naiknya nilai ε .

Solusi optimasi persamaan (5) dan batasan persamaan (6), (7), (8) diselesaikan dengan fungsi *Lagrange*:

$$Q(\vec{w}, b, \alpha_i, \alpha_i^*, \eta_i, \eta_i^*) = L$$

$$L = \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) - \sum_{i=1}^l \alpha_i (\varepsilon + \xi_i + f(\vec{x}_i) - y_i) - \sum_{i=1}^l \alpha_i^* (\varepsilon + \xi_i^* - f(\vec{x}_i) + y_i) - (\eta_i \xi_i + \eta_i^* \xi_i^*) \tag{10}$$

Di mana $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ adalah *Lagrange Multiplier* dan nilainya lebih dari samadengan nol. Sedangkan $f(x_i)$ merupakan persamaan (1). Untuk mendapatkan solusi optimal di atas dilakukan turunan parsial *Lagrange Multiplier* terhadap \vec{w}, b, ξ, ξ^* :

$$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \geq 0$$

$$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum_{i=1}^l (\alpha_i - \alpha_i^*) \varphi(\vec{x}_i) = 0 \tag{11}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^l (\alpha_i - \alpha_i^*) = 0 \tag{12}$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \eta_i = 0 \tag{13}$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \alpha_i^* - \eta_i^* \tag{14}$$

Dari persamaan (11) maka didapatkan:

$$\vec{w} = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \varphi(\vec{x}_i) \tag{15}$$

Kemudian substitusi persamaan (11), (12), (13), (14) ke dalam persamaan (10) dengan memaksimalkan solusi dual $Q(\alpha_i, \alpha_i^*)$:

$$f(\vec{x}) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \vec{w}^T(\vec{x}_i) \varphi(\vec{x}) + b \tag{16}$$

$$f(\vec{x}) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) k(\vec{x}_i, \vec{x}) + b \tag{17}$$

$K(\vec{x}_i, \vec{x})$ merupakan fungsi kernel. Penjelasan fungsi kernel akan dijelaskan pada subbagian berikutnya. Kemudian mencari solusi optimal b yaitu dengan menggunakan kondisi KKT (*Karush-Kuhn-Tucker*) didapatkan sebagai berikut:

$$b = y_i - \vec{w}^T \varphi(\vec{x}_i) - \varepsilon = y_i - \sum_{i=1}^l (\alpha_i - \alpha_i^*) k(\vec{x}_i, \vec{x}) - \varepsilon \tag{18}$$

Fungsi Kernel

Peran dari fungsi kernel adalah memetakan data x di *input space* ke dalam *feature space* yang berdimensi lebih tinggi.

Fungsi kernel yang biasanya digunakan (berdasarkan studi literature) adalah:

1. Kernel linear

$$\varphi(\vec{x}) = K(\vec{x}, \vec{x}_i) = \vec{x}^T \vec{x} \tag{19}$$

2. Kernel polynomial

$$\varphi(\vec{x}) = K(\vec{x}, \vec{x}_i) = (\vec{x}^T \vec{x} + 1)^d \tag{20}$$

3. Kernel RBF

$$\varphi(x) = K(x, x_i) = \exp\left(\frac{-\|x - x_i\|^2}{2\sigma^2}\right) \tag{21}$$

Kinerja Prediksi

Salah satu kriteria untuk mengukur keakuratan suatu model adalah dengan melihat nilai *error* yang dihasilkan. Dalam makalah ini nilai *error* yang digunakan yaitu *Mean Absolute Percentage Error* (MAPE). Dirumuskan sebagai berikut:

$$MAPE = \frac{1}{n} \sum_{i=1}^l \left| \frac{y_i - \bar{y}_i}{y_i} \right| \times 100 \tag{22}$$

Di mana y_i merupakan nilai aktual dan \bar{y}_i merupakan nilai prediksi.

HASIL DAN PERBANDINGAN DENGAN DATA REFERENSI

Data yang digunakan adalah data harga penutupan dari 9 perusahaan pada sektor tiap industri di Indonesia.

Tabel 1. Daftar Perusahaan dari Tiap Sektor Industri

No.	Sektor Industri	Perusahaan
1	Pertanian	Astra Argo Lestari (AALI)
2	Pertambangan	Aneka Tambang (ANTM)
3	Industri Dasar dan Kimia	Semen Indonesia (SMGR)
4	Aneka Industri	Astra International (ASII)
5	Industri Barang Konsumsi	Tiga Pilar Sejahtera Food (AISA)
6	Properti dan Real Estate	Lippo Cikarang (LPCK)
7	Transportasi dan Infrastruktur	Indosat (ISAT)
8	Keuangan	Bank Negara Indonesia (BBNI)
9	Perdagangan	AKR Corpindo (AKRA)

Data yang digunakan yaitu *data training* dari 1 Januari 2005 sampai 31 Desember 2014. Sedangkan *data testing* sebagai komparasi dengan hasil prediksi. Pengolahan data akan menggunakan library SMOreg dengan jumlah *step* sebanyak 6. Tiap *step* merepresentasikan periodisitas dalam bulanan. Sedangkan jumlah *step* menyatakan jumlah prediksi. Sehingga prediksi yang dilakukan adalah untuk 6 periode kedepan.

Tabel 2. Tingkat Akurasi Model

Perusahaan	SVR (RBF) C=1.0, gamma=1.0					
	Mean Absolute Percentage Error (MAPE)					
	1-step ahead	2-step ahead	3-step ahead	4-step ahead	5-step ahead	6-step ahead
AALI	0.2603	0.2941	0.3209	0.345	0.3669	0.3874
ANTM	1.0629	2.9755	3.6434	4.2501	4.8017	5.3093
SMGR	1.9468	2.961	3.7928	4.532	5.2311	5.8999
ASII	2.0592	3.1807	4.0777	4.8587	5.5354	6.1633
AISA	3.5441	4.4261	4.9451	5.3344	5.6889	6.0136
LPCK	4.4496	5.6265	6.5814	7.2714	8.0177	8.6398
ISAT	1.5715	2.2691	2.7825	3.1495	3.4743	3.7289
BBNI	1.6772	2.3973	2.9374	3.4095	3.7887	4.1596
AKRA	1.82	2.9115	3.8218	4.629	5.2022	5.7473

Tabel 3. Hasil Prediksi 6 Periode Kedepan

Perusahaan	SVR (RBF) C=1.0, gamma=1.0					
	Hasil Prediksi					
	1-step ahead	2-step ahead	3-step ahead	4-step ahead	5-step ahead	6-step ahead
AALI	23664.94	23687.92	23729.87	23736.43	23761.63	23793.87
ANTM	1057.24	1044.17	1042.27	1039.64	1039.99	1043.69
SMGR	16192.99	16146.03	16139.13	16129.32	16120.22	16113.78
ASII	7389.92	7360.77	7344.61	7339.67	7333.10	7328
AISA	2099.38	2096.76	2099.96	2098.42	2100.43	2098.89
LPCK	10275.5	10224.1	10176.85	10167.72	10148.2	10138.76
ISAT	4034.1	4018.71	4010.72	4003.72	3995.4292	3990.48
BBNI	6062.9	6058.23	6065.40	6077.89	6100.17	6110.65
AKRA	4182.30	4236.96	4302.07	4351.16	4394.60	4433.19

Tabel 4. Data Aktual 6 Periode Kedepan

Perusahaan	Data Aktual					
	Periode 1	Periode 2	Periode 3	Periode 4	Periode 5	Periode 6
AALI	23250	24650	24300	20350	24800	22950
ANTM	894.575	844.177	726.58	667.782	638.382	571.184
SMGR	14575	14875	13650	12500	13450	12000

ASII	7850	7850	8575	6850	7300	7075
AISA	2150	2200	2095	1760	1825	1870
LPCK	11500	11925	11675	11975	10850	9100
ISAT	4095	4125	4265	4000	3740	4000
BBNI	6250	6875	7225	6425	6875	5300
AKRA	4695	4870	5125	5200	5475	5925

Performansi (tingkat keakurasian) semakin baik apabila nilai MAPE semakin kecil. Dari tabel 2. menunjukkan nilai MAPE yang cukup besar. Hanya pada saham Astra Argo Lestari yang memiliki persentase MAPE kecil. Sehingga ini belum bisa dijadikan pedoman dalam tingkat kepercayaan investor terhadap nilai prediksi saham dari perusahaan tersebut.

Selain baik tidaknya data, keakurasian juga dipengaruhi pada fungsi kernel yang digunakan serta penentuan nilai parameter yang digunakan.

KESIMPULAN

SVR merupakan metode yang dapat mengatasi *overfitting*.^[2,3] Pemilihan fungsi kernel sangat penting. Fungsi kernel ini akan menentukan ruang fitur yang digunakan dalam mencari *hyperplane terbaik*. Ini dibuktikan dengan hasil persentase MAPE yang masih berkisar 5% menunjukkan model prediksi belum akurat. Hal yang perlu dilakukan kedepannya adalah melakukan *cross-check* terhadap penggunaan fungsi kernel serta uji coba dengan nilai parameter yang bervariasi.

REFERENSI

1. A.S. Nugroho, *Support Vector Machine – Teori Aplikasinya dalam Bioinformatika*. 2003.
2. Smola A.J., Scholkopf B., *A Tutorial on Support Vector Regression*. Statistics and Computing Volume 14, pp 199-222 (2004).
3. Basak D., Pal S., Patranabis D.C., *Support Vector Regression*. Neural Informatics Processing-Letter and Reviews Vol. 11 (2007).
4. Welling Max, *Support Vector Regression*. Department of Computer Science, University of Toronto.
5. Scholkopf B., Smola A.J., *Learning with Kernels*. MIT Press, Cambridge (2001).
6. Ho Chia-Hua, Lin Chih-Jen, *Large-scale Linear Support Vector Regression*. Journal of Machine Learning Research 13 (2012).
7. Kuhn Max, Johnson Kjell, *Applied Predictive Modelling*. Springer Science+Business Media New York (2013).
8. Breiman Leo, *Statistical Modeling: the Two Cultures*. Statistical Science Vol. 16 No.3, 199-215 (2001).
9. JinXing Che, *Support Vector Regression Based on Optimal Training Subset and Adaptive Particle Swarm Optimization Algorithm*. Applied Soft Computing 13 (2013).
10. Nghi Dang Huu, Mai Luong Chi, *Training Data Selection for Support Vector Machines Model*. 2011 INTERNATIONAL CONFERENCE ON INFORMATION AND ELETRONICS ENGINEERING, Singapore (2011).
11. Wu Chun-Hsin, *Traver-Time Prediction with Support Vector Regression*. IEEE Transactions on Intelligent Transportation Systems Vol.5 No.4 (2004).