

Web Mining

SELEKSI FITUR (Studi Kasus Klasifikasi Dokumen)

Prodi Teknik Informatika

Universitas Trunojoyo Madura

2024

LATAR BELAKANG

- **Dalam membangun suatu model (prediksi atau klassifikasi)**
 - Meningkatkan performansi model
 - Mendapatkan representasi dari sebuah model
 - Keterbasan sumber daya komputasi
 - Banyaknya fitur/variabel dibandingkan dengan sampel/instance
 - menyederhanakan model sehingga lebih mudah untuk diinterpretasikan oleh peneliti atau pengguna
 - Meningkatkan generalisasi model dengan cara mengurangi *overfitting*

Text	Label
<p>Manager Kajian Kebijakan Eksekutif Nasional Walhi, Boy Jerry Even Sembiring menilai, pemindahan ibu kota ke Penajaman Paser Utara, Kalimantan Timur, tetap akan diikuti dengan beban ekologis Jakarta dan Pulau Jawa.</p> <p>"Pemindahan ibu kota hanya akan memindahkan beban ekologis Jakarta dan Pulau Jawa ke Kalimantan Timur dan lokasi sekitarnya," ujar Boy dalam diskusi</p>	Negatif
<p>Menteri Dalam Negeri Tito Karnavian meyakini pemindahan ibu kota negara ke Kalimantan akan mendongkrak perekonomian Kalimantan secara keseluruhan.</p>	Positif
<p>rencana pemindahan ibu kota yang diputuskan pemerintah sudah tepat. Menurut dia Jakarta memang sudah waktunya diberi ruang untuk bernapas.</p> <p>"Jakarta saat ini sudah terlalu rumit, baik populasinya, polusinya, tingkat kejahatannya, ketimpangan sosialnya, dan seterusnya," kata perempuan yang akrab disapa Mala itu, Selasa</p>	Positif
. . . dst	

Klasifikasi dokumen

Dokumen	Term_1 (X1)	Term_2	...	Term_n-1	Term_n	Label (Y)
1	0.6119	0.6507	0.3215	0.7213	0.3412	Negatif
2	0.6685	0.6178	0.5128	0.2393	0.4715	Positif
3	0.4256	0.5908	0.8071	0.6702	0.4431	Positif
4	0.7273	0.7170	0.2909	0.3294	0.6603	Negatif
5	0.6754	0.5982	0.7754	0.7236	0.8239	Positif
6	0.4012	0.5566	0.1866	0.1114	0.3246	Positif
.... ds...						

Jumlah
dokumen = k

misal
Banyak term = n

$n > k$

Persoalan dalam pembelajaran mesin (contoh)

- **Klasifikasi dokumen**
 - Dokumen dinyatakan dengan **vector space model** dengan dimensi data adalah kosa kata yang ada dalam dokumen
 - Semakin banyak kosakata yang pada dokumen -> semakin tinggi dimensi variabel /fitur (jumlahnya sangat banyak atau melebihi banyaknya dokumen)

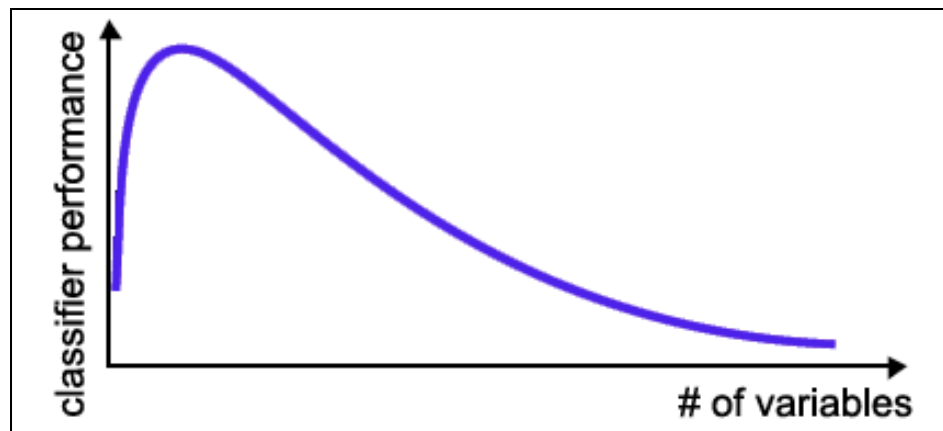
Misal:

Vocabulary ~ 15.000 kata (sehingga setiap dokumen dinyatakan dalam dengan 15.000-dimensi vektor

- **Seleksi Gen dari data microarray**
 - Variabel: koefisien ekspresi gene
 - Tujuan :untuk apat membedakan pasien sehat dari pasien penyakit kanker

PERLU.....

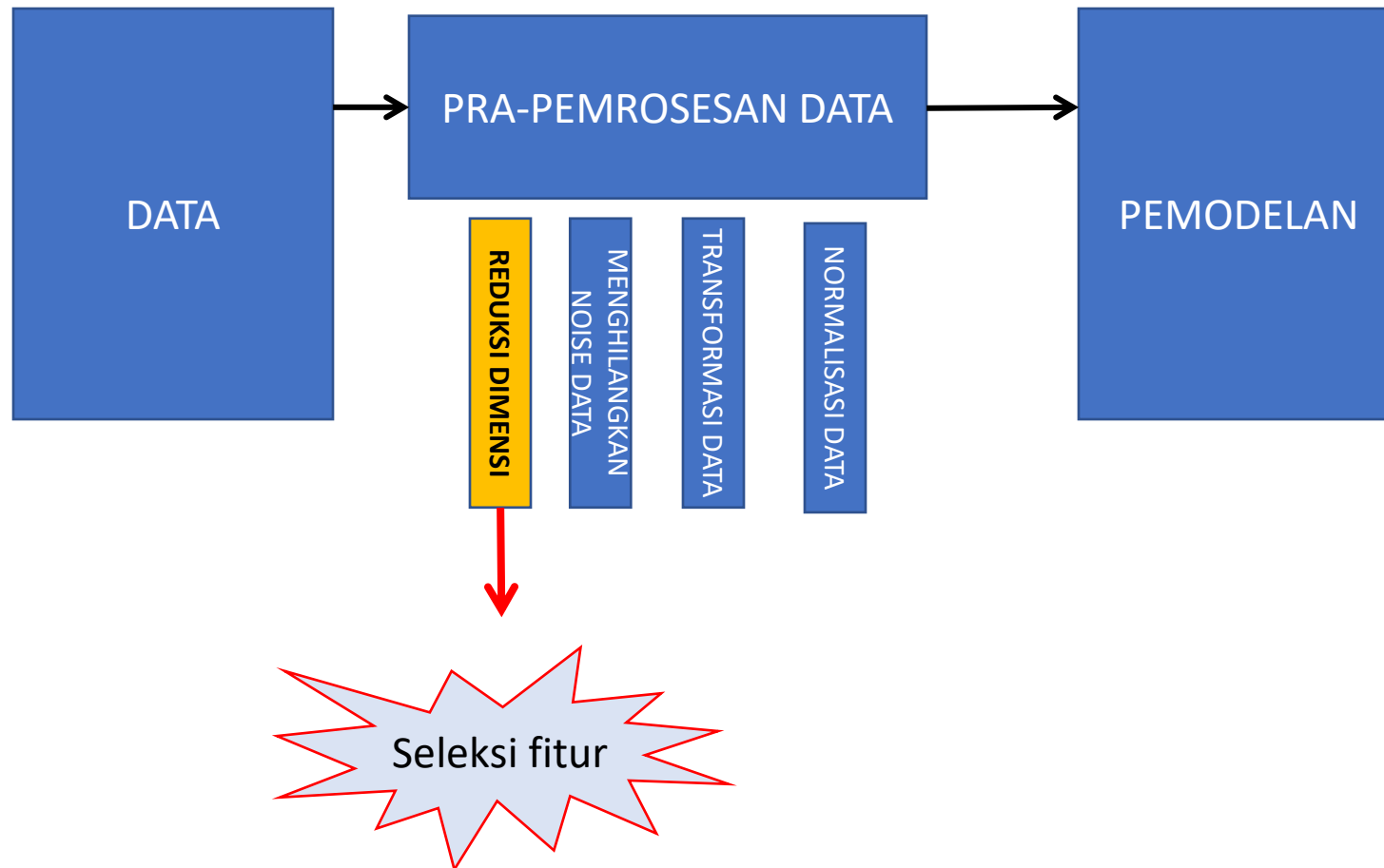
- Ketika jumlah variabel/fitur sangat banyak (dimensi data tinggi) maka perlu reduksi dimensi ----->**SELEKSI FITUR**
- Seleksi fitur sangat signifikan untuk meningkatkan *performance* algoritma pembelajaran
- Memilih fitur-fitur yang relevan dan menghilangkan Redundancy suatu kumpulan fitur



BUTUH

Seleksi fitur

Kerangka *Machine Learning*



Definsi masalah

- Klasifikasi (Supervised Learning):
Diberikan data pelatihan (**training data**)

$$L = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_m, y_m)\} \in X \times Y$$

Clasifier harus menemukan **hypothesis**

$$h \in H : X \rightarrow Y$$

Yang digunakan untuk memberi y untuk data baru x .

Fitur/Variabel

- Data X terdiri dari n fitur :

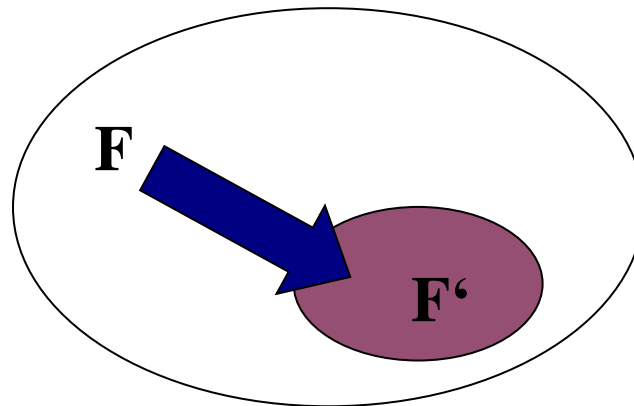
$$X = f_1 \times \dots \times f_i \times \dots \times f_n$$

Seleksi fitur

- Dari fitur yang ada $F = \{f_1, \dots, f_i, \dots, f_n\}$
Seleksi fitur adalah mencari $F' \subseteq F$
yang memaksimalkan hasil klasifikasi

Seleksi fitur

- Seleksi fitur:



$$\{f_1, \dots, f_i, \dots, f_n\} \xrightarrow{f.selection} \{f_{i_1}, \dots, f_{i_j}, \dots, f_{i_m}\}$$

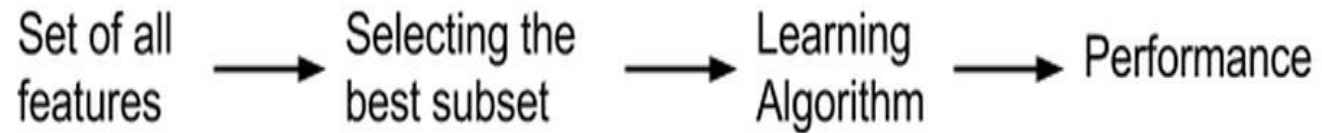
$$i_j \in \{1, \dots, n\}; j = 1, \dots, m$$

$$i_a = i_b \Rightarrow a = b; a, b \in \{1, \dots, m\}$$

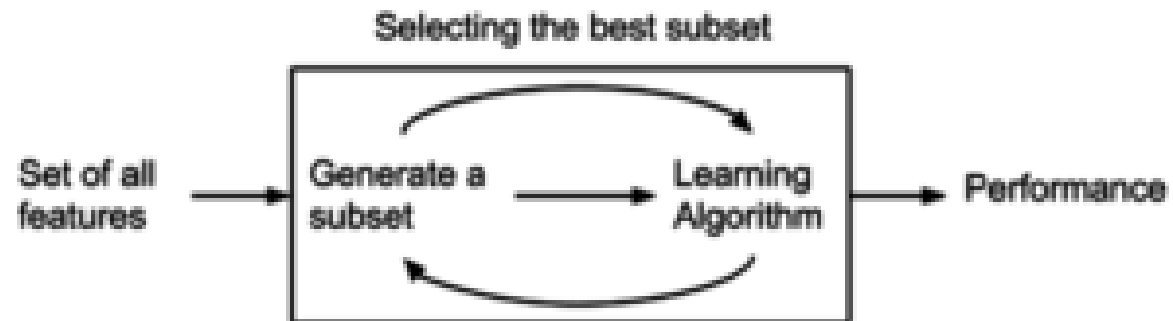
Seleksi Fitur

- Metode Filter
 - Metode ini mengevaluasi setiap fitur secara independen dari algoritma machine learning. Contoh teknik yang digunakan adalah korelasi Pearson, Chi-Square, dan ANOVA
- Metode Wrapper
 - Metode ini menggunakan algoritma machine learning untuk mengevaluasi kombinasi fitur dan memilih yang terbaik berdasarkan kinerja model. Contoh tekniknya adalah forward selection, backward elimination, dan recursive feature elimination
- Metode Embedded
 - Metode ini mengintegrasikan proses seleksi fitur ke dalam pelatihan model. Contoh tekniknya adalah Lasso Regression dan Decision Trees

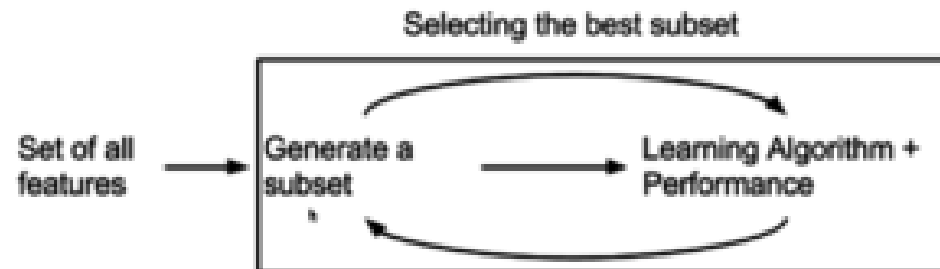
Seleksi Fitur



Metode Filter



Metode Wrapper



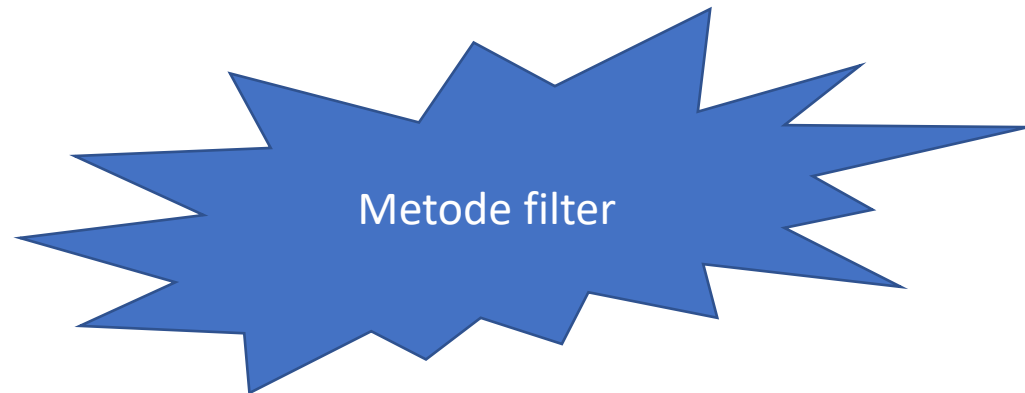
Metode Embedded

Seleksi Fitur

- memilih kualitas fitur atau fitur yang optimal didasarkan pada
 - **ukuran informasi (information measures)** ,
 - ukuran jarak (distance measures),
 - **ukuran kebergantungan (dependence measures),**
 - ukuran kekonsistenan (consistency measures).
- Metode filter dilakukan terpisah dari model pembelajaran proses pemilihan fitur dilakukan sebelum tahapan pembelajaran model.
- Metode ini juga disebut dengan metode pemeringkatan fitur. Setelah melakukan pemeringkatan nilai ukuran, dipilih fitur berdasarkan batas ambang yang ditentukan,

Merangking fitur/variable ranking

- Dari fitur F
Variable Ranking adalah proses mengurutkan nilai dari fungsi skore (yang mengukur relevansi fitur) $S:F \rightarrow \mathbb{R}$
- Skor tinggi menyatakan fitur yang relevan.



Metode filter

- Jenis teknik seleksi fitur, diantaranya adalah
 - **Mutual/Information Gain** (Kullback–Leibler divergence),
 - *Chi-Squared*,
 - ANOVA (*Analysis of Variance*), dll.

Information Gain

- *Information Gain* dalam machine learning digunakan untuk mengukur seberapa relevan / berpengaruh sebuah feature terhadap hasil pengukuran.
- Pengukuran didasarkan pada Entropy sebelum dan sesudah pemisahan.
- *Information Gain (IG)* dikenal juga dengan sebutan ***Mutual Information (MI)*** dalam kasus untuk mengetahui dependency antara dua variable (x,y).

Procedure Information Gain data mining feature selection Algorithm (IGFS)

```
var1  sf1 /* store selected feature. Initially empty*/
var2  Th  /* hold threshold value*/
var3  f(i) /* contains the ith feature of the data set */
1: begin
2: IGFS.build ();
3: begin
4:   sf1 = {}; /* create array of var1*/
5:   for loop i=1 to int of features
6:     INF=compute (IG) for the feature /* store computed features*/
7:     Gain (i) =INF /* compute Gain (i) */
8:   end for
9:   Th= threshold value /* hold threshold value*/
10:  For i= 1 to number of features
11:    If gain (i) > Th then
12:      Sfl=sf1+f {i} /* store and compute every feature in dataset*/
13:    end if
14:  end for
15: end
```

Sumber:

https://www.researchgate.net/publication/340099203_A_Composite_Hybrid_Feature_Selection_Learning-Ba

MUTUAL INFORMATION

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x, y)}{P(x) P(y)} \right)$$

dimana X dan Y adalah variabel

$P(x)$, $P(y)$, dan $P(x, y)$ adalah probabilitas

Jika X adalah kontinu dan Y adalah diskrit

$$I(X, Y) = H(X) + \sum_{i=1}^k p(Y = i) H(X|Y = i)$$

$H(X)$ adalah entropy


$$H(X) = - \sum_{i=1}^n P(x_i) \log P(x_i)$$

Vector Space Model

term_1	term_2	term_3	term_4	Opini
255	0	0	3	positif
255	255	10	16	positif
20	255	50	10	negatif
50	255	235	76	negatif
0	50	255	0	negatif
50	245	60	34	negatif
234	50	0	11	positif
0	10	235	21	positif
0.20	0.06	0.10	0.00	MI

Seleksi fitur

term_1	term_2	term_3	Opini
255	0	0	positif
255	255	10	positif
20	255	50	negatif
50	255	235	negatif
0	50	255	negatif
50	245	60	negatif
234	50	0	positif
0	10	235	positif
0.20	0.06	0.10	MI



term_4 dihapus
karena tidak
relevan

```
selector = SelectKBest(mutual_info_classif, k=3)
```

Membagi data latih dan uji

term_1	term_2	term_3	Opini
255	0	0	positif
255	255	10	positif
20	255	50	negatif
50	255	235	negatif
0	50	255	negatif
50	245	60	negatif
234	50	0	positif
0	10	235	positif

```
X_train, X_test, y_train, y_test = train_test_split(X_fitur_baru, y, test_size=0.2, random_state=0)
```

Performansi Model

Confusión Matrix		PREDICTED PATTERNS	
		0	1
ACTUAL PATTERNS	0	True Positives TP	False Negatives FN
	1	False Positives FP	True Negatives TN

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$F1Score = 2 * \frac{precision * recall}{precision + recall}$$

$$Accuracy = \frac{\sum TP + TN}{\sum TP + FP + FN + TN}$$

<https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397>

<https://towardsdatascience.com/multi-class-metrics-made-simple-part-i-precision-and-recall-9250280bddc2>

SELESAI

Metode Chi-square

- The χ^2 features selection code builds a [contingency table](#) from its inputs X (feature values) and y (class labels). Each entry i, j corresponds to some feature i and some class j , and holds the sum of the i 'th feature's values across all samples belonging to the class j . It then computes the χ^2 test statistic against expected frequencies arising from the empirical distribution over classes (just their relative frequencies in y) and a uniform distribution over feature values.
- This works when the feature values are frequencies (of terms, for example) because the sum will be the total frequency of a feature (term) in that class. There's no discretization going on.
- It also works quite well in practice when the values are tf-idf values, since those are just weighted/scaled frequencies.