26 November 2020

Tips dan tricks memakai machine learning untuk pattern recognition

Dr. Anto Satriyo Nugroho

Pusat Teknologi Informasi & Komunikasi,
Badan Pengkajian dan Penerapan Teknologi
URL: http://asnugroho.net Email: anto.satriyo@bppt.go.id

Name : Anto Satriyo Nugroho, Dr.Eng.
Birthday: 21-October-1970
Education: Nagoya Institute of Technology, Japan
B.Eng (1995), M.Eng.(2000), Dr.Eng.(2003)
B.Eng : STMDP2 scholarship from Pak Habibie, M.Eng+Dr.Eng: Monbukagakushou scholarship
Core Competence:
Pattern Recognition & Image Processing
Publications: http://asnugroho.net/

Work Experiences

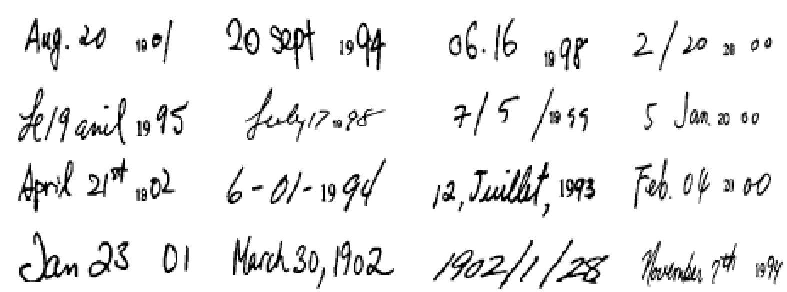
1989-present	: Agency for the Assessment & Application of Technology
2003-2007	: Visiting Professor at Chukyo University, Japan
2007	: Member of ICT Technical Team of General Election Comission (KPU)
2012	: Technical Team of e-KTP (Ministry of Home Affairs)
2015	: Research Consultant at Puslitbang Mabes Polri in Scientific Crime Investigation studies
2017-2019	: President of Indonesian Assoc. for Pattern Recognition
2018-2020	: Governing Board Member of International Association for Pattern Recognition (IAPR), representing Indonesia

Outline

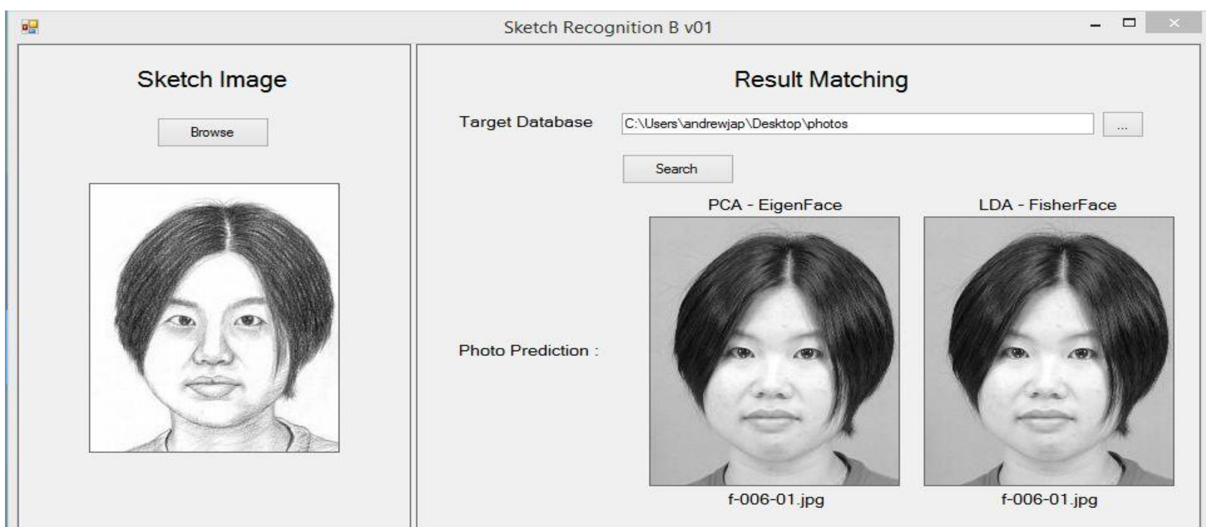
1. Apakah Pattern Recognition / Pengenalan Pola ?
2. Arsitektur Pattern Recognition
3. Review ringkas tentang berbagai metode klasifikasi
4. Memahami berbagai jenis atribut
5. Feature Subset Selection : memilih fitur yang relevan
6. Imbalanced Dataset Problem
7. Performance Evaluation
8. Referensi

Apakah Pattern Recognition ?

- Pattern Recognition : proses mengasosiasikan suatu *pattern* (pola) dengan kategori
- Contoh 1 : mengenali tulisan tangan



- Contoh 2 : mengenali wajah dari sketsa yang dibuat



Outline

1. Apakah Pattern Recognition / Pengenalan Pola ?
2. **Arsitektur Pattern Recognition**
3. Review ringkas tentang berbagai metode klasifikasi
4. Memahami berbagai jenis atribut
5. Feature Subset Selection : memilih fitur yang relevan
6. Imbalanced Dataset Problem
7. Performance Evaluation
8. Referensi

Arsitektur Pattern Recognition

→ Observation → Preprocessing → Feature Extraction/ Feature Subset Selection → Classifier → Kategori

- Observasi : proses mengubah pattern ke dalam format yang bisa diolah oleh mesin
- Preprocessing : melakukan proses awal agar mudah diolah pada tahap selanjutnya
- Feature Extraction : memakai <u>seluruh</u> hasil dari observasi dan memetakan bagian yang bermanfaat ke dimensi yang lebih rendah lewat suatu fungsi
- Feature Subset Selection : memilih subset “terbaik” dari atribut/feature data ditinjau dari kontribusinya terhadap class-separability
- Classifier/Klasifikasi : memetakan hasil reduksi dimensi ke kategori tertentu

Outline

1. Apakah Pattern Recognition / Pengenalan Pola ?
2. Arsitektur Pattern Recognition
3. **Review ringkas tentang berbagai metode klasifikasi**
4. Memahami berbagai jenis atribut
5. Feature Subset Selection : memilih fitur yang relevan
6. Imbalanced Dataset Problem
7. Performance Evaluation
8. Referensi

Ulasan Ringkas Berbagai Metode Klasifikasi

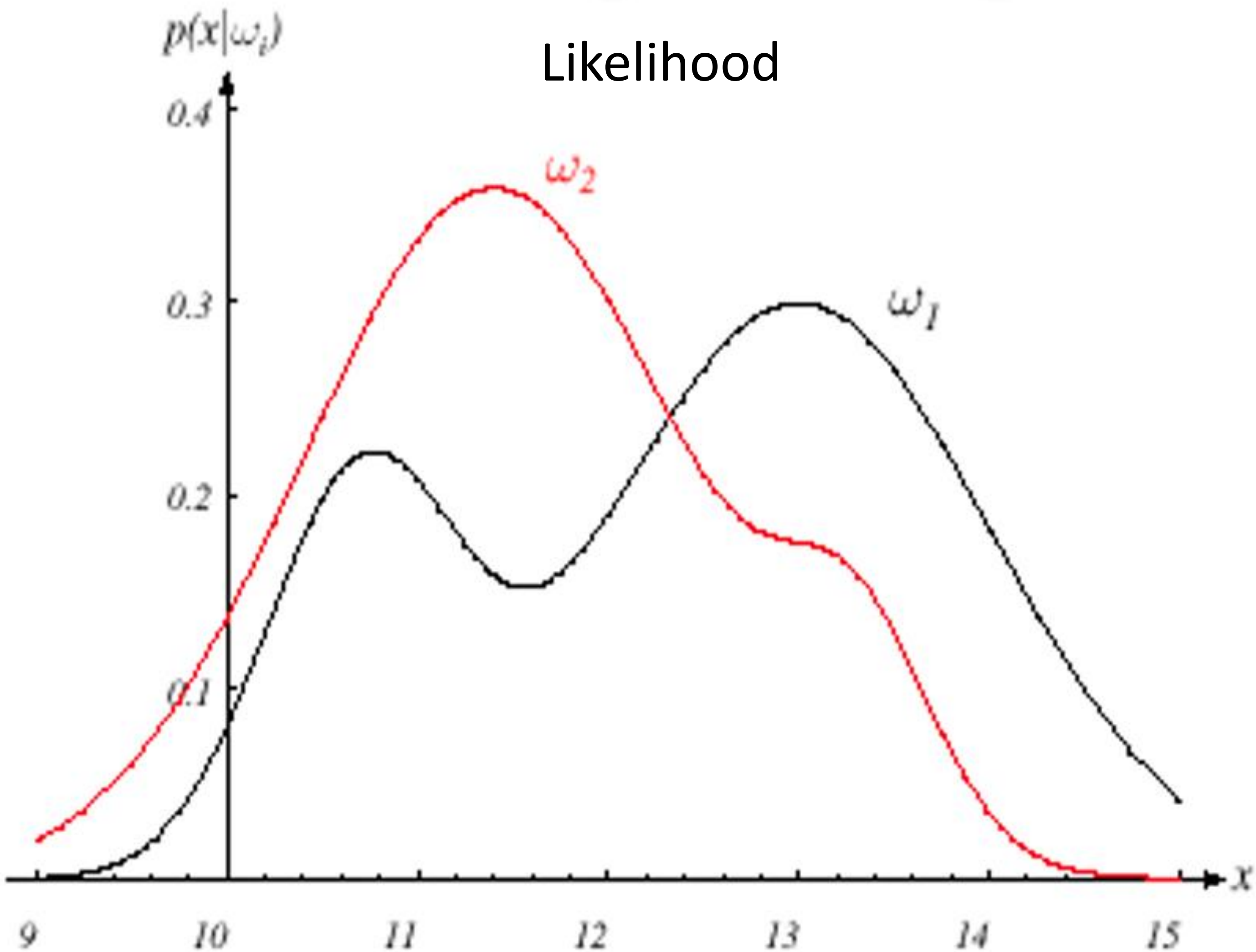
→ Observation → Preprocessing → Feature Extraction/ Feature Subset Selection → Classifier → Kategori

- Naïve Bayes
- K-Nearest Neighbor Classifier
- Multilayer Perceptron
- Deep Learning (contoh : Residual Neural Network yang dikembangkan Microsoft)

Naïve Bayes

- Klasifikasi dilakukan dengan menghitung Posterior Probability suatu class. Untuk menghitung Posterior Probability, diperlukan informasi Prior Probability dan Likelihood
- $Posterior = \frac{Prior \times Likelihood}{Evidence}$
- Likelihood (continuous attribute): $P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

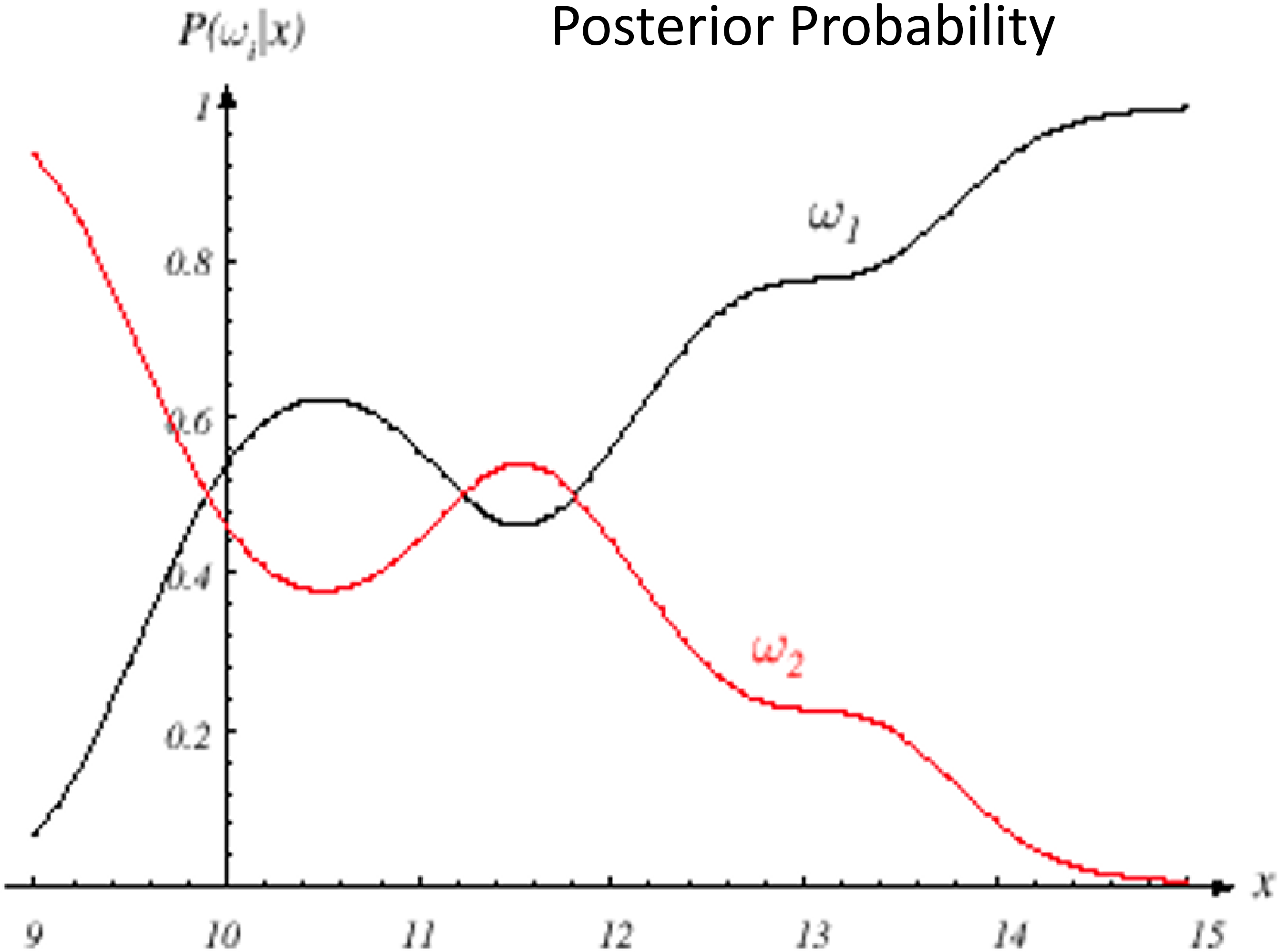
Naïve Bayes



$p(x|\omega_i)$
Likelihood
0.4
0.3
0.2
0.1
ω_2
ω_1
9
10
11
12
13
14
15
x

Naïve Bayes

Posterior Probability



K-Nearest Neighbor Classifier

- Training set terdiri dari *n* sampel dan *c* class diformulasikan sebagai berikut

$$(\boldsymbol{x}_1, \theta_1), (\boldsymbol{x}_2, \theta_2), \cdots, (\boldsymbol{x}_n, \theta_n)$$

$$\theta_p \in \{\omega_1, \omega_2, \cdots, \omega_c\} \qquad (p = 1, \cdots, n)$$

1-Nearest Neighbor classification rule didefinisikan:

$$\min_{p=1,\cdots,n} \{D(\boldsymbol{x}, \boldsymbol{x}_p)\} = D(\boldsymbol{x}, \boldsymbol{x}_k) \quad \Longrightarrow \quad \boldsymbol{x} \in \theta_k$$

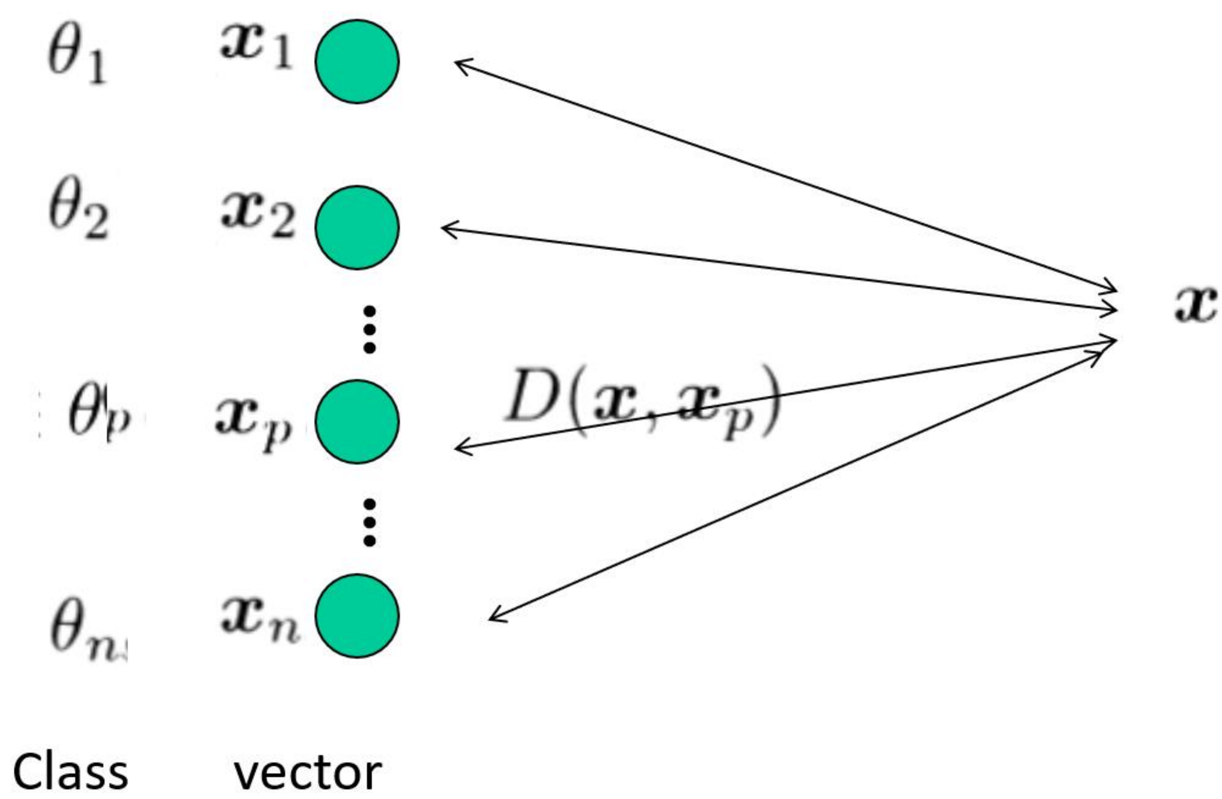
$$\boldsymbol{x}_k \in \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$$

$$\theta_k \in \{\theta_1, \theta_2, \cdots, \theta_n\}$$

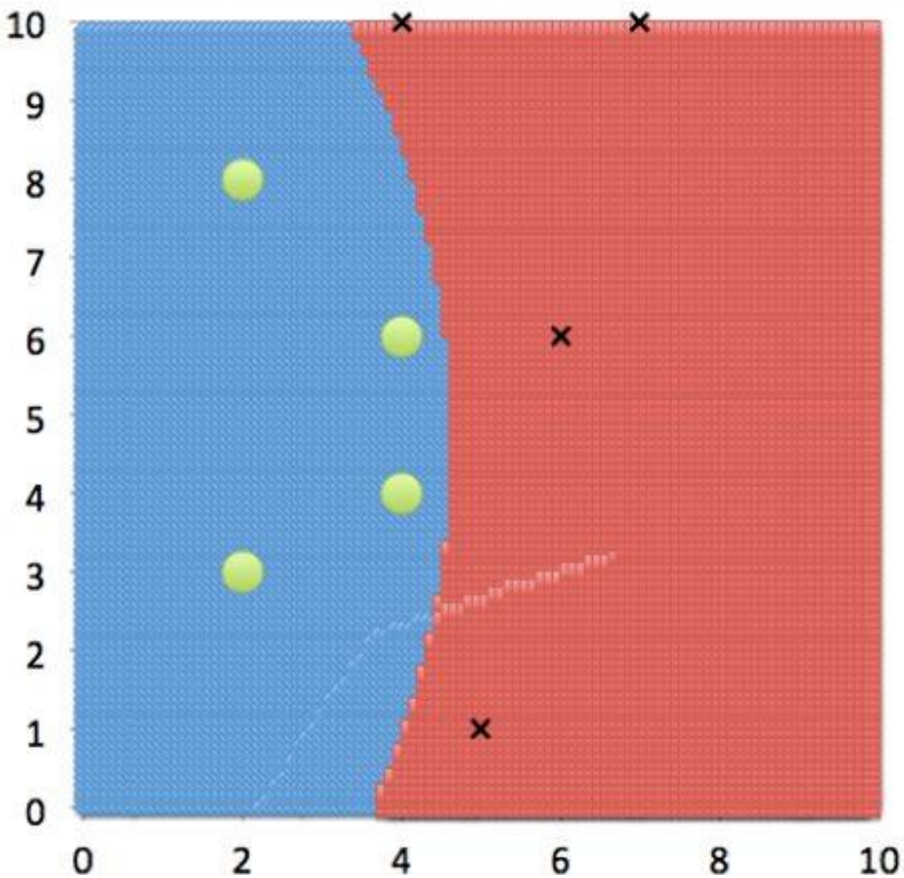
- memerlukan storage untuk menyimpan seluruh data

K-Nearest Neighbor Classifier

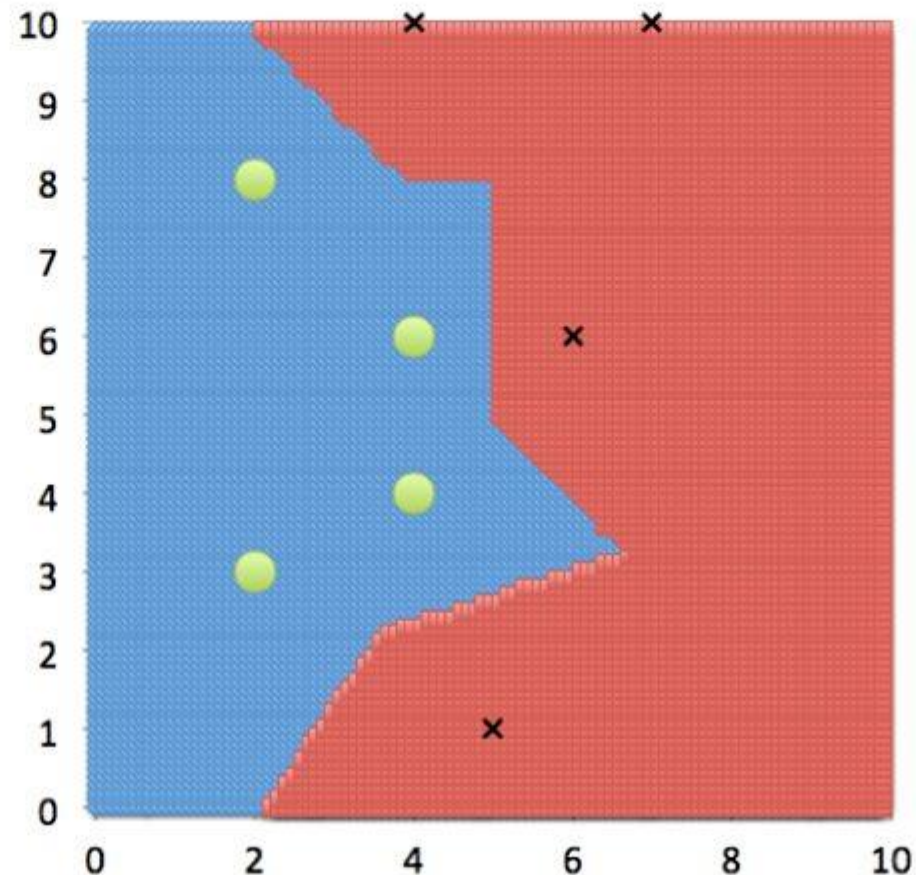
$$\min_{p=1,\cdots,n} \{D(\boldsymbol{x}, \boldsymbol{x}_p)\} = D(\boldsymbol{x}, \boldsymbol{x}_k) \quad \Longrightarrow \quad \boldsymbol{x} \in \theta_k$$



Class Border



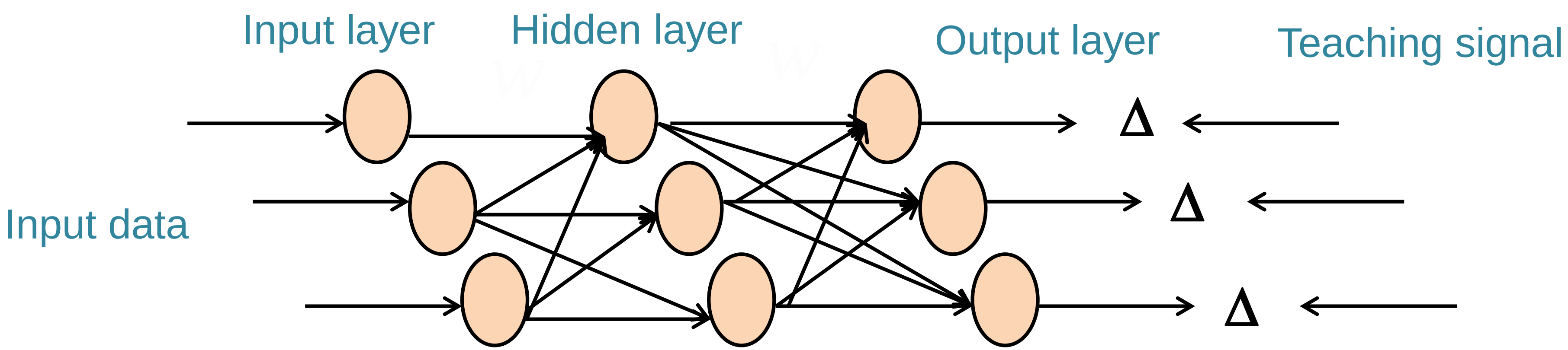
Class border by Naïve Bayes



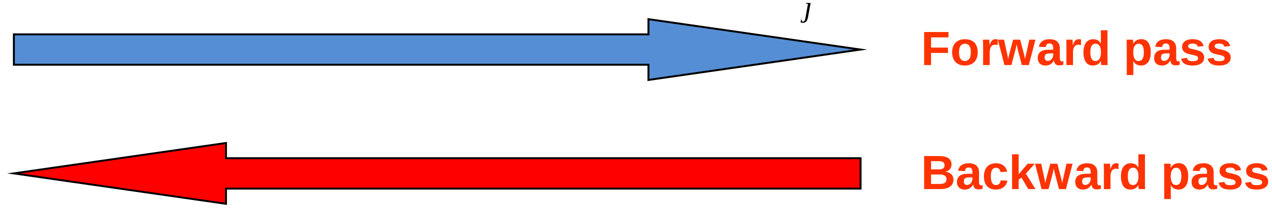
Class border by Nearest Neighbor

Multilayer Perceptron + Backpropagation

- Tiga layer network yang dilatih dengan backpropagation algorithm (Rumelhart)



$$O_k = f(net_k) = \frac{1}{1+e^{-net_k}} \qquad net_k = \theta_k + \sum_j w_{kj} H_j$$

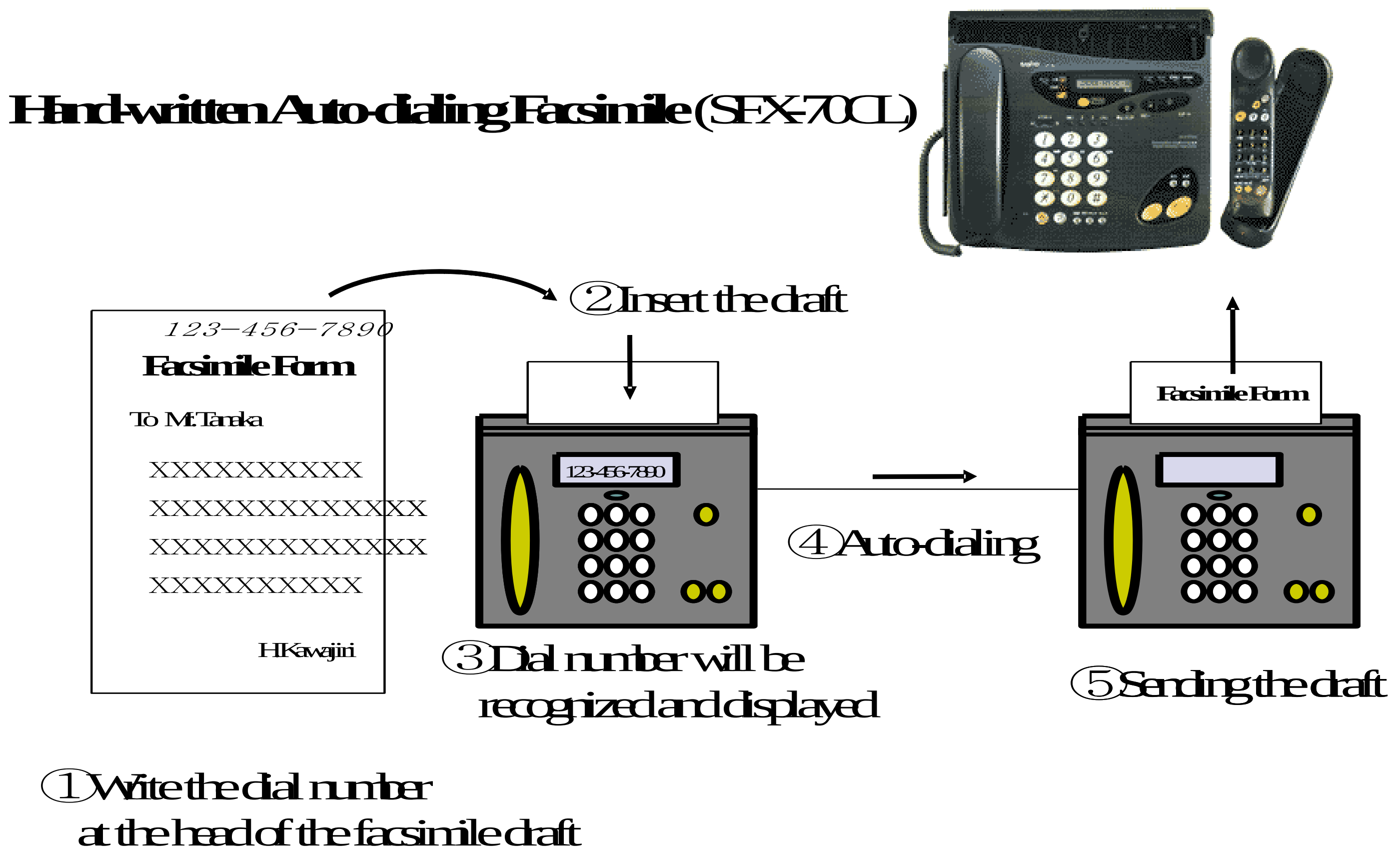


$$w_{new} = w_{old} + \Delta w_{ji}$$

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}} = \eta \delta_j x_i$$

$$\delta_j = H_j (1 - H_j) \sum_k w_{kj} \delta_k$$

Handwriting Character Recognition

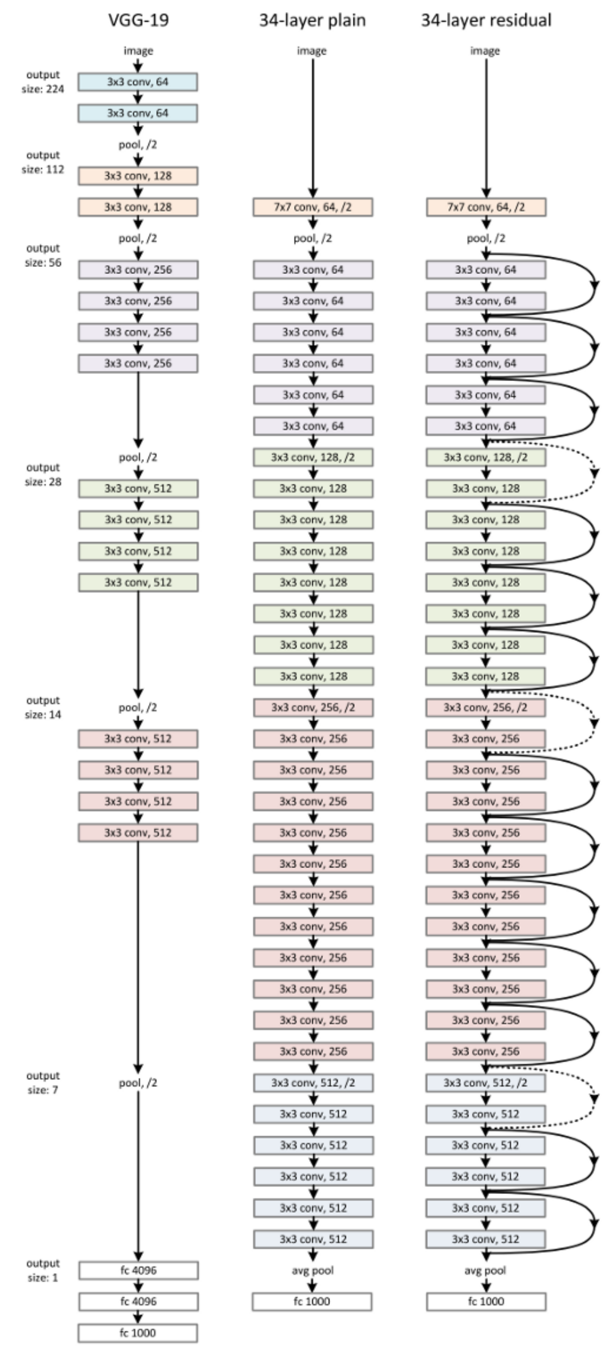


Hand-written Numeric Character Recognition for Facsimile Auto-dialing by Large Scale Neural Network CombNET-II, H. Kawajiri, Y. Takatoshi, T. Junji, A.S. Nugroho and A. Iwata, Proc. of 4th.International Conference on Engineering Applications of Neural Networks, pp.40-46, June 10-12, 1998, Gibraltar

Deep Learning : ResNet-34

- AlexNet memiliki 5 convolutional layers, Visual Geometry Group (VGG) network memiliki 19 layers sedangkan GoogLeNet (inception) memiliki 22 layer
- Residual Network dikembangkan oleh Microsoft
- Versi original : terdiri dari 152 layer
- Saat ini kami memakai ResNet-34 untuk face recognition (Kerjasama dengan Badan Pembinaan Ideologi Pancasila, Kejaksaan Agung dsb)

- https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035



Alur menyelesaikan masalah

→ Observation → Preprocessing → Feature Extraction/ Feature Subset Selection → Classifier → Kategori

- Memahami karakteristik masalah yang diselesaikan
  - Missing data
  - Apakah ada data nominal ?
  - Normalisasi data
  - Class imbalancy
- Menentukan ekstraksi fitur/seleksi fitur (handcrafted feature). Namun apabila memakai deep learning, ekstraksi fitur dilakukan secara otomatis pada modul classifier
- Menentukan metode klasifikasi
- Menentukan metode evaluasi

Apakah cukup memakai classifier “terkini” ?

- Perlunya menganalisa karakteristik data agar pengolahan berlangsung secara <u>benar</u> dan optimal
- Data apabila dalam jumlah sedikit, umumnya metode classifier akan memiliki akurasi yang sama. Tidak banyak perbedaan klasifikasi boundary yang dibentuk.
- Data dalam jumlah sedikit : antara lain data biomedis, bencana alam
- Data dalam jumlah besar ?
  - Lihat dulu, apakah data tidak ada yang redundant ?
  - Apakah linearly separable ?
- Jumlah data tiap class apakah “seimbang” atau terdapat kasus *imbalanced* ?
- Apakah dimensi data bisa direduksi ? Mungkinkah proses pelatihan dilakukan dengan data yang lebih kecil dimensinya ?
- Apabila system diagnosis (contoh) akan dipakai oleh dokter, apakah hasil ekstraksi fitur diperlukan ?

Memahami data dan karakteristiknya menjadi syarat utama sebelum metode klasifikasi ditentukan

Outline

1. Apakah Pattern Recognition / Pengenalan Pola ?
2. Arsitektur Pattern Recognition
3. Review ringkas tentang berbagai metode klasifikasi
4. **Memahami berbagai jenis atribut**
5. Feature Subset Selection : memilih fitur yang relevan
6. Imbalanced Dataset Problem
7. Performance Evaluation
8. Referensi

Empat Jenis Atribut

- Ada empat kategori atribut (disebut juga fitur/feature)
  - Nominal
    - Contoh: ID numbers, warna mata, zip codes
  - Ordinal
    - Contoh: peringkat (e.g., tingkat pedasnya suatu makanan 1-5), tinggi {tinggi, medium, pendek}
  - Interval
    - Contoh: tanggal kalender, suhu dalam Celcius atau Fahrenheit.
  - Ratio
    - Contoh: panjang, waktu, hitungan, suhu dalam Kelvin

- Perlu kehati-hatian ketika memakai atribut nominal

Menangani Atribut Nominal

Instance No.	Color (1:red, 2:green, 3:blue)
A	1
B	2
C	3

Suppose we use column 2 as feature value

Distance (A,B) = 1
Distance (A,C) = 2 → wrong
Distance (B,C) = 1

Menangani Atribut Nominal

Instance No.	Color (1:red, 2:green, 3:blue)
A	1 (1, 0, 0)
B	2 (0, 1, 0)
C	3 (0, 0, 1)

Suppose we use codes in brackets as feature value (1-of-c coding)

Distance (A,B) = 1.4
Distance (A,C) = 1.4
Distance (B,C) = 1.4

Perlunya normalisasi data

Weight (kg)	Height (cm)	Class
70	170	Soccer
120	190	Sumo

Test datum X: weight: 80 kg height: 185 cm

Distance with Soccer : 18.0
Distance with Sumo : 40.3

Thus, the nearest neighbor of X is Soccer player

Perlunya normalisasi data

Weight (kg)	Height (mm)	Class
70	1700	Soccer
120	1900	Sumo

Test datum X: weight: 80 kg height: 1850 mm

Distance with Soccer : 150.3
Distance with Sumo : 64.0

Thus, the nearest neighbor of X is Sumo player

Perlunya normalisasi data

- Convert the values the feature to have the same range, e.g. [0,1]
- Suppose, the feature is x, and the new feature is y

$$y = \frac{x - \min}{\max - \min}$$

Perlunya normalisasi data

Back to the example. Suppose the minimum and maximum value of each feature are as follows:

	Weight (kg)	Height (cm)
minimum	50	160
maximum	150	200

then we can transform the input into :

Perlunya normalisasi data

Weight (kg)	Height (mm)	Class
0.2	0.25	Soccer
0.7	0.75	Sumo

Test datum X: weight: 80 kg height: 1850 cm is also normalized to become

weight: 0.3 height: 0.625

Distance with Soccer : 0.39
Distance with Sumo : 0.42

Thus, the nearest neighbor of X is Soccer player

Outline

1. Apakah Pattern Recognition / Pengenalan Pola ?
2. Arsitektur Pattern Recognition
3. Review ringkas tentang berbagai metode klasifikasi
4. Memahami berbagai jenis atribut
5. **Feature Subset Selection : memilih fitur yang relevan**
6. Imbalanced Dataset Problem
7. Performance Evaluation
8. Referensi

Feature Subset Selection

- Definisi: memilih subset “terbaik” dari atribut/feature data ditinjau dari kontribusinya terhadap class-separability

- Dua hal yang perlu dibahas :
  - Bagaimana subset fitur dipilih ? (lihat taxonomi)
  - Bagaimana mengevaluasi subset fitur ?

Taksonomi Feature (Subset) Selection

- Feature Selection — Jain& Zongker(1997)
  - Statistical Pattern Recognition
    - suboptimal
      - single solution
        - deterministic
          - *PTA(l,r)*
          - *Floating (SFFS, SBFS)*
          - *SFS, SBS*
          - *Max-Min*
        - stochastic
      - many solutions
        - deterministic
          - *beam search*
        - stochastic
          - *GA*
    - optimal
      - *Exhaustive search*
      - *branch-and-bound*
  - Artificial Neural Network
    - *node pruning*

Anil Jain, Douglas Zongker, “Feature Selection: Evaluation, Application, and Small Sample Performance,“ IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 19, Issue 2, pp.153--158, February 1997

Sequential Forward Selection(SFS)

sequentially select a feature that become the most significant to the current condition

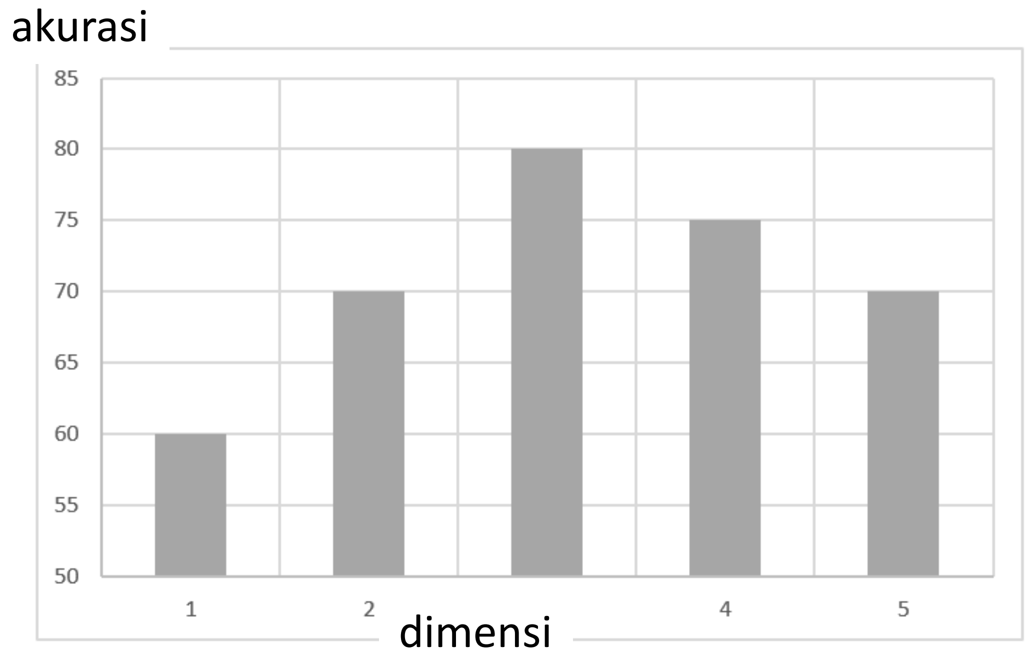
Input: *F* – full set, *U* – *measure*

initialize: S = { }
repeat
(1) *f* = FindNxt(F)
(2) *S* = *S* U { f }
(3) *F* = *F* – { f }
until *S* satisfies *U* or *F* = { }
Output: *S*

A.W,Whitney,"A direct method of nonparametric measurement selection," IEEE Trans. Comput.20,pp.1100-1103, 1997.

Sequential Forward Selection(SFS)

a 50%	b 60%	c 55%	d 40%	e 50%	{b} terbaik
ba 55%	bc 60%	bd 70%	be 65%		{b, d} terbaik
bda 80%	bdc 70%	bde 75%			{b, d, a} terbaik
bdac 70%	bdae 75%				{b, d, a, e} terbaik
bdaec 70%					{b, d, a, e, c)



Metode ini memakai pendekatan *wrapper*. Kualitas subset fitur yang dipilih dievaluasi memakai akurasi suatu *classifier.*

Sequential Backward Selection (SBS)

sequentially remove a feature that is the most non-significant to the present condition

Input: F – full set, U – *measure*

initialize: S = { } */* S holds the removed features */*

repeat

(1) f = GetNxt(F)

(2) $F = F - \{ f \}$

(3) $S = S \cup \{ f \}$

until *F does not* satisfy *U* or *F* = { }

Output: F U { f }

Marill,T,D.M.Green,"On the effectiveness of receptors in recognition system," IEEE Trans. Inform. Theory 9, pp.11-17, 1963.

Sequential Floating Forward Selection

Sequentially applying several SBS steps after SFS

- Solusi terhadap nesting problem pada algoritma SFS dan SBS
- Hasil terbaik dibandingkan metode yang lain (Jain & Zonker, 1997)
- Computational Complexity still a problem

P.Pudil, J.Novovicova, J.Kittler," Floating Search Methods in Feature Selection," Pattern Recognition Letters, vol.15, no.11, pp.279-283, 1994.

Kriteria Evaluasi Subset

- **Wrapper**
  - Memakai akurasi classifier untuk mengevaluasi performance dari subset yang dipilih
  - Masalah :
    - bagaimana memilih classifier
    - *Computational complexity*
- **Filter**
  - Berbasis distance/information measures yang merupakan *intrinsic properties* pada data
  - Cheaper in time complexity
  - Tepat dipakai untuk data berdimensi sangat tinggi

Fisher Criterion

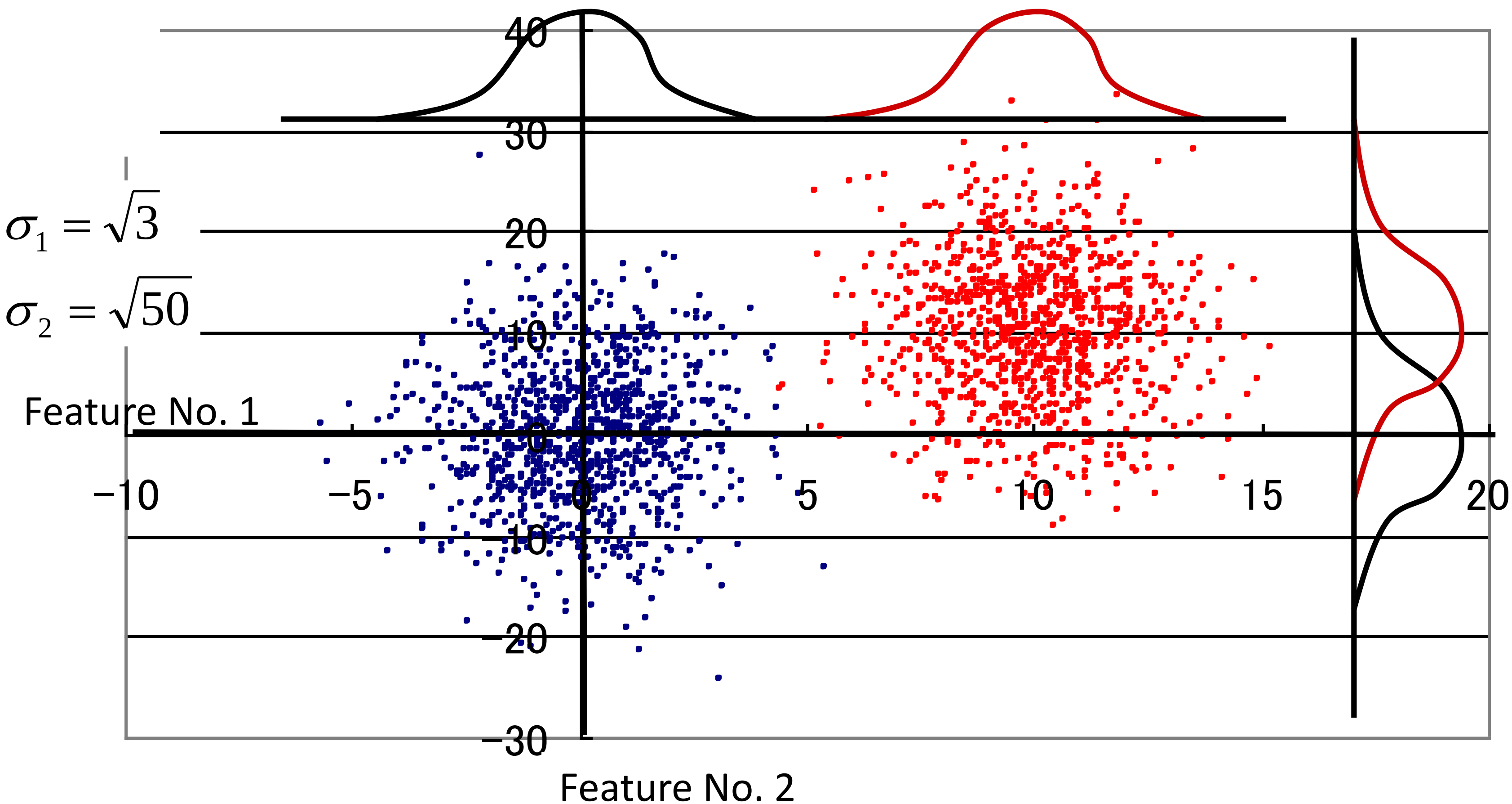
Kriteria Fisher memilih salah satu feature secara independent yang paling baik dalam memisahkan dua class.

$$F(x_j) = \frac{n_{+1} n_{-1}}{n_{+1} + n_{-1}} \frac{(\mu_{j,+1} - \mu_{j,-1})^2}{n_{j,+1}\sigma^2_{j,+1} + n_{j,-1}\sigma^2_{j,-1}}$$

Score of the j^{th} feature of vector x

Num. of patterns in Negative Class

Num. of patterns in Positive Class



- F(Feature No.1) = 2.9
- F(Feature No.2) = 0.7
- F(Feature No.1) > F(Feature No.2)
- Feature No.1 is better than No.2

Outline

1. Apakah Pattern Recognition / Pengenalan Pola ?
2. Arsitektur Pattern Recognition
3. Review ringkas tentang berbagai metode klasifikasi
4. Memahami berbagai jenis atribut
5. Feature Subset Selection : memilih fitur yang relevan
6. **Imbalanced Dataset Problem**
7. Performance Evaluation
8. Referensi

Imbalanced Dataset Problem

- Definisi : suatu kelas/kategori memiliki jumlah sampel jauh lebih banyak daripada kelas/kategori yang lain
- Contoh : munculnya super-cooling fog, prediksi abnormality pada penyakit
- Solusi
  - Modifikasi pada sisi data training : upsampling, downsampling
  - Modifikasi pada sisi metode
  - Kombinasi
- Hati-hati dengan formula untuk mengevaluasi hasil : secara umum geometric mean lebih disukai daripada arithmetical mean

$$Akurasi = \sqrt{RR_1 RR_2 RR_3 \ldots RR_n}$$

- Contoh : kelas positif : 90 sample (semua diprediksi benar)
  kelas negative: 10 sample (2 diprediksi benar)

Arithmetic Mean : akurasi = (90+2)/100 = 92 %

Geometric Mean : akurasi = $\sqrt{100\% \times 20\%} = 45\%$

Meteorological Forecasting

- Predicting fog event based on meteorological observation
- The prediction was held every 30 minutes and the result was used to support aircraft navigation
- The number of fog events was very small compared to no fog events which can be considered as a pattern classification problem involving imbalanced training sets
- Observation was held every 30 minutes, in Long.141.70 E, 42.77 Lat., 25 m above sea level by Shin Chitose Meteorological Observatory Station (Hokkaido Island, Japan)
- Fog Event is defined for condition where
  - Range of Visibility < 1000 m
  - Weather shows the appearance of the fog
- Part of this research was a participation in 1999 Fog Forecasting Contest organized by Neuro Computing Technical Group of IEICE-Japan

Meteorological Forecasting

Observed Information

No.	Meteorological Information	No.	Meteorological Information
1	Year	14	Weather
2	Month	15	Cloudiness (1st layer)
3	Date	16	Cloud Shape (1st layer)
4	Time	17	Cloud Height (1st layer)
5	Atmospheric Pressure [hPA]	18	Cloudiness (2st layer)
6	Temperature [ºC]	19	Cloud Shape (2st layer)
7	Dew Point Temperature [ºC]	20	Cloud Height (2st layer)
8	Wind Direction [º]	21	Cloudiness (3st layer)
9	Wind Speed [m/s]	22	Cloud Shape (3st layer)
10	Max.Inst.Wind Speed [m/s]	23	Cloud Height (3st layer)
11	Change of Wind (1) [º]	24	Cloudiness (4st layer)
12	Change of Wind (1) [º]	25	Cloud Shape (4st layer)
13	Range of Visibility	26	Cloud Height (4st layer)

Example :

1984 1 1 4.5 1008 0.0 –7.0 270 6 –1 –1 –1 9999 85 0 2 10 0 4 25 –1 –1 –1 –1 –1 –1

Imbalanced Problem

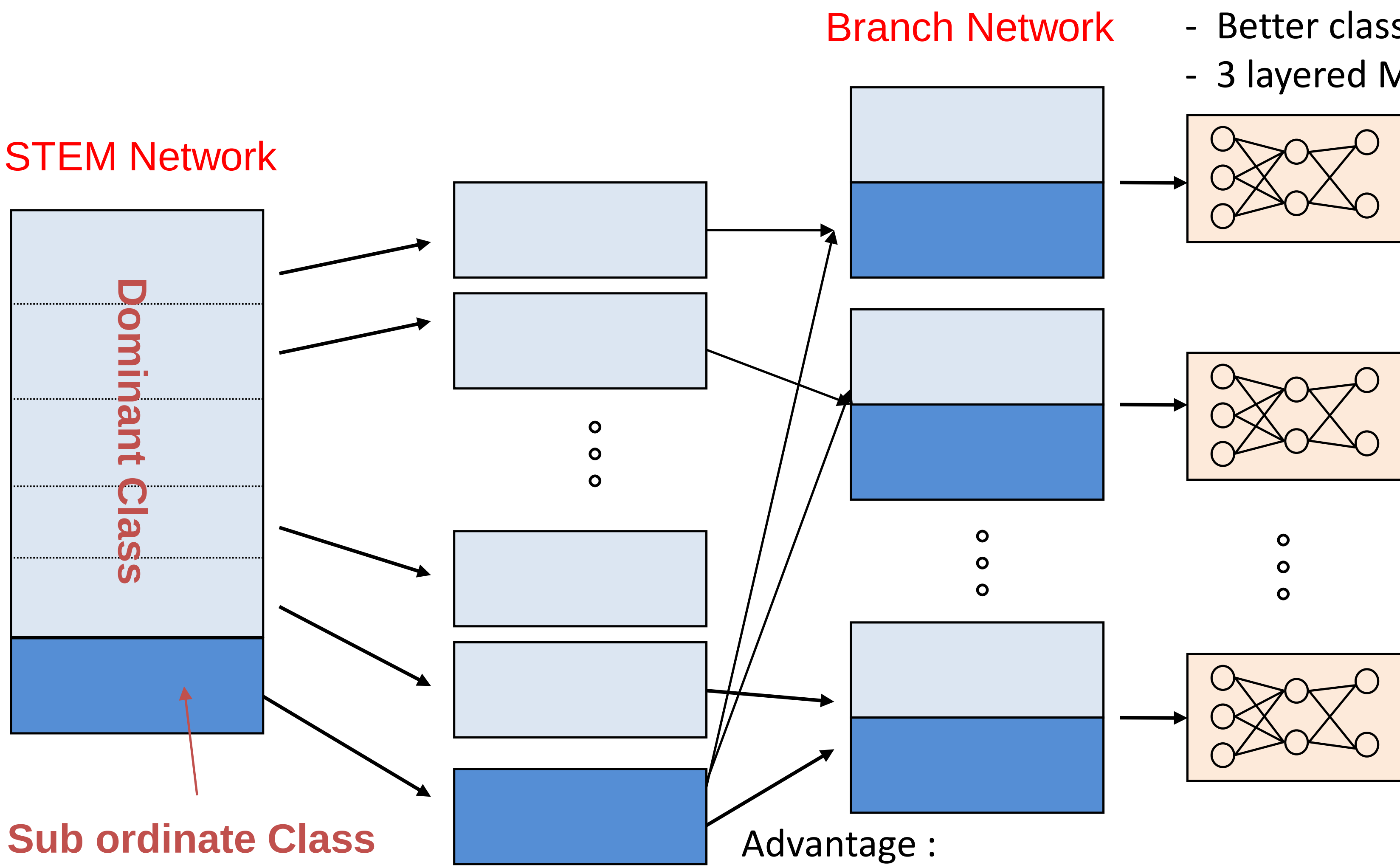
- Observation was held every 30 minutes, in Long.141.70° E, 42.77°Lat., 25 m above sea level by Shin Chitose Meteorological Observatory Station (Hokkaido Island, Japan)
- The observed items consist of atmospheric pressure, dew point temperature, wind direction, wind speed, change of wind, range of visibility, weather, cloudiness, cloud shape, cloud height, etc. of total 26 items, and provided in numeric expression
- Fog Event is defined for condition where
  - Range of Visibility < 1000 m
  - Weather shows the appearance of the fog (expressed as value of 40-49)

Year	1984	1985	1986	1987	1988	1989
Number of Fog Event	467	426	314	275	282	251
Number of NoFog Event	16961	17033	17130	17172	17260	17218
Ratio	1 : 36.3	1 : 40.0	1 : 54.5	1 : 62.4	1 : 61.2	1 : 68.6

Year	1990	1991	1992	1993	1994	1995
Number of Fog Event	220	220	389	211	298	288
Number of NoFog Event	17199	17272	17163	17301	17211	17222
Ratio	1 : 78.1	1 : 78.5	1 : 44.1	1 : 82	1 : 57.8	1 : 59.8

Metode yang diusulkan

- Better classification
- 3 layered MLP (BP)



Stem Network is trained by Self Growing Algorithm to do VQ of the dominant class

Advantage :

- reducing the imbalancy
- branch modules can be trained individually, that may reduce the training time

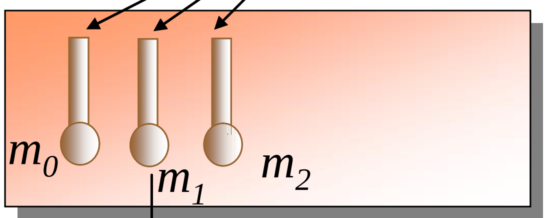
Self-Growing Algorithm

Input vector : $\vec{x}$

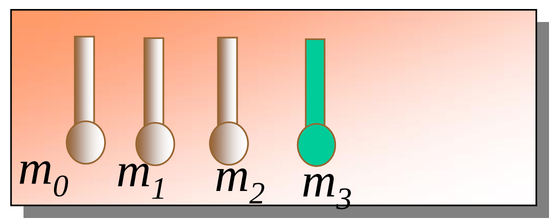
IJCNN92-Baltimore, June 7-11, 1992

Calculate similarity between $\vec{x}$ and the stem neurons, and find the stem neuron with the best similarity

Suppose that neuron m_1 gives the best similarity expressed as $sim(m_1,\vec{x})$



$sim(\vec{x}, m_1) < r_{th}$

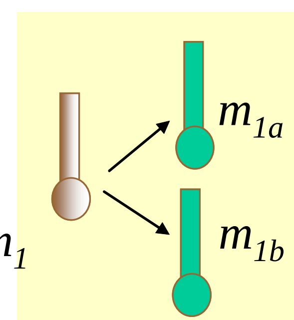


NEURON CREATION

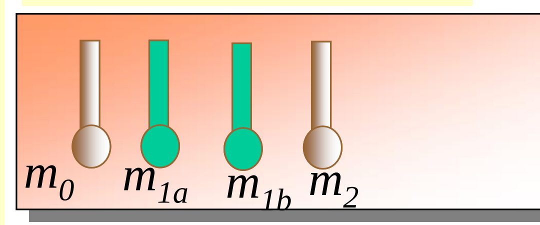
Create a new stem neuron, with $\vec{x}$ as initial reference vector

If $sim(\vec{x}, m_1) <$ similarity threshold r_{th}, then $\vec{x}$ is recorded as the member of group corresponding to neuron m_1

If inner potential (member) of neuron m_1 exceeds a threshold h_{th} then divide the neuron m_1 Into 2 neurons : m_{1a} and m_{1b}



NEURON DIVISON



Experimental Results

	CombNET-II	1-NN	MLP (BP)
Correct Prediction Rate (for 1989) Num. Of wrong predictions	99.10 % 158	99.06 % 164	98.44 % 272
Num. of synapses (training) Ratio Num. of synapses (prediction) Ratio	150876 18.2 9742 1.2	7001360 843.3 7001360 843.3	8302 1.0 8302 1.0
Network Structure	Stem Neuron : 18 Branch : 80-100-2	2 classes 87517 templ.	80-100-2
Time needed to train the network	3 days (5 machines) 19.6 hrs x 18	-	3 weeks (402:17:38)

Proposed Method	Fog Events (539 correct)		Num. of false prediction.
	Predictions	Correctly Pred.	
CombNET-II	622	374	**370**
Probabilistic NN	169	127	445
Modified Counter Propagation NN	908	178	734

Experimental Results

This study **won the first prize award** in the 1999 Fog Forecasting Contest organized by Neurocomputing Technical Group of IEICE-Japan


最優秀賞

名古屋工業大学
Anto Satriyo Nugroho,
黒柳 奨，岩田 彰 殿

あなたは本研究会が主催する気象データコンテストに表記の成績を修められたことを証します。

平成11年3月28日

電子情報通信学会
ニューロコンピューティング研究専門委員会
委員長 石川 眞澄

Outline

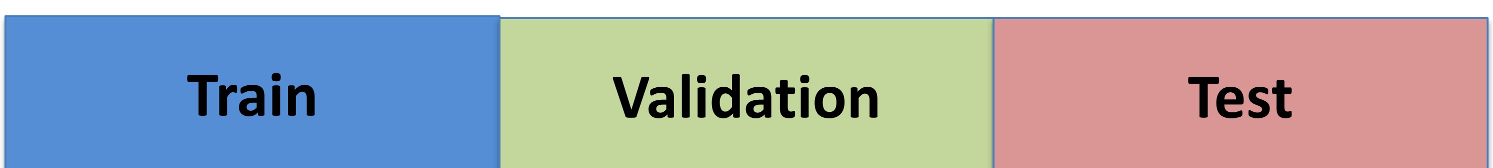
1. Apakah Pattern Recognition / Pengenalan Pola ?
2. Arsitektur Pattern Recognition
3. Review ringkas tentang berbagai metode klasifikasi
4. Memahami berbagai jenis atribut
5. Feature Subset Selection : memilih fitur yang relevan
6. Imbalanced Dataset Problem
7. **Performance Evaluation**
8. Referensi

Model Assessment

- Generalization : prediksi kapabilitas suatu classifier pada data test yang independent
- Model Selection : perkiraan kinerja berbagai model yang berbeda untuk menentukan model terbaik
- Model Assessment : memakai model yang dipilih, untuk mengukur kinerjanya (generalization error) dari tingkat error pada data yang lain dengan yang dipakai di atas

Best approach

- The data is divided into 3 parts : training set, validation set and testing set
  - The training set is used to fit the model
  - the validation set is used to estimate prediction error for model selection
  - the test set is used for assessment of the generalization error of the final chosen model. Ideally, the test set should be kept in a “vault,” and be brought out only at the end of the data analysis.
- Suitable for a data-rich situation



Hold-Out (H) Method

- Data are divided into two parts :
  - Training Set : The training set is used to fit the model
  - Testing Set : for assessment of the generalization error of the model

- Suitable for data-rich situation

K-fold Cross Validation (CV) Method

1	2	3	4	5
Train	Train	Validation	Train	Train

- For the k th part (third above), we fit the model to the other K − 1 parts of the data, and calculate the prediction error of the fitted model when predicting the k th part of the data. We do this for k = 1, 2, . . . ,K and combine the K estimates of prediction error
- K is usually 5 or 10
- Leave-One-Out Cross Validation : K = N (number of samples)

Penutup

- Tips dan Trick dalam pattern recognition sangat banyak dibahas dalam paper akademik maupun buku teks pattern recognition/datamining/Artificial Intelligence
- Pemahaman masalah secara benar sangat penting sebelum menentukan metode classifier yang akan dipakai