

Rozpoznawanie obiektów na podstawie zredukowanego zbioru cech

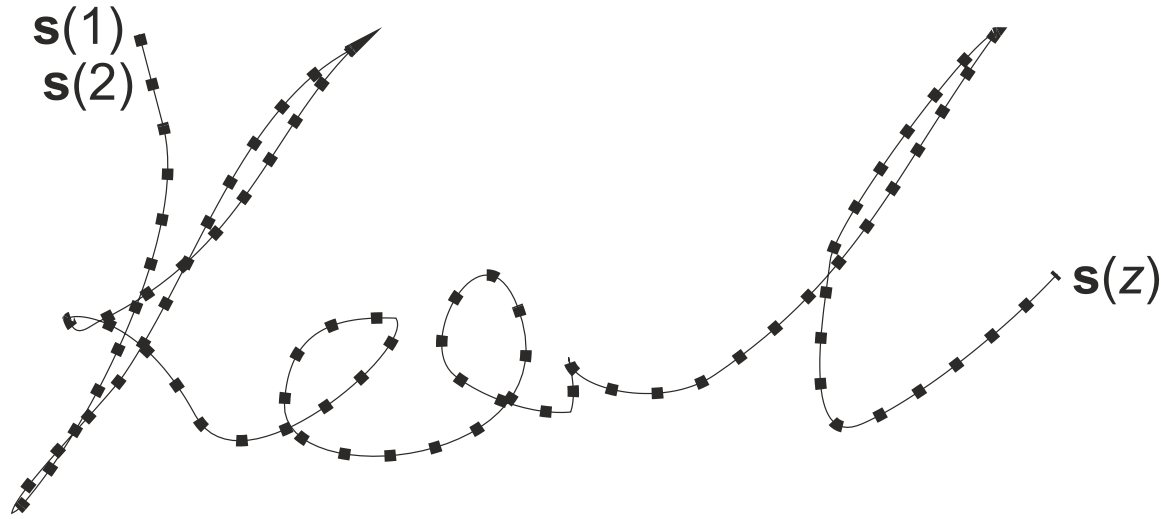
Piotr Porwik
Uniwersytet Śląski
w Katowicach

??

„It is obvious that more does not mean better,
especially in the case of classifiers!!” *)

*) XIAO-HUA ZHOU et al..Statistical Methods in Diagnostic
Medicine. Wiley-Interscience, 2002,

Cechy podpisu



$$S = \{s(1), s(2), \dots, s(z)\} .$$

$$s(t) \quad t = 1, \dots, z$$

$$F = \{f_1, f_2, \dots, f_u\}$$

3 $s(t) = [f_1^t, f_2^t, \dots, f_u^t]$

Czy klasyfikacja jest zawsze możliwa?

NIE!



Podpis kapitana Bonaparte w roku 1793 (ma 24 lata)



Generał dywizji (1796) (ma 27 lat)



Napoleon jak cesarz Francji 1804 (ma 35 lat)



Dowódca Francuzów pod Austerlitz 1805 (ma 36 lat)



Dowódca Francuzów pod Moskwą 1812 (ma 43 lata)

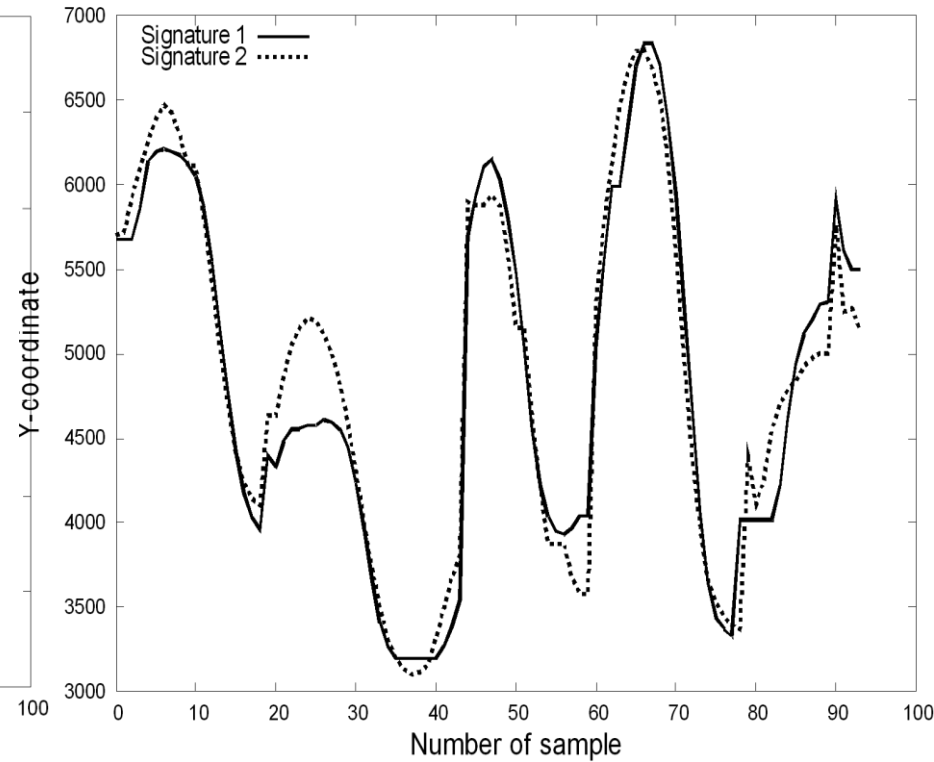
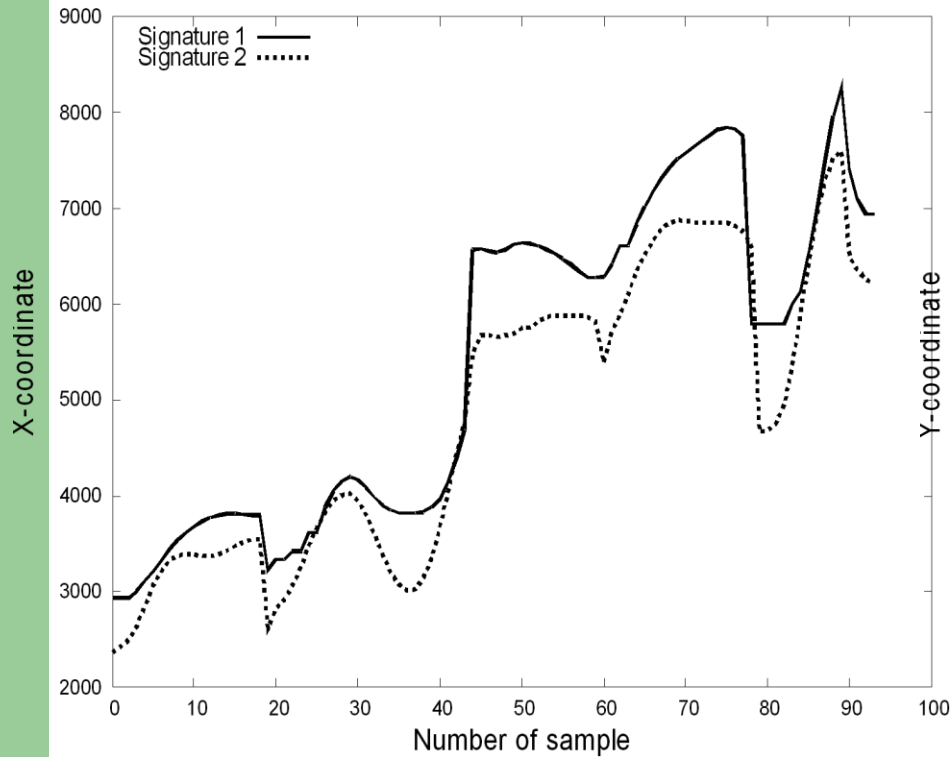


Przed śmiercią na wyspie Św. Heleny 1821 (ma 52 lata)

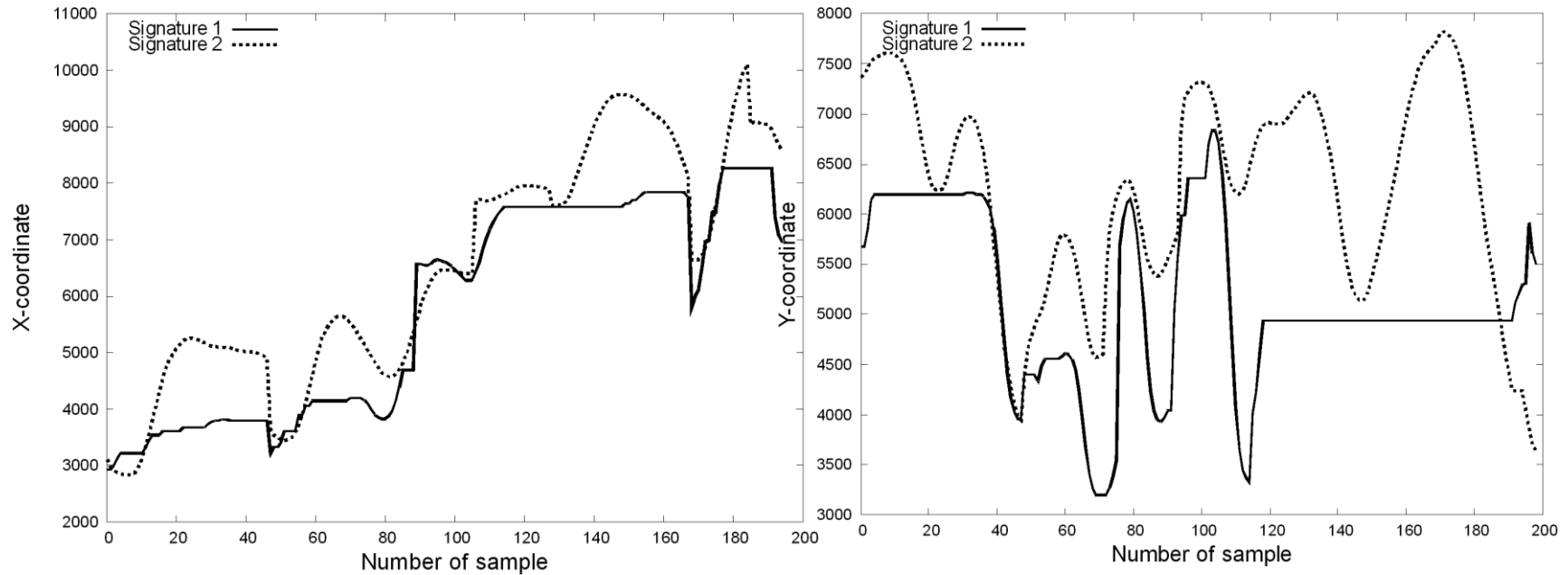


1769 Korsyka – 1821 Św. Helena

Cech obiektu – tutaj podpis (1)



Cechy obiektu –tutaj podpis (2)



Miary podobieństwa (1)

1	Euclidean	12	Jaccard
2	Gower	13	Fidelity
3	Minkowski	14	Bhattacharyya
4	City Block	15	Hellinger
5	Cosine	16	Matusita
6	Kulczynski	17	Pearson
7	Canberra	18	Neyman χ^2
8	Czekanowski	19	Squared χ^2
9	Intersection	20	Symmetric χ^2
10	Clark	21	Kullback–Leibler
11	Lorentzian	22	Kumar-Hassebrook

Miary podobieństwa (2)

$$ER^2(f_i) = \frac{\left[\sum_{i=1}^z (p_i - \bar{p}_i)(q_i - \bar{q}_i) \right]^2}{\sum_{i=1}^z (p_i - \bar{p}_i)^2 \sum_{i=1}^z (q_i - \bar{q}_i)^2}$$

$$d_{Euc}(f_i) = \left(\sum_{i=1}^z |p_i - q_i|^2 \right)^{1/2}$$

$$S_{Jac}(f_i) = \frac{\sum_{i=1}^z p_i q_i}{\sum_{i=1}^z p_i^2 + \sum_{i=1}^z q_i^2 - \sum_{i=1}^z p_i q_i}$$

Sung-Hyug Cha (2007) Comprehensive survey on distance/similarity measures between probability density functions. Int. J. of Mathematical Models and Methods in Applied Sciences, vol. 1, pp. 300-307.

Nowa koncepcja podobieństwa (1)

Zbiór cech: $f_m \in F \quad m=1,\dots,u$

Zbiór metod :

(sposoby wyznaczania podobieństwa) $\omega_j \in M \quad j=1,\dots,k$

Zbiór wszystkich kombinacji „cecha-metoda”:

$$FM = \{\varepsilon_i \prec (f_m, \omega_j)_i : f_m \in F, \omega_j \in M\}, \quad i=1,\dots,u \cdot k,$$

Niech nowy współczynnik nazywa się *Sim*

Nowa koncepcja podobieństwa (2)

Podpisy oryginalne: $\pi_1 = \{S_1, S_2, \dots, S_c\}$

Podpisy fałszywe: $\pi_2 = \{S_1^\Delta, S_2^\Delta, \dots, S_d^\Delta\}$

$$\mathbf{X} = \left[S_1 \leftrightarrow S_2, \dots, S_1 \leftrightarrow S_c, \dots, S_{c-1} \leftrightarrow S_c \right]_{(uk) \times \binom{c}{2}} = [\mathbf{x}_1, \dots, \mathbf{x}_{\binom{c}{2}}]$$

$$\mathbf{Y} = \left[\left[S_1 \leftrightarrow S_1^\Delta \right], \dots, \left[S_1 \leftrightarrow S_d^\Delta \right], \dots, \left[S_c \leftrightarrow S_d^\Delta \right] \right]_{(uk) \times (cd)} = [\mathbf{y}_1, \dots, \mathbf{y}_{cd}]$$

c – liczba podpisów oryginalnych,

d – liczba podpisów fałszywych,

u – liczba zarejestrowanych cech,

k – liczba użytych miar podobieństwa (metod pomiaru)

Nowa koncepcja podobieństwa (3)

Współczynnik *Sim* pomiędzy podpisami oryginalnymi:

$$\mathbf{X} \ni \mathbf{x}_1 = [S_1 \leftrightarrow S_2] = \begin{bmatrix} Sim(S_1, S_2)^{(f_1, \omega_1)_{1,1}} \\ \cdot \\ Sim(S_1, S_2)^{(f_1, \omega_k)_{1,k}} \\ \cdot \\ Sim(S_1, S_2)^{(f_u, \omega_1)_{u,1}} \\ \cdot \\ Sim(S_1, S_2)^{(f_u, \omega_k)_{u,k}} \end{bmatrix}_{(u \cdot k) \times 1}$$

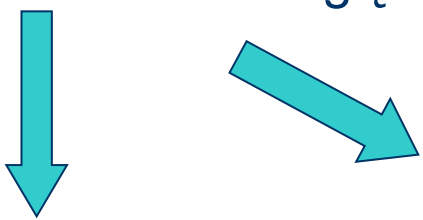
Nowa koncepcja podobieństwa (4)

Współczynnik Sim pomiędzy podpisami prawdziwymi i fałszywymi:

$$\mathbf{Y} \ni \mathbf{y}_1 = [S_1 \leftrightarrow S_1^\Delta] = \begin{bmatrix} Sim(S_1, S_1^\Delta)^{(f_1, \omega_1)_{1 \cdot 1}} \\ \cdot \\ Sim(S_1, S_1^\Delta)^{(f_1, \omega_k)_{1 \cdot k}} \\ \cdot \\ Sim(S_1, S_1^\Delta)^{(f_u, \omega_1)_{1 \cdot u}} \\ \cdot \\ Sim(S_1, S_1^\Delta)^{(f_u, \omega_k)_{u \cdot k}} \end{bmatrix}_{(u \cdot k) \times 1}$$

Nowa koncepcja podobieństwa (5)

Macierze X oraz Y mogą być jednak bardzo duże!


$$(u \cdot k) \times \begin{pmatrix} c \\ 2 \end{pmatrix}$$
$$(u \cdot k) \times (c \cdot d)$$

Rozwiązanie: redukcja rozmiarów macierzy. Na przykład:

- PCA
- SVD
- Metoda Hotellinga

Dwie pierwsze metod są dobrze znane i opisane w literaturze. Metoda Hotellinga jest mniej znana ale tutaj daje najlepsze rezultaty.

Szkic metody Hotellinga (1)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdot & x_{1n} \\ x_{21} & x_{22} & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ x_{p1} & x_{p2} & \cdot & x_{pn} \end{bmatrix} = \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$$

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdot & y_{1m} \\ y_{21} & y_{22} & \cdot & y_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ y_{p1} & y_{p2} & \cdot & y_{pm} \end{bmatrix} = \mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$$

Szkic metody Hotellinga (2)

Elementy macierzy \mathbf{X} oraz \mathbf{Y} są próbkami dwóch populacji:

$$\mathbf{x}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{y}_i \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Rozkład jest normalny, a parametry tego rozkładu są nieznane. Można je jednak estymować:

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{X} \cdot \mathbf{j} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \cdot \\ \bar{x}_p \end{bmatrix}$$

$$\bar{\mathbf{y}} = \frac{1}{m} \mathbf{Y} \cdot \mathbf{g} = \frac{1}{m} \sum_{i=1}^m \mathbf{y}_i = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \cdot \\ \bar{y}_p \end{bmatrix}$$

$$\mathbf{S}_1 = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$\mathbf{S}_2 = \frac{1}{m-1} \sum_{i=1}^m (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^T$$

Szkic metody Hotellinga (3)

* - test Bartletta

Jeśli kowariancje **S1** oraz **S2** są homogeniczne*, to:

$$V_1 = \frac{S_1(n-1) + S_2(m-1)}{n+m-2} \quad (\text{wariancja wspólna})$$

w przeciwnym przypadku:

$$V_2 = \frac{S_1}{n} + \frac{S_2}{m} \quad (\text{wariancje oddzielnie})$$

Statystyka Hotellinga: $T^2 = (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T V_i^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}}) \quad i = \{1, 2\}$

Test Hotellinga można sprowadzić do testu F :

$$\tilde{F} = \frac{n+m-p-1}{p(n+m-2)} T^2, \quad \longrightarrow \quad \tilde{F} \sim F_{p, n+m-p-1, \alpha}$$

Poziom istotności

Szkic metody Hotellinga (4)

Algorithm 1

Let ε_i be the i^{th} element in the set FM ;

Let p be the cardinality of the set FM ;

Let $\mathbf{J} = [\mathbf{X} \ \mathbf{Y}]$, where \mathbf{X} and \mathbf{Y} are the matrices include the values *Sim* of composed features, ordered according to formulas (9) and (11). Each i -th row of matrix \mathbf{J} contains *Sim* coefficients associated with appropriate $\varepsilon_i = (f_m, \omega_j)_i \in FM$ pair, where $i = 1, 2, \dots, p$ and $p = u \cdot k$.

Let $T^2(\varepsilon_1, \dots, \varepsilon_p)$ be the Hotelling's statistics for the matrix \mathbf{J} ;

Let U_i be the necessity of the ε_i calculated as a difference of two preceding Hotelling's statistics;

Done=FALSE;

Repeat

for $i=1$ **to** p **do**

$$U_i = T^2(\varepsilon_1, \dots, \varepsilon_p) - T^2(\varepsilon_1, \dots, \varepsilon_{i-1}, \varepsilon_{i+1}, \dots, \varepsilon_p)$$

end

$$j = \arg \min_{i=1..p} (U_i)$$

$$\tilde{F} = (n + m - p - 1) \cdot \frac{U_i}{1 + T^2(\varepsilon_1, \dots, \varepsilon_p) - U_j}$$

If $\tilde{F} < F_{1, n+m-p-1, \alpha}$ **then**

 Remove the j^{th} row of the matrix \mathbf{J} ;

 Remove ε_j from the set FM ;

$p := p - 1$;

else

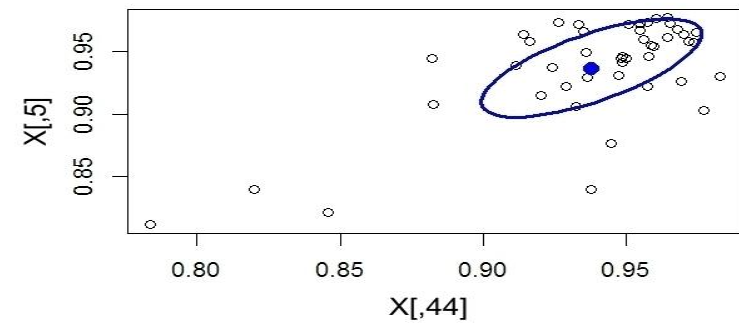
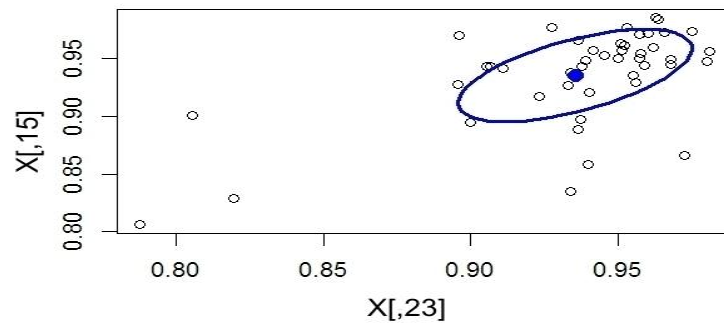
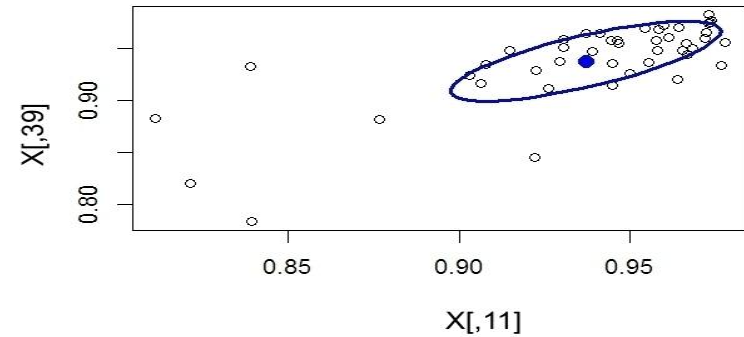
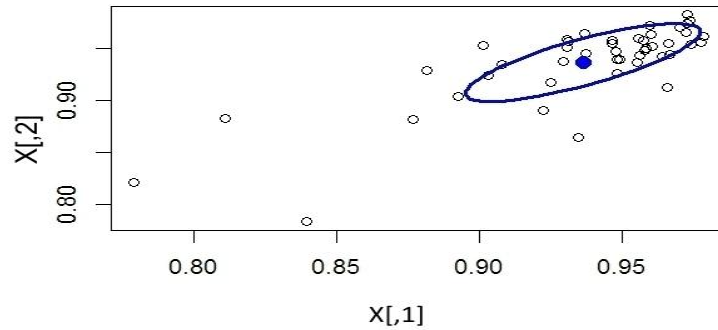
 Done= TRUE;

end

Until Done;

Czy nowe dane są prawidłowe ?

- Tak, bo:

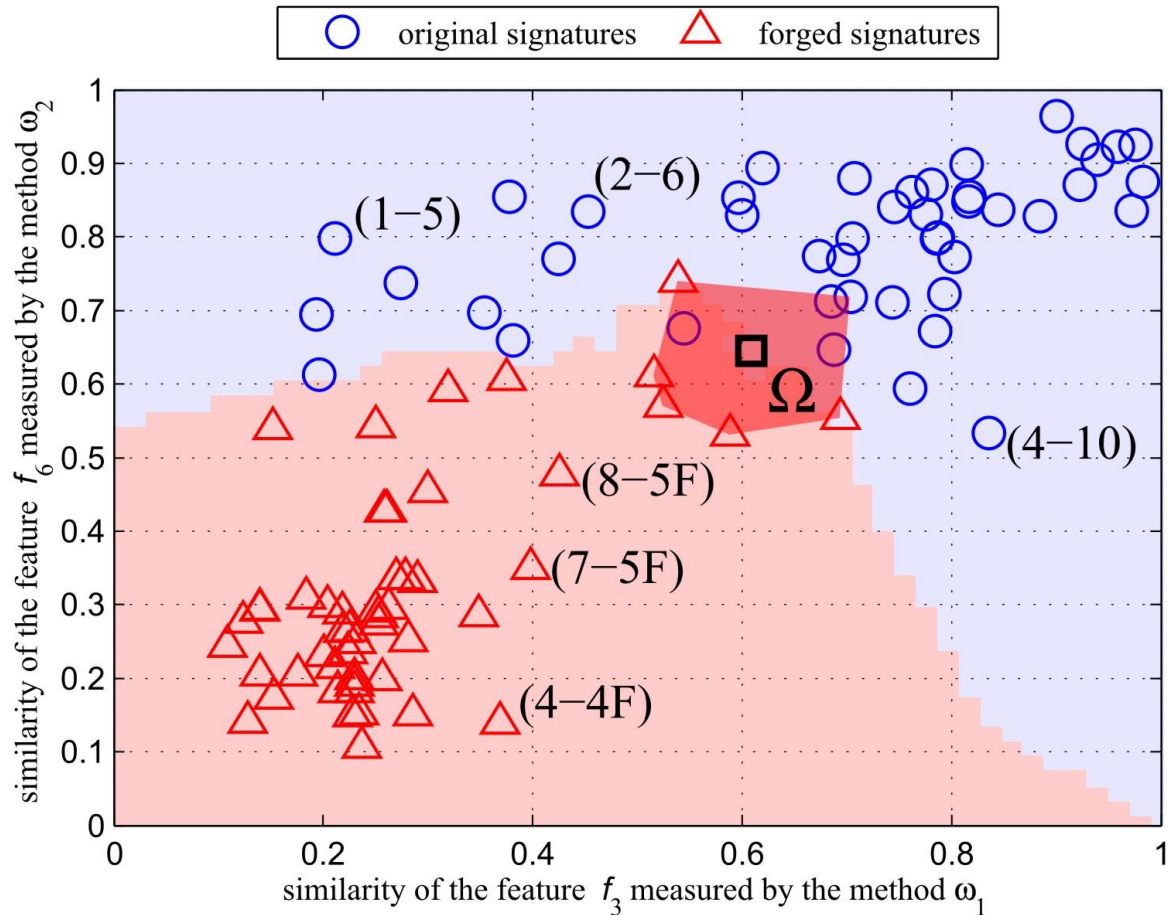


Bazy danych

- SVC 2004 <http://www.cse.ust.hk/svc2004>
- MCYT <http://atvs.ii.uam.es/mcytscores.html>
- SigComp2011 http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List
- 4NSigComp2012 http://www.iapr-tc11.org/mediawiki/index.php/Datasets_List
- SigWiComp2013 <http://www.dfki.de/afha/2013/SigWiComp.html>
- Baza własna <http://biometrics.us.edu.pl>

Bazy zawierają podpisy oryginalne i sfałszowane.
W eksperymencie badano 1600 podpisów z podziałem
na podpisy uczące i testowe

Testy praktyczne, (k -NN), Hotelling



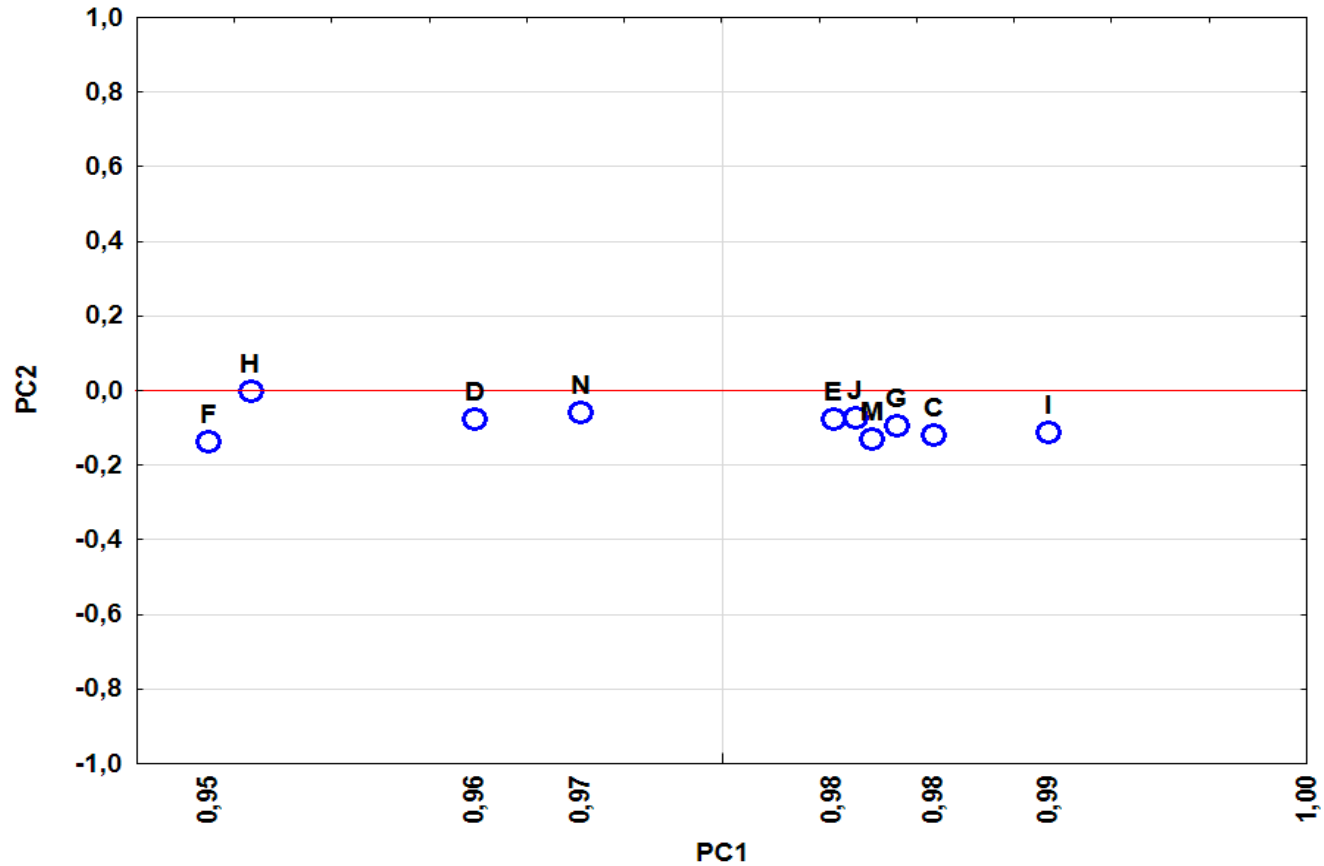
Testy praktyczne, PCA

TABLE 3
The most important factor loadings (absolute values below 0.1 were omitted).

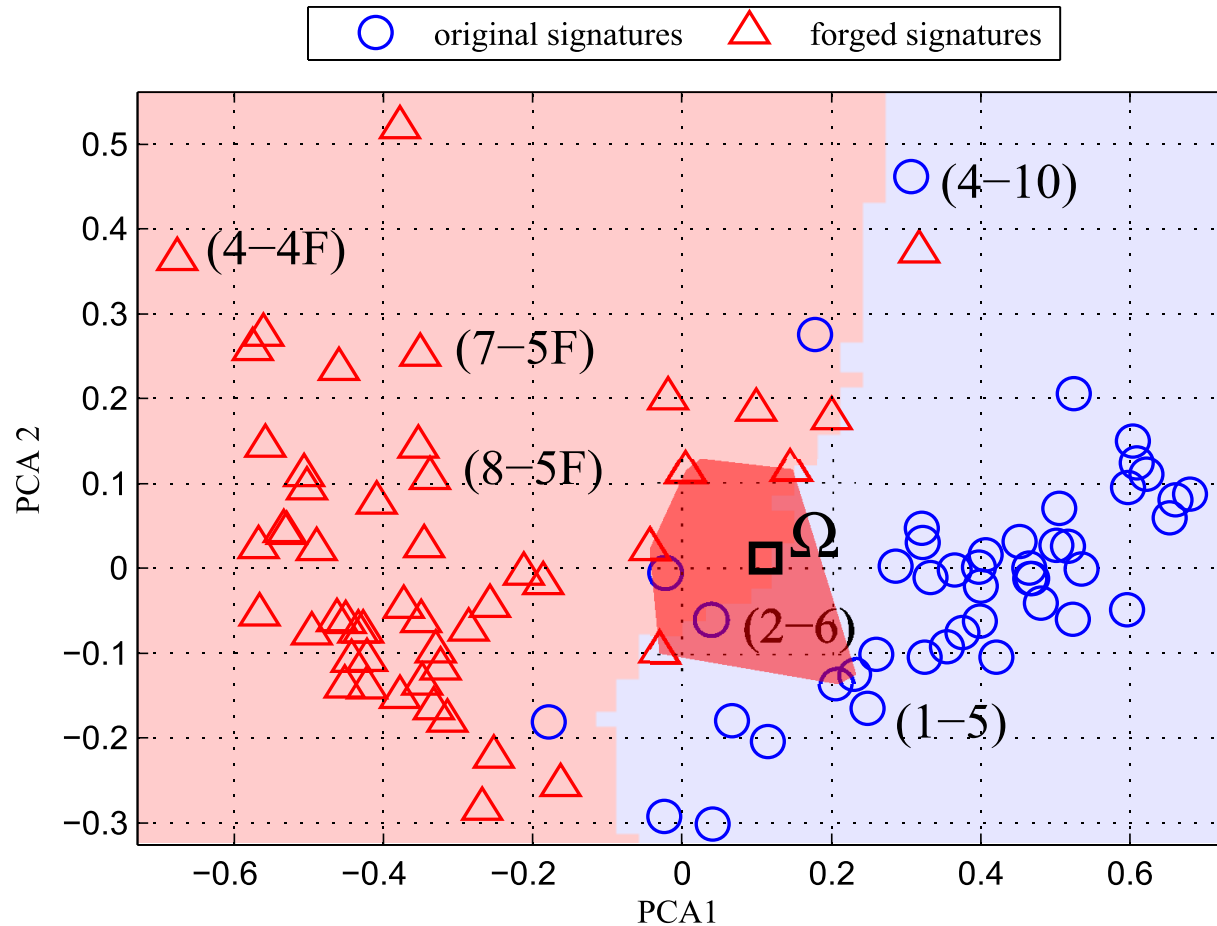
Feature-method ^{a)}	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7
A- (f_6, ω_1)	0.887	-0.102	0.359	-0.260			
B- (f_6, ω_2)	0.890	0.119	0.359	0.172		-0.149	
C- (f_7, ω_1)	0.984	-0.116					
D- (f_7, ω_2)	0.964		-0.183				
E- (f_1, ω_1)	0.980		-0.106	-0.103			
F- (f_1, ω_2)	0.953	-0.134	-0.198				
G- (f_3, ω_1)	0.982					0.104	
H- (f_3, ω_2)	0.955			0.265			
I- (f_4, ω_1)	0.989	-0.111					
J- (f_4, ω_2)	0.981						0.129
K- (f_5, ω_1)	0.861	0.393	-0.125		0.282		
L- (f_5, ω_2)	0.665	0.719			-0.154		
M- (f_2, ω_1)	0.981	0.127	-0.037			0.105	
N- (f_2, ω_2)	0.969		0.129			0.138	
Explained variance [%]	87.78	5.90	2.77	1.50	0.97	0.71	0.37

^{a)} The letters A...N represent corresponding "feature-method" pairs

Testy praktyczne, PCA



Testy praktyczne (k -NN), PCA



Jakie klasyfikatory stosować ?

Pytania dodatkowe. Dla nowego typu danych:
czy jakość klasyfikacji jest podobna kiedy klasyfikator pracuje na:
-danych zredukowanych,
-danych nie zredukowanych, (to może być duży zbiór!)
-danych surowych (pierwotne cechy, bez przetworzenia)

Aby odpowiedzieć na te pytania trzeba przeprowadzić odpowiednie eksperymenty!

$$Accuracy = \frac{\textit{liczba poprawnie klasyfikowanych}}{\textit{liczba wszystkich prób}}$$

Zastosowane klasyfikatory

- *k*-Nearest Neighbours Classifier (*k*-NN)
- Probabilistic Neural Network (PNN) and PSO method,
- Random Forests - forest of random trees (RanF),
- Random Tree - tree that considers *K* randomly chosen attributes at each node (RanT),
- J48 - C4.5 decision tree (J48),
- PART - PART decision list (PART),
- Naive Bayes classifier using estimator classifiers,
- RIDOR (RDR) Ripple-Down Rule learner (implementation the RDR technique in the WEKA package),

Wyniki klasyfikacji

TABELA 1

False Accepted Rate (FAR)

Różna liczba podpisów prawdziwych/falszywych oraz różne liczby wymiaru wektora cech.

Liczba podpisów w zbiorze:		FAR [%]						
$\pi_1 \cup \pi_2$		Wymiar wektora cech w metodzie PCA (wybór ręczny)			Wymiar wektora cech w metodzie SVD (wybór ręczny)			
π_1 (<i>genuine</i>)	π_2 (<i>forged</i>)	FMS*	2	6	12	2	6	12
3	1	6.34	8.14	6.32	6.75	6.14	7.12	5.62
5	3	1.08	2.22	2.62	2.21	2.52	2.12	2.25
10	4	1.67	3.14	3.44	3.29	3.15	3.21	3.94

*) (F)eature-(S)imilarity (M)etod (FSM) i klasyfikator k -NN

Wyniki klasyfikacji

TABELA 2
False Rejection Rate (FRR)

Różna liczba podpisów prawdziwych/falszywych oraz różne liczby wymiaru wektora cech.

Liczba podpisów w zbiorze: $\pi_1 \cup \pi_2$		FRR [%]						
π_1 (<i>genuine</i>)	π_2 (<i>forged</i>)	FMS*	Wymiar wektora cech w metodzie PCA (wybór ręczny)			Wymiar wektora cech w metodzie SVD (wybór ręczny)		
			2	6	12	2	6	12
3	1	7.14	5.31	6.55	5.40	6.33	5.60	4.90
5	3	2.53	2.58	2.78	4.94	4.35	4.78	4.05
10	4	2.60	2.83	3.06	5.25	3.96	4.84	4.60

*) (F)eature-(S)imilarity (M)ethod (FSM) i klasyfikator k -NN

Wyniki klasyfikacji

Użyte narzędzia Matlab, R, KNIME

TABELA 3

Klasyfikator	Nie zredukowany zbiór danych			Zredukowany zbiór danych		
	FAR [%]	FRR [%]	Accuracy [%]	FAR [%]	FRR [%]	Accuracy [%]
<i>k</i> -NN	54.25	0.00	72.87	1.08	2.53	97.92
PNN+PSO	14.65	2.13	83.21	0.11	0.53	99.34
RanF	22.00	2.75	87.62	5.75	0.00	97.12
RanT	40.25	19.75	70.00	18.75	2.00	89.62
J48	71.25	19.25	54.75	62.25	5.75	66.00
PART	71.25	19.25	54.75	62.25	5.75	66.00
NBayes	13.50	0.50	93.00	3.50	0.00	98.25
RIDOR	31.75	25.75	71.25	15.75	0.00	92.12

Wyniki klasyfikacji

TABELA 4
FAR, FRR and Accuracy
dla różnych klasyfikatorów pracujących na surowych danych

Klasyfikator	FAR [%]	FRR [%]	Accuracy [%]
<i>k</i> -NN	25.22	23.92	75.83
PNN+PSO	12,46	17.85	82,02
RanF	23.45	18.47	80.25
RanT	32.50	24.79	73.66
J48	34.29	20.16	75.75
PART	35.26	20.60	75.16
NBayes	14.21	27.11	81.58
Ridor	29.27	24.84	74.33

Porównanie z innymi (przybliżona ocena)

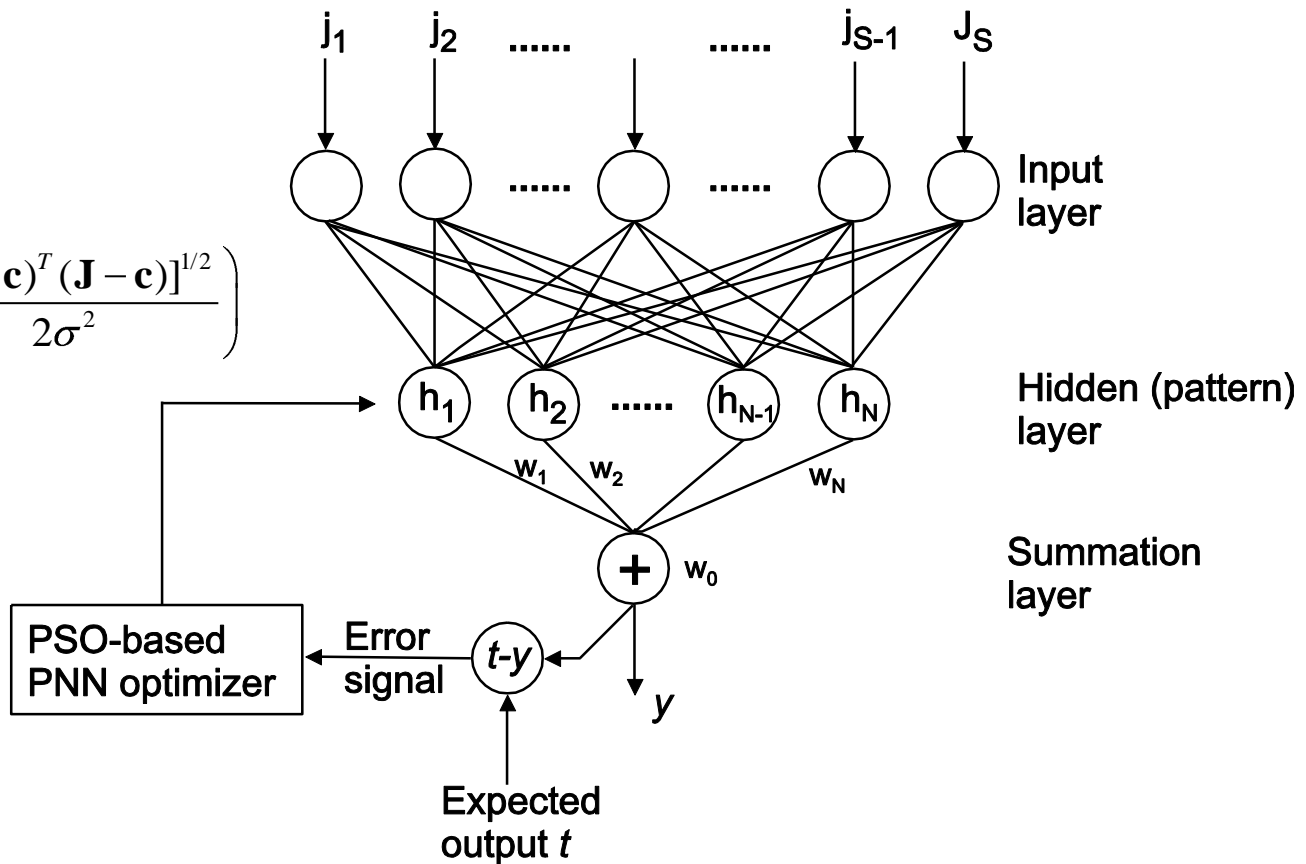
TABELA 5
PORÓWNANIE Z INNYMI ROZWIĄZANAMI ¹⁾

The approach	FAR	FRR	Signature recognition system	
			<i>off-line</i>	<i>on-line</i>
Proponowane rozwiązanie PNN+PSO	0.11	0.53		+
Exterior Contours and Shape Features	6.90	6.50	+	
HMM and Graphometric Features	23.00	1.00	+	
Virtual Support Vector Machine	13.00	16.00	+	
Genetic Algorithm	1.80	8.51	+	
Variable Length Segmentation and HMM	4.00	12.00		+
Dynamic Feature of Pressure	6.80	10.80		+
Consistency Functions	1.00	7.00		+
On line SRS - Digitizer Tablet	1.10	3.09		+

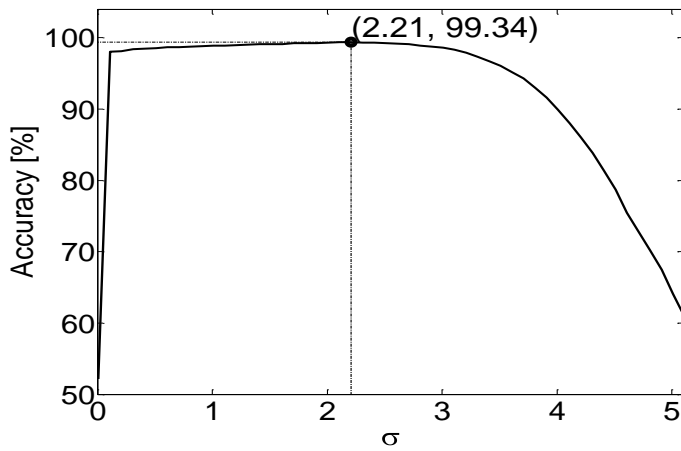
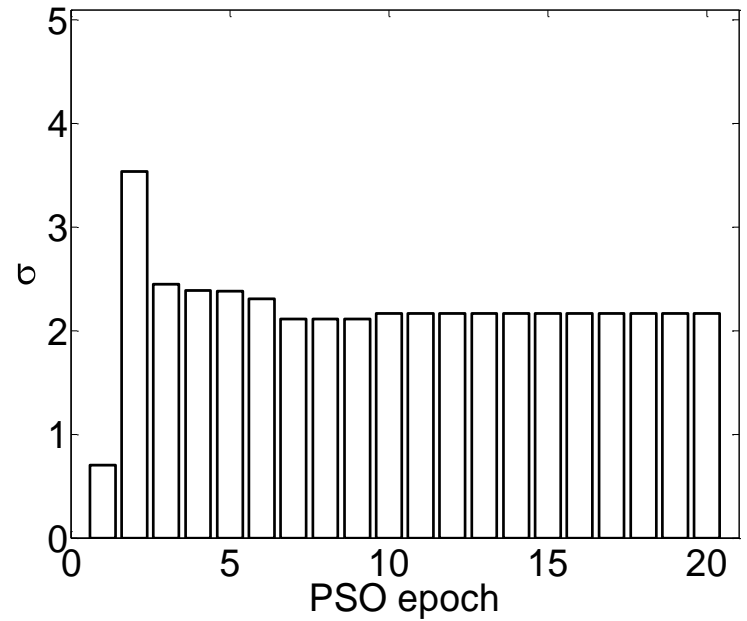
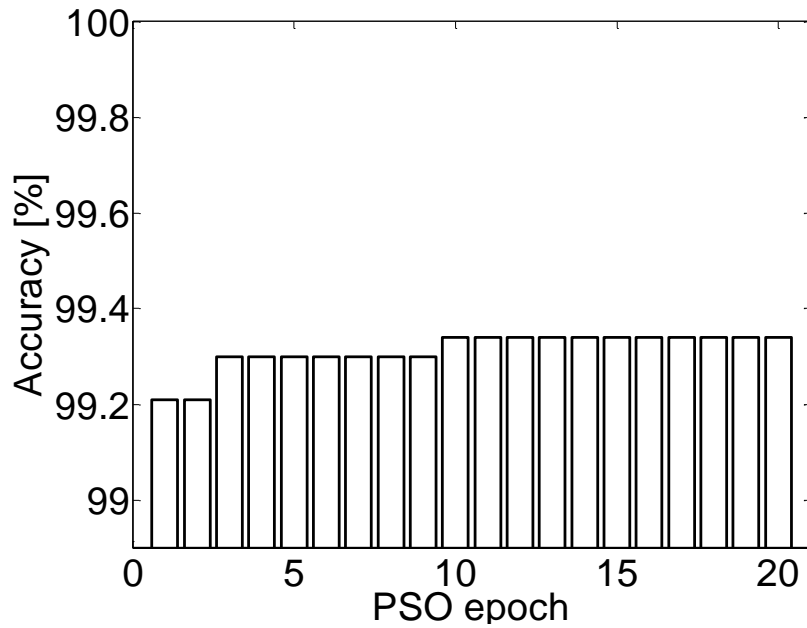
1) D. Impedovo, G. Pirlo, Automatic signature verification: the state of the art. IEEE Trans. on Syst. Man. and Cybernetics. Part C: Applications and Reviews, vol. 38v no. 5, 2013, pp. 609-635.

PNN+PSO

$$G(\mathbf{J}, \mathbf{c}) = \exp\left(-\frac{[(\mathbf{J} - \mathbf{c})^T (\mathbf{J} - \mathbf{c})]^{1/2}}{2\sigma^2}\right)$$



Trening sieci





Dziękuję za uwagę