Secondary Structure Prediction with Support Vector Machines

Ward, McGuffin, Buxton, Jones

Bioinformatics, 2003

Súhrn

- Cieľ štúdie: vývoj spoľahlivej metódy pre problém predikcie sekundárnej štruktúry proteínov
- Použitá metóda: binárny klasifikátor SVM
- Dosiahnuté výsledky: 77%-ná úspešnosť, zrovnateľná s "state-of-the-art" PSIPRED

Predikcia sekundárnej štruktúry

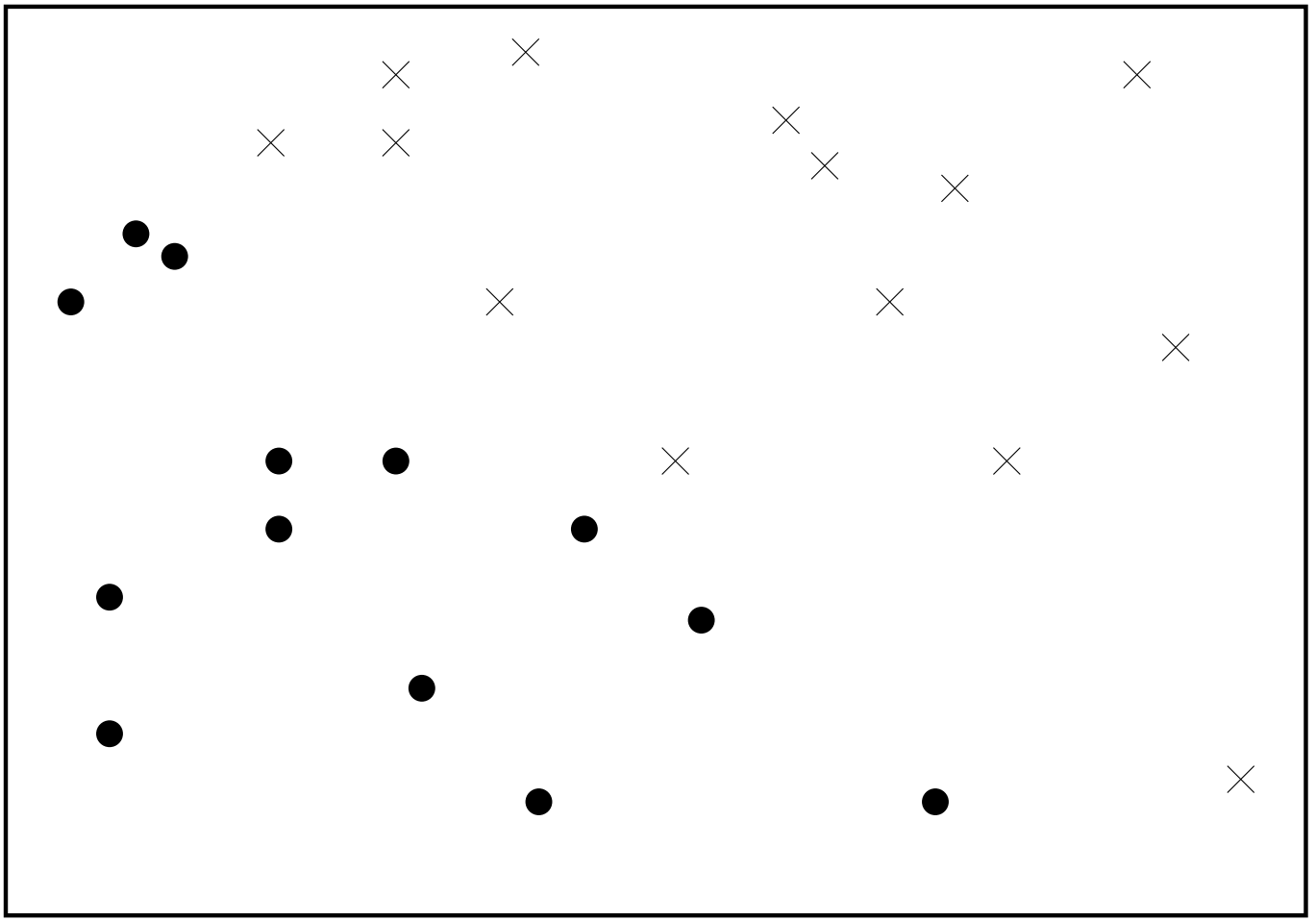
- Viacvrstvé neurónové siete (PSIPRED, PROFsec)

Aplikácie SVM

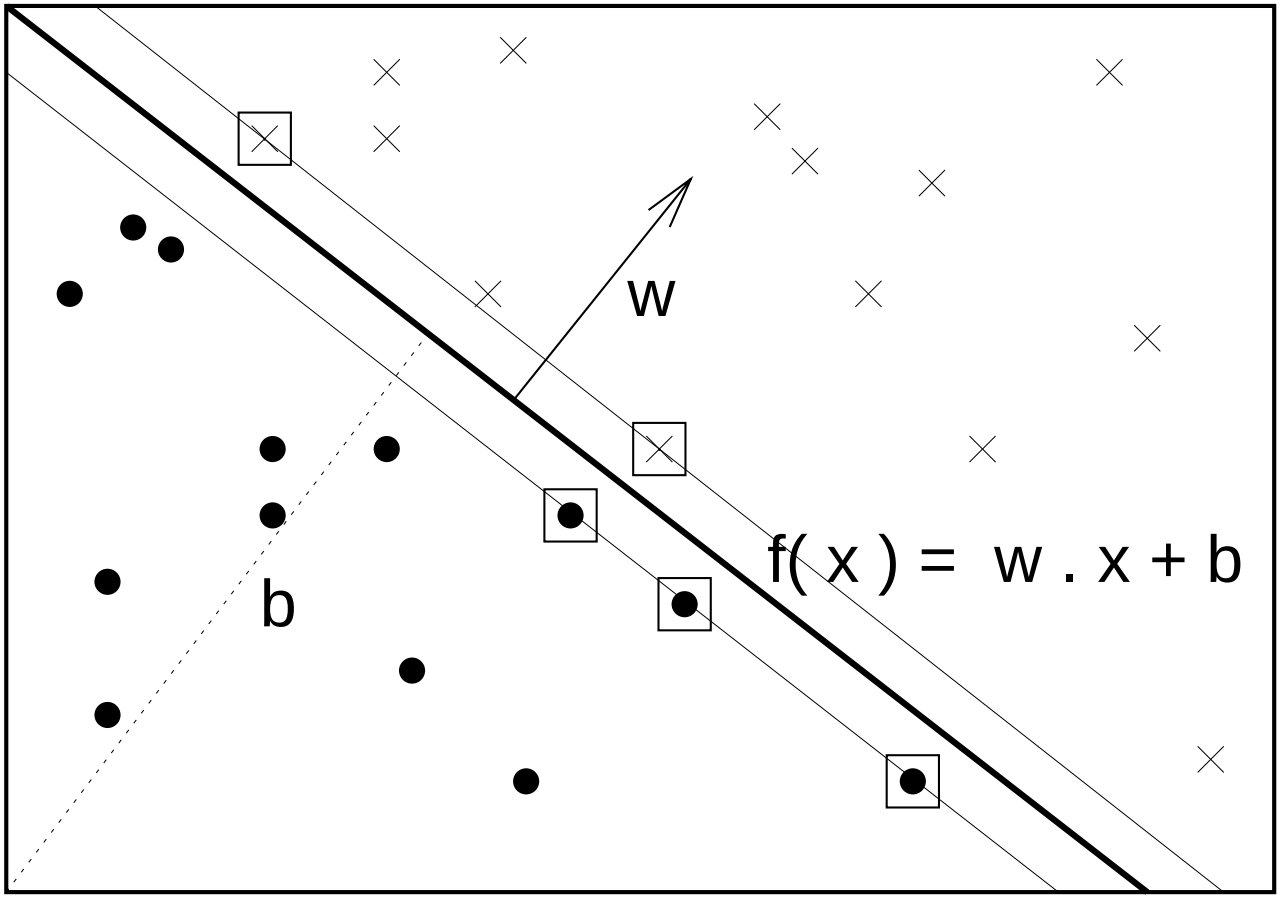
- Rozpoznávanie proteínovej translácie
- Anotácia funkcií génov

Support Vector Machines

- Vynašiel V. Vapnik v 1992
- Založené na štatistickej *VC teórii* (Vapnik-Chervonenkova teória)
- Dobre preskúmaná teoretická časť
- Všeobecný klasifikátor



Obr. 1: 2 triedy objektov

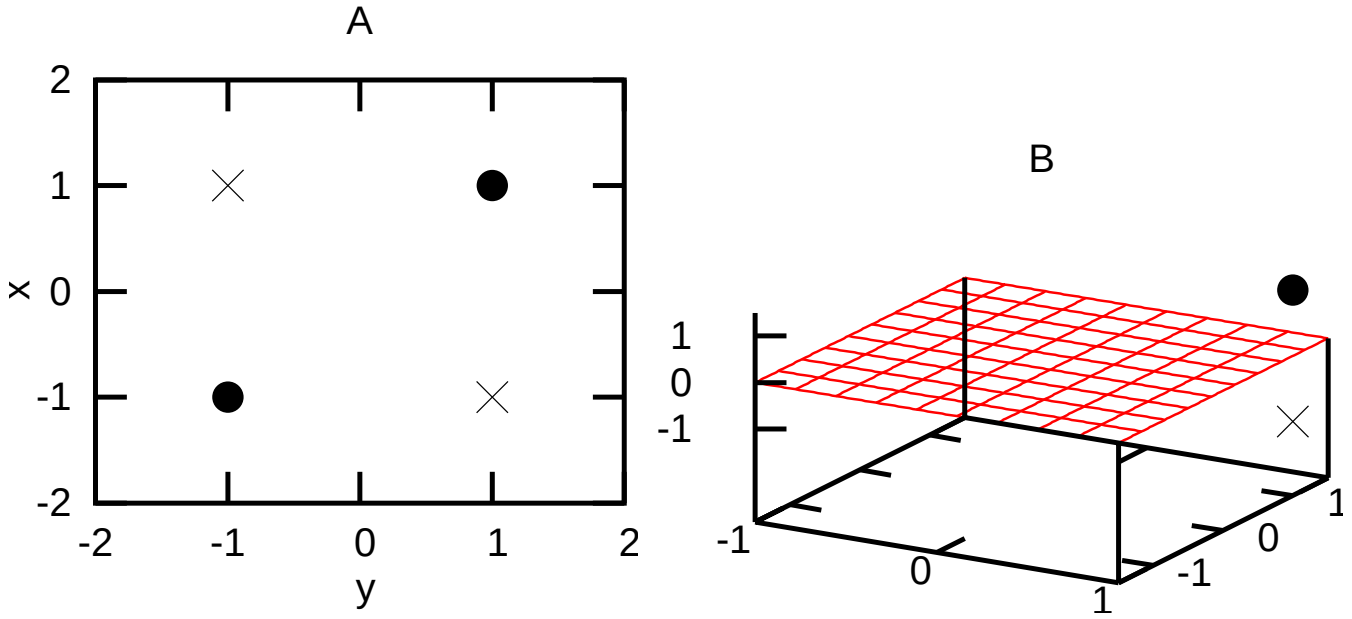


- SVM $f(\vec{x}) = \vec{w}.\vec{x} + b$
  - $f(\vec{x}) = \pm 1$
  - $\vec{x}$, $\vec{w}$, b

Linearita

- SVM dokáže pracovať len s lineárnymi problémami
- Skoro všetky problémy sú nelinárne
- Jadrové funkcie (Kernel functions)
  - nie *každá* funkcia môže byť jadrovou!

XOR problém:

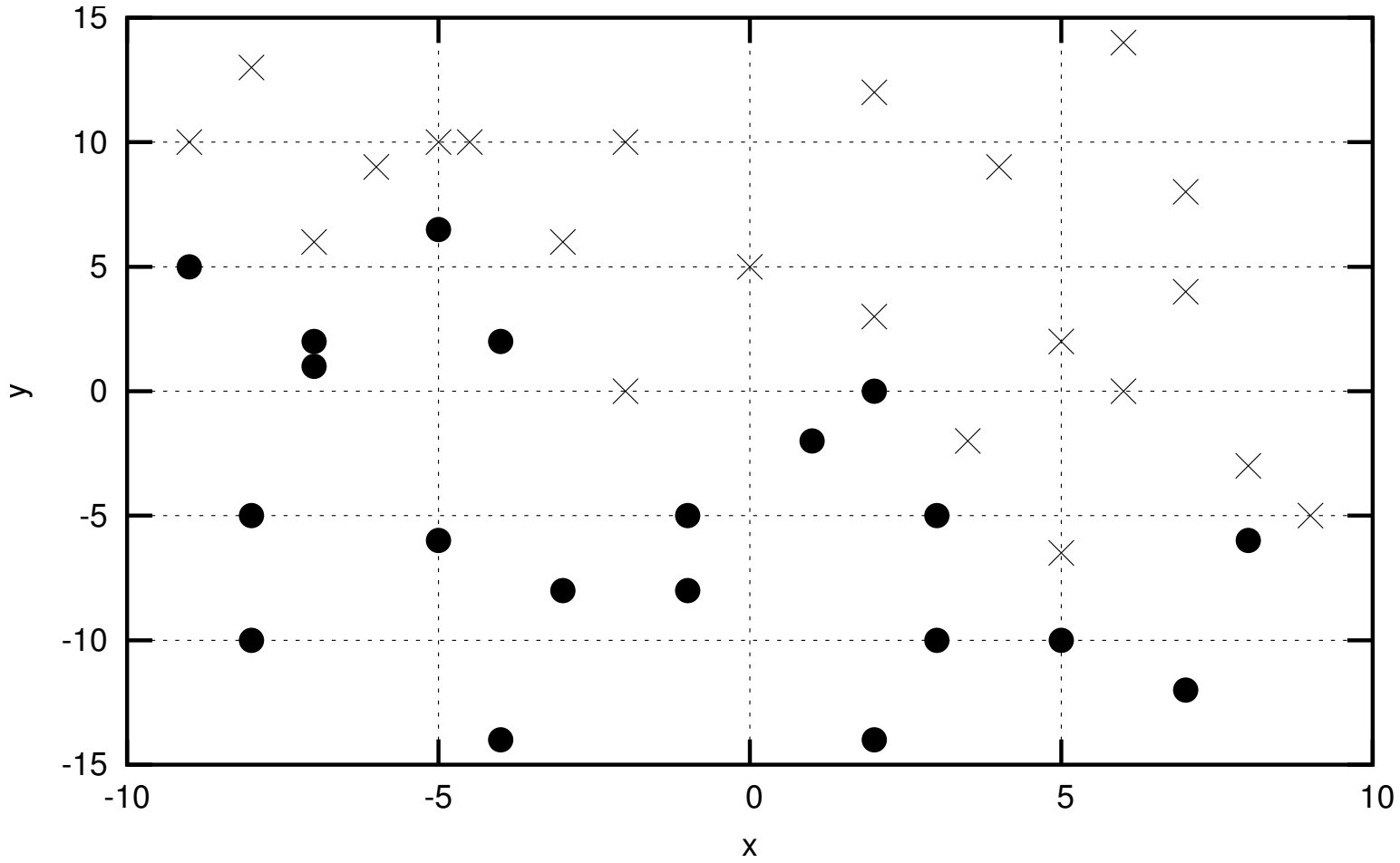


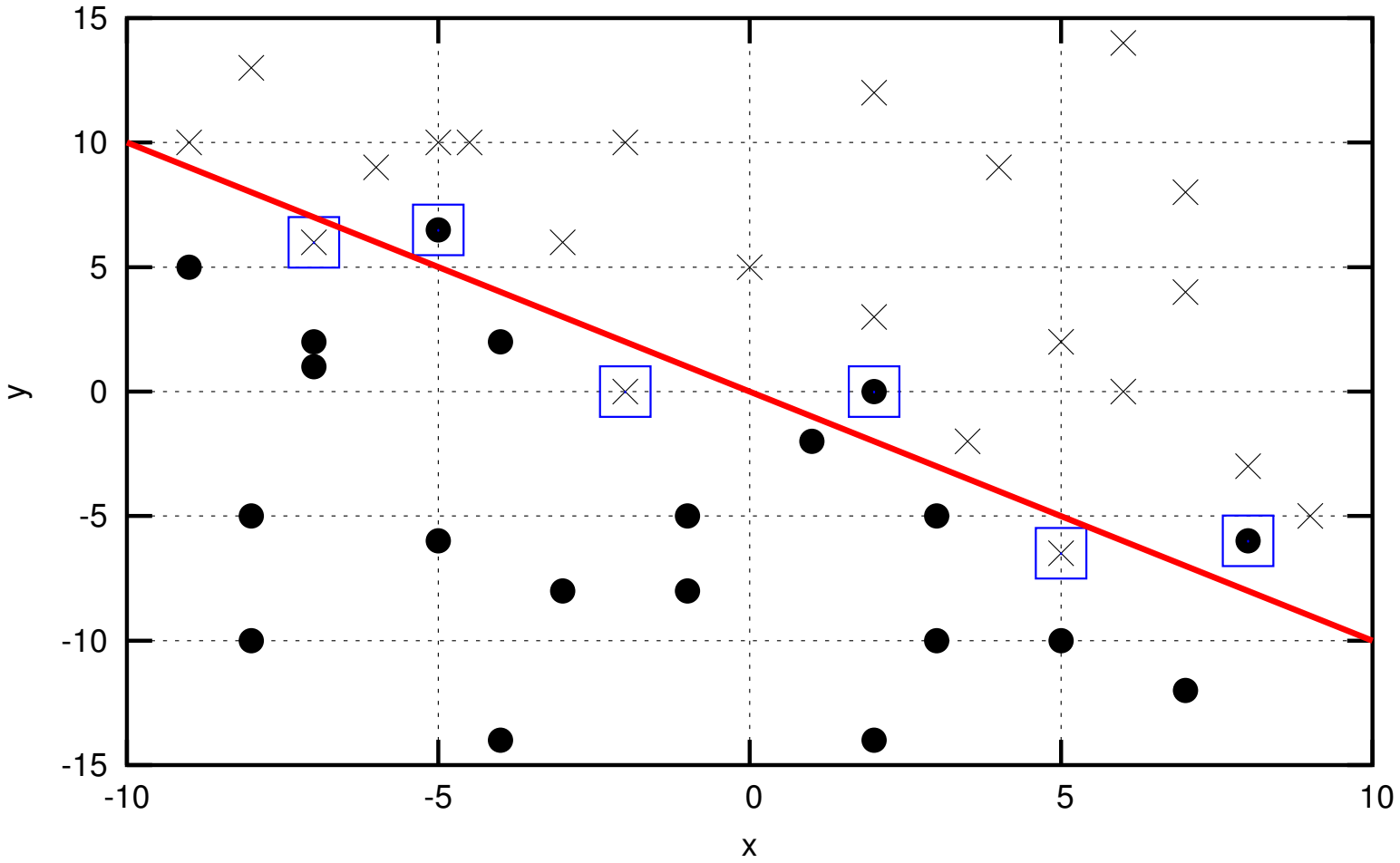
Použitá polynomiálna jadrová funkcia:

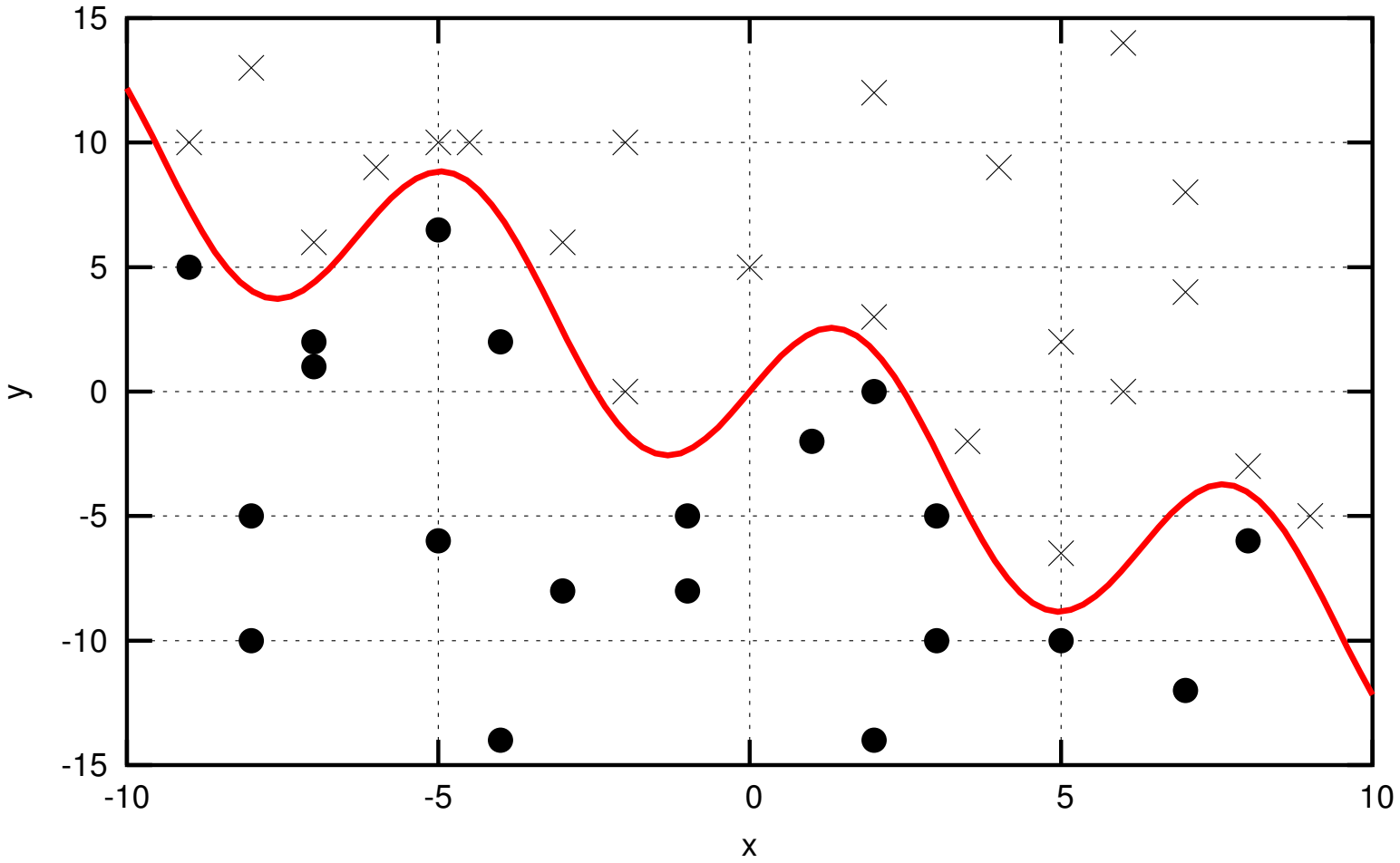
$$k(\vec{u},\vec{v}) = (\vec{u},\vec{v})^2 = \Big(u_1 \;.\; v_1 + u_2 \;.\; v_2\Big)^2 = \ldots =$$

$$\Big(\Big(u_1^2, u_2^2, \sqrt{2} \;.\; u_1 \;.\; u_2\Big) \;.\; \Big(v1^2, v_2^2, \sqrt{2} \;.\; v_1 \;.\; v_2\Big)\Big)$$

Príklad




15
10
5
0
-5
-10
-15
y
-10
-5
0
5
10
x


15
10
5
0
-5
-10
-15
y
-10
-5
0
5
10
x

Proteínové dáta

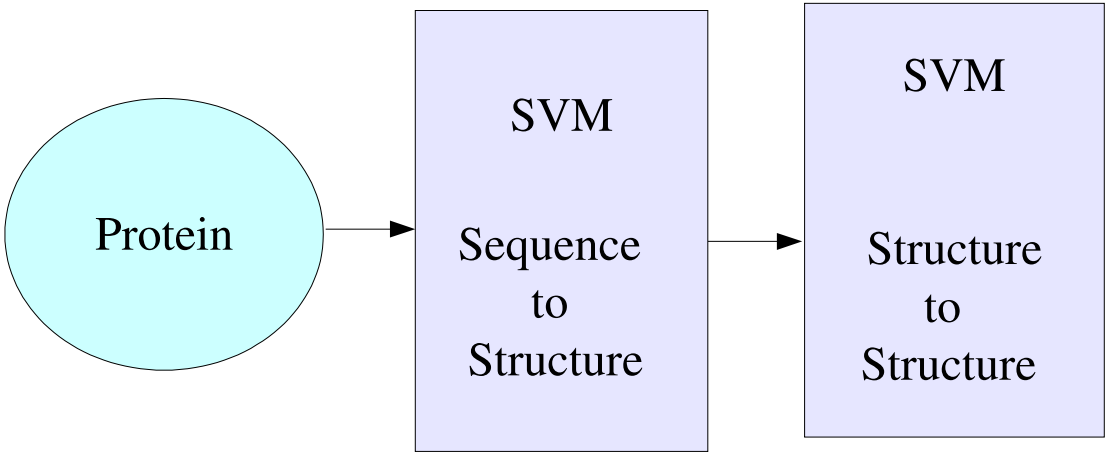
- Klastrovanie – max. 25% podobnosť medzi klastrami (zhlukami)
- Reprezentatívna vzorka obsahovala 1460 proteínov
- 10 násobné krížové overovanie (10-fold cross-validation)

Predspracovanie

- Každý proteín rozdelený na úseky dĺžky 15 (aminokyselín)
- SVM ako vstup dostane maticu 20x15
- 20 – počet možných aminokyselín
- Triedy proteínov:
  - H – α-helix, 3_{10}-helix
  - E – sheet, isolated β-bridge
  - C – π-helix, turn, bend a ostatné

Multiklasifikácia pomocou binárneho SVM

- Jeden proti všetkým (HxEC, ExHC, CxHE)
- Jeden proti jednému (HxE, HxC, ExC)
- + pravdepodobnostná modifikácia predchádzajúcich metód
- Výsledná hodnota získaná hlasovaním komisie klasifikátorov


Protein
SVM
Sequence
to
Structure
SVM
Structure
to
Structure

Sequence-to-Structure

(jednotlivé klasifikátory)

Trieda A	**Trieda B**	**#SV**	**Presnosť (%)**
C	H, E	55,0	77,7
H	C, E	40,9	86,4
E	H, C	36,5	85,6
C	H	46,1	84,2
C	E	48,5	81,3
H	E	36.0	88.0

Výsledky

Klasifikátor	**Presnosť (%)**
Sequence-to-Structure	
Jeden proti všetkým	74,54
Modifikácia MaxProb SVM	74,52
Jeden proti jednému	74,04
Modifikácia Pair-wise couple SVM	74,24
PSIPRED	73,28
Structure-to-Structure	
Pair-wise couple SVN	75,44
PSIPRED	74,72

Zvýšenie presnosti

- Z dátovej množiny 1460 proteínov odstránené kolízne reťazcové zlomy (unresolved chain breaks)
- Upravené dáta – 1095 proteínov
- 3-zložková krížová validácia — **77,07%**

Komisia rôznych klasifikátorov

- Výsledky na 121 nových proteínoch

Klasifikátor	**Presnosť (%)**
SVM	74,92
PSIPRED	74,97
PROFsec	74,90
Komisia	76,17

Záver

- Výhody SVM:
  - porovnateľné výsledky ako PSIPRED
  - menšia trénovacia množina (1460) oproti PSIPRED (5000)
- Nevýhody SVM:
  - väčšie požiadavky na výkon a pamäť počítača