Support vector machines and Applications in bioinformatics

Jean-Philippe.Vert@mines.org

Ecole des Mines de Paris
Computational Biology group

Ecole des Mines de Paris

- 1770 persons (250 academics, 400 PhD students, 670 undergraduates/M.S.)
- Excellent formation (undergraduate and graduate)
- 19 research centers (earth science, energy, mechanics, applied maths, economics)
- 21.5 Million euros of research contracts (through Armines)
Computational biology at the Ecole des Mines de Paris

- Expertise in statistics, machine learning, data mining...
- Projects: functional genomics, learning from heterogeneous data, virtual screening of chemical compounds, microarray data and pathway analysis...
Overview

1. Pattern recognition and Support Vector Machines
2. Signal peptide detection
3. Virtual screening of small molecules
4. Analysis of microarray data with pathways information
Partie 1

Pattern recognition
and
Support Vector Machines
The pattern recognition problem
The pattern recognition problem

- Learn from labelled examples a discrimination rule
The pattern recognition problem

- Learn from labelled examples a discrimination rule
- Use it to predict the class of new points
Pattern recognition examples

- Medical diagnosis (e.g., from microarrays)
- Drugability/activity of chemical compounds
- Gene function, structure, localization
- Protein interactions
Support Vector Machines for pattern recognition
Support Vector Machines for pattern recognition

- Object $x$ represented by the vector $\Phi(\vec{x})$ (feature space)
Support Vector Machines for pattern recognition

- Object $x$ represented by the vector $\Phi(x)$ (feature space)
- Linear separation with large margin in the feature space
Support Vector Machines for pattern recognition

- Object $x$ represented by the vector $\Phi(\vec{x})$ (feature space)
- Linear separation with large margin in the feature space
The kernel trick for SVM

- The separation can be found without knowing $\Phi(x)$. Only the following kernel matters:

$$K(x, y) = \Phi(x) \cdot \Phi(y)$$

- Simple kernels $K(x, y)$ can correspond to complex $\Phi$

- SVM work with any sort of data as soon as a kernel is defined
Kernels

• A kernel can be thought of as a measure of similarity.

• There are mathematical conditions to ensure that a function \( K(x, y) \) is a valid kernel (it must be symmetric positive semidefinite).

• As soon as \( K(., .) \) is a valid kernel, SVM can be used for pattern recognition.
Advantages of SVM

- Works well on real-world applications
- Large dimensions, noise OK
- Can be applied to any kind of data as soon as a kernel is available
Partie 2

Signal peptide cleavage site detection
Secretory pathway

- Nascent protein
- ER
- Golgi
- mRNA
- Signal peptide
- Cell surface (secreted)
- Lysosome
- Plasma membrane
- Nucleus
- Chloroplast
- Mitochondrion
- Peroxisome
- Cytosole
## Signal peptides

<table>
<thead>
<tr>
<th>Protein</th>
<th>-1</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>MKANAKTIIGMALAISHTAM</td>
<td>EE...</td>
</tr>
<tr>
<td>(2)</td>
<td>MKQSTIALALLPLLFTVPVKA</td>
<td>RT...</td>
</tr>
<tr>
<td>(3)</td>
<td>MKATKLVLGAVILGSTLLAG</td>
<td>CS...</td>
</tr>
</tbody>
</table>

## Signal peptides

<table>
<thead>
<tr>
<th>Protein</th>
<th>Sequence</th>
<th>-1</th>
<th>+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>MKANAKTIAGMIALAISHTAMA</td>
<td>EE...</td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>MKQSTIALALLPLLFTPVTKA</td>
<td>RT...</td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>MKATKLVLGAVILGSTLLAG</td>
<td>CS...</td>
<td></td>
</tr>
</tbody>
</table>


- 6-12 hydrophobic residues (in yellow)
- (-3,-1): small uncharged residues
The classification problem(s)

- Problem 1:
  Given an aminoacids windows:

  \[ [x_{-8}, x_{-7}, \ldots, x_{-1}, x_1, x_2] = \text{ILGSTLLACS} \]

  is there a cleavage site between \( x_{-1} \) and \( x_1 \)?
The classification problem(s)

- **Problem 1**: Given an aminoacids windows:
  \[
  [x_{-8}, x_{-7}, \ldots, x_{-1}, x_1, x_2] = \text{ILGSTLLACS}
  \]
  is there a cleavage site between \(x_{-1}\) and \(x_1\) ?

- **Problem 2**: Given an protein sequence, **does it contain a signal peptide?**
Current methods : Problem 1

- **Weight matrix method**: compute the score of a window by:

\[ s(ILGSTLLACS) = s_{-8}(I) + s_{-7}(L) + \ldots + s_{2}(S) \]

where \( s_i \) have been trained from example to discriminate between windows with or without cleavage site (Von Heijne)

- **Neural networks** (Brunak et al.)
SVM approach (PSB 2002)

- We need a kernel \( K(w_1, w_2) \) between 2 windows
- It is possible to transform a weight matrix into a kernel (technical, see paper)
- Experiment: 1,418 positive examples, 65,216 negative examples, cross-validation
Result: ROC curves

Interpolated Kernel

Product Kernel (Bayes)
Remarks

- The weight matrix is used to define the geometry of the feature space (through the kernel)

- The SVM algorithm learns a linear discrimination in this space
Problem 2: signal peptide detection

- Classical approach: move a window along the sequence, check whether it looks like a typical signal peptide

- SVM approach: we need a string kernel $K(p_1, p_2)$ for variable-length protein sequences

- String kernel examples: Fisher kernel (Jaakkola et al. 99), spectrum and mismatch kernels (Leslie et al. 02), local alignment kernel (Vert et al. 03)
Local alignment kernel

- For two strings $x$ and $y$, a local alignment $\pi$ with gaps is:

  $\begin{align*}
  \text{ABCD} & \quad \text{EF} \quad \text{G} \quad \text{HI} \quad \text{JKL} \\
  \text{MNO} & \quad \text{EFP} \quad \text{ORGS} \quad \text{I} \quad \text{TUVWX}
  \end{align*}$

- The score is:

  $$s(x, y, \pi) = s(E, E) + s(F, F) + s(G, G) + s(I, I) - s(\text{gaps})$$
Smith-Waterman (SW) score

\[ SW(x, y) = \max_{\pi \in \Pi(x, y)} s(x, y, \pi) \]

- This is not a kernel in general
- But the following is a valid kernel:

\[ K_{LA}^{(\beta)}(x, y) = \sum_{\pi \in \Pi(x, y)} \exp(\beta s(x, y, \pi)), \]
SCOP superfamily recognition benchmark
Partie 3

Virtual screening of small molecules
The problem

- **Objects** = chemical compounds (formula, structure..)

- **Problem** = predict their:
  - drugability
  - pharmacocinetetics
  - activity on a target etc...
Classical approaches

- Use molecular descriptors to represent the compounds as vectors
- Select a limited numbers of relevant descriptors
- Use linear regression, NN, nearest neighbour etc...
SVM approach

- We need a kernel $K(c_1, c_2)$ between compounds
SVM approach

- We need a kernel $K(c_1, c_2)$ between compounds
- One solution: inner product between vectors
SVM approach

- We need a kernel $K(c_1, c_2)$ between compounds
- One solution: inner product between vectors
- Alternative solution: define a kernel directly using graph comparison tools
Example: graph kernel

\[
\begin{align*}
&\text{N} & \text{C} & \text{C} & \text{C} & \text{C} & \text{C} & \text{O} \\
& & \text{C} & & & & & \text{O} \\
& & & & & \text{N} & & \\
\end{align*}
\]
Example: graph kernel

Extract random paths
Example: graph kernel

Extract random paths
Example: graph kernel

- Let $H_1$ be a random path of a compound $c_1$
- Let $H_2$ be a random path of a compound $c_2$
- The following is a valid kernel:

$$K(c_1, c_2) = \text{Prob}(H_1 = H_2).$$
Remarks

- Interesting preliminary results in mutagenesis prediction (benchmark dataset)
Remarks

- Interesting preliminary results in mutagenesis prediction (benchmark dataset)
- Two compounds are compared in terms of their common substructures
Remarks

- Interesting preliminary results in mutagenesis prediction (benchmark dataset)
- Two compounds are compared in terms of their common substructures
- What about kernels for the 3D structure?
Partie 4

Analysis of microarray data with pathways information
Genes encode proteins which can catalyse chemical reactions

Nicotinamide Mononucleotide Adenylyltransferase With Bound Nad$^+$
Chemical reactions are often parts of pathways

From http://www.genome.ad.jp/kegg/pathway
Microarray technology monitors RNA quantity

(From Spellman et al., 1998)
Comparing gene expression and protein network

Gene network

Expression profiles

Are there “correlations”?
In yellow: a candidate pattern, and the correlation coefficient with each gene profile.
The correlation function with interesting patterns should vary smoothly on the graph.
Summary

Diffusion kernel

Linear kernel

Kernel CCA
Data

- **Gene network**: two genes are linked if they catalyze successive reactions in the KEGG database.

- **Expression profiles**: 18 time series measures for the 6,000 genes of yeast, during two cell cycles.
First pattern of expression
Related metabolic pathways

50 genes with highest $s_2 - s_1$ belong to:

- Oxidative phosphorylation (10 genes)
- Citrate cycle (7)
- Purine metabolism (6)
- Glycerolipid metabolism (6)
- Sulfur metabolism (5)
- Selenoaminoacid metabolism (4), etc...
Related genes
Related genes
Related genes
Opposite pattern
Related genes

- RNA polymerase (11 genes)
- Pyrimidine metabolism (10)
- Aminoacyl-tRNA biosynthesis (7)
- Urea cycle and metabolism of amino groups (3)
- Oxidative phosphorylation (3)
- ATP synthesis (3), etc...
Related genes

RNA polymerase II (Saccharomyces cerevisiae)
Related genes
Related genes

UREA CYCLE AND METABOLISM OF AMINO GROUPS

[Diagram showing the urea cycle and metabolism of amino groups with various enzymes and pathways labeled]
Extensions

- Can be used to **extract features** from expression profiles (preprint 2002)

- Can be generalized to **more than 2 datasets** and other kernels

- Can be used to extract **clusters of genes** (e.g., operon detection, *ISMB 03* with Y. Yamanishi, A. Nakaya and M. Kanehisa)
Conclusion
Conclusion

- Kernels offer a versatile framework to represent biological data
- SVM and kernel methods work well on real-life problems, in particular in high dimension and with noise
- Encouraging results on real-world applications
- Many opportunities in developing kernels for particular applications