Multi-Class Protein Fold Recognition using Large Margin Logic based Divide and Conquer Learning

Huma Lodhi[1]
Stephen Muggleton[1]
Mike J E Sternberg[2]

[1] Department of Computing, Imperial College London

[2] Centre for Bioinformatics, Imperial College London

28 June 2009

Introduction

- Introduction
- Support Vector Inductive Logic Programming
- Decision List based Support Vector Inductive Logic Programming
- Experiments and Results
- Conclusion

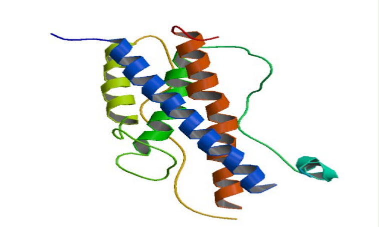
Protein Fold Recognition: Multi-class Learning Problem

Alphabet {A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V }
Protein: A finite sequence of characters from alphabet of 20 amino acids

Protein Folding

```
FPTIPLSRLFQNAMLRAHRLHQLAFDTYEE
FEEAYIPKEQKYSFLQAPQASLCFSESIPT
PSNREQAQQKSNLQLLRISLLLIQSWLEPV
GFLRSVFANSLVYGASDSDVYDLLKDLEEG
IQTLMGRLEDGSPRTGQAFKQTYAKFDANS
HNDDALLKNYGLLYCFRKDMDKVETFLRIV

QCRSVEGSCGF
```



Challenges

- Structured data
- Skewed class distribution

Distinguishing Characteristics: Kernel Methods & Inductive Logic Programming

Inductive Logic Programming (ILP)

- Ease of incorporation of background knowledge
- Expressive language formalism

Kernel Methods (KMs)

- High generalization ability
- Strong theoretical foundation

Problem

Methodologies for regression estimation and multi-class pattern classification

Logic based Kernel Learning

- Handles arbitrary type of data
- Methodology for regression estimation
- Algorithms for multi-class pattern classification

Support Vector Inductive Logic Programming (SVILP): An Instance of Logic based Kernel Learning

- At the intersection of Support Vector Machines and Inductive Logic Programming

Learning with SVILP

- A set of rules $\mathcal{H}$ is obtained from an ILP system, where a first order rule, $h \in \mathcal{H}$, can be viewed as a boolean function of the form, $h : D \rightarrow \{0, 1\}$
- A subset $H \in \mathcal{H}$ is selected

Feature Map

The subset of rules defines a mapping ϕ

$$\phi : d \rightarrow \left(\sqrt{\pi(h_1(d))}, \sqrt{\pi(h_2(d))}, \ldots, \sqrt{\pi(h_t(d))} \right)^T$$

- A kernel function is constructed by using the selected set of rules

SVILP Kernel

$k(d_i, d_j) = \langle \phi(d_i), \phi(d_j) \rangle = \sum_{l=1}^{t} \sqrt{\pi(h_l(d_i))} \sqrt{\pi(h_l(d_j))}$

Learning with SVILP

- Construct Gaussian RBF kernels in ILP space

$$k_{RBF}(d_i, d_j) = \exp\left(\frac{-\|(\phi(d_i) - \phi(d_j)\|^2}{2\sigma^{\,2}}\right)$$

$$\|(\phi(d_i) - \phi(d_j)\| = \sqrt{k(d_i, d_i) - 2k(d_i, d_j) + k(d_j, d_j)}$$

- Learning is performed by using an SVM in conjunction with the SVILP kernel.

Learning with SVILP: Low Dimensional Embedding

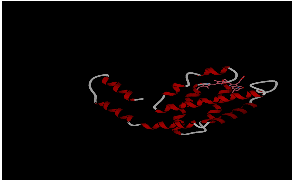
- For each rule compute goodness of fit by using compression

  $$C = \frac{PT * (ps - (ng + c))}{ps}$$

  $ps =$ number of positive examples correctly deducible from the rule
  $ng =$ number of negative examples that satisfy the conditions of the rules
  $c =$ length of the rule
  $PT =$ total number of positive examples
- Select k rules with positive compression values.

SVILP Kernel

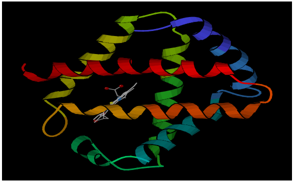
Protein domain 1alla



Relationally encoded features of protein domain 'd1alla_'.

```
dom_t(d1alla_).
len(d1alla_, 161). nb_alpha(d1alla_,7).
nb_beta(d1alla_,0). has_pro(d1alla_h1).
sec_struc(d1alla_, d1alla_h3).
unit_t(d1alla_h3).
sst(d1alla_h3,4,4,a,104,9,h,0.443,
3.003,116.199, [v,t,p,i,e,e,i,g,v]).
unit_hmom(d1alla_h2, hi).···
```

Protein domain 2hbg



Relational encoded features of protein domain 'd2hbg__'.

```
dom_t(d2hbg__).
len(d2hbg__, 147). nb_alpha(d2hbg__,6).
nb_beta(d2hbg__,0). has_pro(d2hbg__h5).
sec_struc(d2hbg__, d2hbg__h2).
unit_t(d2hbg__h2).
sst(d2hbg__h2,3,3,blank,40,7,h,0.540,
1.812, 213.564,
[q,m,a,a,v,f,g]). ···
```

SVILP Kernel

fold(Globinlike,A)	$\leftarrow$ adjacent(A,B,C,1,h,h), adjacent(A,C,D,2,h,h), coil(B,C,4). /*A domain is classified 1 (belongs to Fold 'Globinlike') if helices B(at position 1) and C are adjacent, C (at position 2) and D are adjacent and length of loop connecting B and C is 4.*/
fold(Globinlike,A)	$\leftarrow$ adjacent(A,B,C,1,h,h), has_pro(C). /*A domain is classified 1 if helices B(at position 1) and C are adjacent and C has proline.*/
fold('Globinlike',A)	$\leftarrow$ adjacent(A,B,C,1,h,h), coil(B,C,4), nb_α_interval(4=<(A=<8)). /*A domain is classified 1 if helices B (at position 1) and C are adjacent, number of α helices are in range [4,8] and length of loop connecting B and C is 4*/.

Feature Map and SVILP Kernel

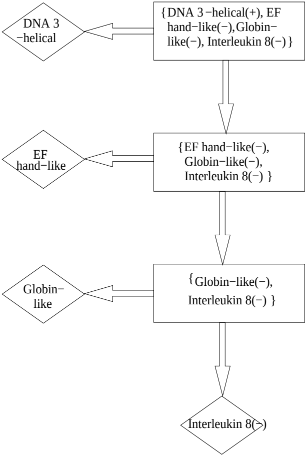
$$\phi(d1alla\_) = \phi(d1) = \begin{pmatrix} 1*1 & 1*1 & 1*1 \end{pmatrix}^T = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}^T$$

$$\phi(d2hbg\_\_) = \phi(d2) = \begin{pmatrix} 1*1 & 0*1 & 1*1 \end{pmatrix}^T = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}^T$$

$$k(d1, d2) = k(d2, d1) = 2,\ k(d1, d1) = 3 \text{ and } k(d2, d2) = 2$$

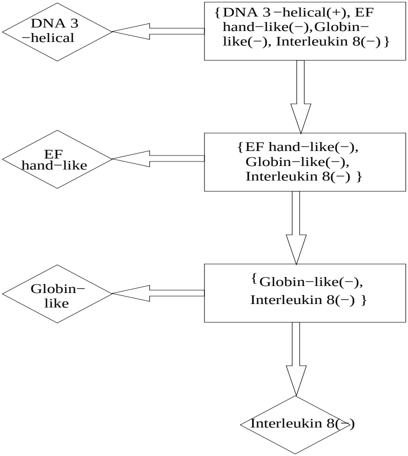
Multi-class Classification: Decision List based SVILP (DL_SVILP)

```
Require: A set of training examples $d_i \in D$ and $c_i \in \{1, 2, \ldots, r\}$ and a vector $index$ that represents learned structure of the list.
  for $j = 1$ to $r - 1$ do
    $p = index[j]$   /* Select a class $p$ from $r$ classes */
    /* Formulate the binary class problem by assigning label '1' to examples of class $p$ and '-1' to examples of remaining classes */
    $f_i : D_i \rightarrow \{1, -1\}$   /* Induce a binary classification function $f_i$ by applying SVILP to set $D_i$ */
    $D_{i+1} = D_i \setminus D_p$   /* Reduce the size of set $D_i$ by removing the examples belonging to class $p$ */
  end for
  return $f_i$ for $i = 1, \ldots, r - 1$
```


DNA 3 −helical
{DNA 3−helical(+), EF hand−like(−),Globin−like(−), Interleukin 8(−)}
EF hand−like
{EF hand−like(−), Globin−like(−), Interleukin 8(−) }
Globin−like
{Globin−like(−), Interleukin 8(−) }
Interleukin 8(−)

Multi-class Classification: Decision list-based SVILP (DL_SVILP)

1. Begin at the root node
2. Apply the classifier associated with the node to example *d*
3. Travel down the edge labeled by the classifier's output
4. If the edge is labeled positive output the class associated with the leaf. If the edge is labeled negative repeat steps 2 and 3 until the last positive edge is reached. Output the label given by the node.


{DNA 3−helical(+), EF hand−like(−),Globin−like(−), Interleukin 8(−)}
DNA 3 −helical
{EF hand−like(−), Globin−like(−), Interleukin 8(−) }
EF hand−like
{Globin−like(−), Interleukin 8(−) }
Globin−like
Interleukin 8(−)

Learning underlying Structure for DL_SVILP

```
Require: Training set, $d_1, d_2, \ldots, d_n$, validation set, $d'_1, d'_2, \ldots, d'_s$, $r$ classes and a logic based kernel learning algorithm (such as SVILP)
  for $j = 1$ to $r$ do
    /* Formulate the binary class problem by assigning label '1' to examples of class $j$ and '-1' to examples of remaining classes */
    /* Induce a binary classification function by applying SVILP to training data, $d_1, d_2, \ldots, d_n$ */
    /* Apply the learned function to validation set, $d'_1, d'_2, \ldots, d'_s$ */
    /* Measure performance of classifier by using expression */
    $S[j]' = W_P * P^- + W_N * N^+$
    where $P =$ total number of positive example, $N =$ total number of negative examples, $P^- =$ number of misclassified positive examples, $N^- =$ number of misclassified negative examples, $W_P = \frac{N}{P}$ and $W_N = 1$
    $index[j]' = j$
  end for
  /* Sort list $S'$ in ascending order and reorder list $index'$ accordingly */
  $S = sort(S')$
  $index = reorder(index')$
  return $index$ and $S$
```

Evaluation Measures

$$P_j = \#\text{examples in class } j$$

$$P = \sum_{j=1}^{j=k} P_j = \#\text{examples in } k\text{classes}$$

$$TP_j = \#\text{correctly classified examples in class } j$$

$$\text{Accuracy}_j = \frac{TP_j}{P_j}$$

$$\text{Overall accuracy (OA)} \; = \frac{\sum_{j=1}^{j=k} TP_j}{P}$$

Experiments: Recognizing Protein Folds

- 381 protein domains
- 20 folds of SCOP categorized into 4 structural classes, namely α, β, α/β $\alpha + \beta$
- SCOP folds:
  1: DNA 3-helical, 2: EF hand-like, 3: Globin-like, 4: 4-Helical cytokines, 5: Lambda repressor, 6: Ig beta-sandwich, 7: Tryp ser proteases, 8: OB-fold, 9: SH3-like barrel, 10: Lipocalins, 11: α/β *TIM*-barrel, 12: Rossmann-fold, 13: P-loop, 14: Periplasmic II, 15: α/β-Hydrolases, 16: Ferredoxin-like, 17: Zincin-like, 18: SH2-like, 19: β-Grasp, and 20: Interleukin.

Experiments: Recognizing Protein Folds

Table: 5-fold cross-validated accuracies for 20 SCOP folds.

Fold	#Exm	MC_ILP	DL_SVILP	MC_SVM
α				
1	30	93.3 ± 4.6	66.7 ± 8.6	43.3 ± 9.1
2	14	28.6 ± 12.1	57.1 ± 13.2	14.3 ± 9.4
3	13	46.2 ± 13.8	53.9 ± 13.8	46.2 ± 13.8
4	10	10.0 ± 9.5	30.0 ± 14.5	0.0 ± 0.0
5	10	40.0 ± 15.5	40.0 ± 15.5	30.0 ± 14.5
OA		55.8 ± 5.7	54.6 ± 5.7	31.2 ± 5.3
β				
6	45	73.3 ± 6.6	88.9 ± 4.7	68.9 ± 6.9
7	21	57.1 ± 10.8	90.5 ± 6.4	66.7 ± 10.3
8	20	0.0 ± 0.0	35.0 ± 10.7	25.0 ± 9.7
9	16	43.8 ± 12.4	75.0 ± 10.8	68.8 ± 12.0
10	14	64.3 ± 12.8	71.4 ± 12.1	71.4 ± 12.1
OA		52.6 ± 4.6	75.9 ± 4.0	61.2 ± 4.5

Experiments: Recognizing Protein Folds

Table: 5-fold cross-validated accuracies for 20 SCOP folds.

Fold	#Exm	MC_ILP	DL_SVILP	MC_SVM
α/β				
11	55	52.7 ± 6.7	76.4 ± 5.7	56.4 ± 6.7
12	21	52.4 ± 10.9	90.5 ± 6.4	28.6 ± 9.7
13	14	28.6 ± 12.1	50.0 ± 13.4	21.4 ± 11.0
14	13	7.7 ± 7.4	38.5 ± 13.5	0.0 ± 0.0
15	12	0.0 ± 0.0	8.3 ± 8.0	16.7 ± 10.8
OA		39.1 ± 4.6	64.4 ± 4.5	36.5 ± 5.0
$\alpha+\beta$				
16	26	53.9 ± 9.8	69.2 ± 9.1	34.6 ± 9.3
17	13	15.4 ± 10.0	53.9 ± 13.8	30.8 ± 12.8
18	13	7.7 ± 7.4	53.8 ± 13.8	38.5 ± 13.5
19	12	0.0 ± 0.0	25.0 ± 12.5	33.3 ± 13.6
20	9	77.8 ± 13.9	66.7 ± 15.7	22.2 ± 13.9
OA		32.9 ± 5.7	54.8 ± 5.6	32.9 ± 5.6
OA		45.4 ± 2.6	64.0 ± 2.5	42.3 ± 2.5

Experiments: Recognizing Protein Folds

Table: Accuracy ± standard deviation for 45 protein folds.

Fold	MC_ILP	DL_SVILP
α	57.78 ± 5.21	62.22 ± 5.11
β	33.64 ± 4.57	45.79 ± 4.82
α/β	56.45 ± 4.45	62.90 ± 4.33
$\alpha + \beta$	66.67 ± 5.41	72.62 ± 5.27
All	52.84 ± 2.48	60.25 ± 2.43

Conclusion

- Logic based multi-class classification method
- Accurate solutions to protein fold recognition probmen