Designing Abstract Features for Sequence Classication Tasks

Cornelia Caragea, Joint Work with Adrian Silvescu and Vasant Honavar

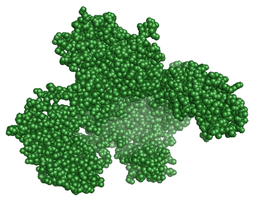
Computer Science Department, Iowa State University

IOWA STATE UNIVERSITY

Real World Applications

Bioinformatics: protein localization?

$\cdots$ PVKLKPGMDGPKVKQWPLTEEKIKA $\cdots$



Sequence Data:

From Data to Knowledge

Machine learning offers an approach to the design of algorithms for training computer programs to *efficiently and accurately classify text and biological sequence data*

Challenge: Data representation provided to a learner

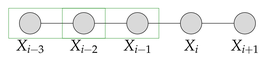
- The data representation has to be:
  - rich enough to capture distinctions that are relevant from the standpoint of learning
  - but not so rich as to make the task of learning harder due to *overfitting*

Approaches to Feature Construction

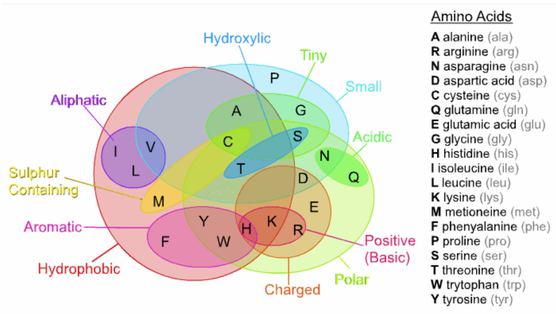
Let $\mathbf{x} = (x_0, x_1, \cdots, x_{t-1})$ be a sequence over a finite set $\mathcal{X}$, $\mathbf{x} \in \mathcal{X}^*$

- Super-structuring:
  - Is the operation of generating all the contiguous sub-sequences of a certain length k from $\mathbf{x}$ ("super-structures" or k-grams): $(x_{i-k}, \cdots, x_{i-1})$ for $i = k, \cdots, t$

$k = 2$:



  - Helps model dependencies between neighboring elements in a sequence
- Abstraction:
  - Is the operation of grouping "similar" entities to generate more abstract entities



Amino Acids

A alanine (ala)
R arginine (arg)
N asparagine (asn)
D aspartic acid (asp)
C cysteine (cys)
Q glutamine (gln)
E glutamic acid (glu)
G glycine (gly)
H histidine (his)
I isoleucine (ile)
L leucine (leu)
K lysine (lys)
M metioneine (met)
F phenyalanine (phe)
P proline (pro)
S serine (ser)
T threonine (thr)
W trytophan (trp)
Y tyrosine (tyr)

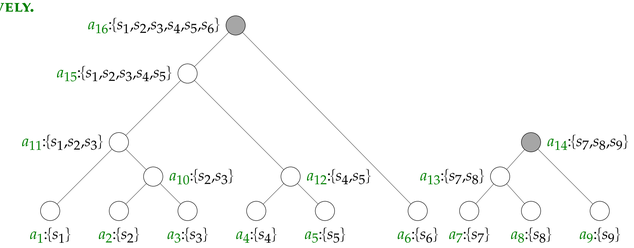
Our Approach

Combining super-structuring and abstraction to construct new features!

Constructing Abstractions over K-Grams

- Greedy agglomerative procedure
- Initially map each abstraction to a k-gram
- Recursively group pairs of abstractions until m abstractions are obtained

Constructing two Abstractions $\mathcal{A} = \{a_{16}, a_{14}\}$ on a set $\mathcal{S} = \{s_1, \cdots, s_9\}$ of 2-grams over an alphabet of size 3. The abstractions a_1 to a_9 correspond to the 2-grams s_1 to s_9, respectively.



Distance between two abstractions a_u and a_v

Let A denote a random variable that takes values in a set of abstractions $\mathcal{A} = \{a_1, \cdots, a_m\}$.
Goal: find a set of abstractions s.t. the reduction in the mutual information between A and the class variable Y, $I(A, Y)$, is minimized at each step of the greedy procedure.
We have shown that the reduction in $I(A, Y)$ due to a merge $\{a_u, a_v\} \rightarrow a_w$ of the greedy procedure is given by: $\boxed{\delta I(\{a_u, a_v\}, a_w) = (p(a_u) + p(a_v)) \cdot JS_{\pi_u, \pi_v}(p(Y|a_u), p(Y|a_v)) \geqslant 0}$ where

$$JS_{\pi_1,\pi_2}([p_1(\mathcal{Y})], [p_2(\mathcal{Y})]) = \pi_1 KL(p_1(\mathcal{Y}) || p(\mathcal{Y})) + \pi_2 KL(p_2(\mathcal{Y}) || p(\mathcal{Y}))$$

Hence, the distance between two abstractions is as follows:

$$d_{\mathcal{D}}(a_u, a_v) = \delta I(\{a_u, a_v\}, a_w) \text{ where } a_w = \{a_u \cup a_v\}$$

Feature selection

- alternative approach to reducing the number of k-grams to m k-grams
- we used mutual information between the class variable and k-grams to rank the k-grams

Task: Protein Subcellular Localization Prediction

- **plant** data set [Emanuelsson *et al.*, 2000]
  - 940 protein sequences classified into: *chloroplast*, *mitochondrial*, *secretory pathway/signal peptide*, and *other*
- **non-plant** data set [Emanuelsson *et al.*, 2000]
  - 2738 protein sequences classified into: *mitochondrial*, *secretory pathway/signal peptide*, and *other*

Experiments

We compare Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers trained using:

- unigrams: a bag of letters representation of protein sequences, no super-structuring, abstraction or feature selection (UNIGRAM);
- super-structuring: a bag of k-grams ($k = 3$) representation of protein sequences (SS);
- super-structuring and feature selection: a bag of m k-grams ($k = 3$) chosen using feature selection from the bag of k-grams obtained by super-structuring (See Section 3 for details) (SS+FSEL);
- super-structuring and abstraction: a bag of m abstractions over k-grams ($k = 3$) obtained using the combination of super-structuring and abstraction (See Section 2 for details) (SS+ABS).

Results

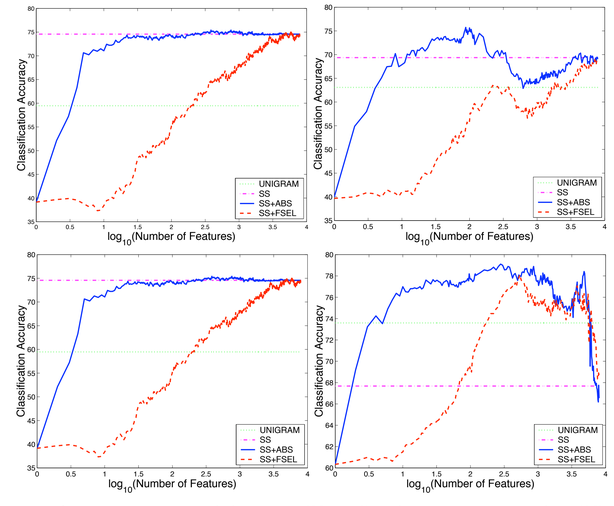


Figure: Comparison of super-structuring and abstraction (SS+ABS) with super-structuring alone (SS), super-structuring and feature selection (SS+FSEL) and UNIGRAM on the **plant** and **non-plant** data sets using Naïve Bayes (NB) (left column), and Support Vector Machines (SVM) with linear kernel (right column). The plots show the accuracy as a function of the number of features used in the classification model, ranging from 1 to $\approx 8,000$ on both data sets. The x axis shows the number of features on a logarithmic scale.

Analysis of Abstractions

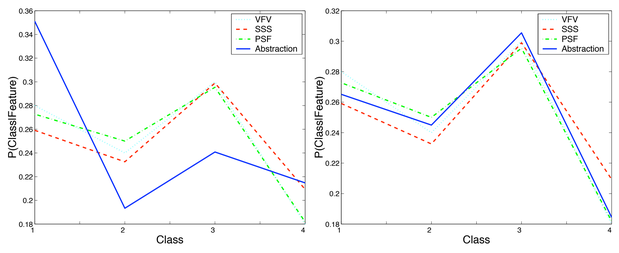


Figure: Class probability distributions induced by one of the m abstractions, namely a_i, and by three 3-grams, namely "VFV", "SSS", and "PSF", on the **plant** data set, where $m = 10$ and $i = 1$ (left); and $m = 100$ and $i = 3$ (right). The three 3-grams are initially sampled from a_3 (when $m = 100$). The number of classes in the data set is 4.

Conclusions

- We have shown that:
  - combining super-structuring and abstraction makes it possible to construct predictive models that use significantly smaller number of features than those obtained using super-structuring alone.
  - abstraction in combination with super-structuring yields better performing models than those obtained by feature selection in combination with super-structuring.

Acknowledgements: This work is supported in part by a grant from the National Science Foundation (NSF 0711356) to Vasant Honavar.