PR2: Peptide Classification

Description:

****
This is a study partner team assignment. However, each partner must submit at least once to CLP and each student must submit their own copy of the report.
****

Overview and Assignment Goals:

The objectives of this assignment are the following:

- Experiment with different feature extraction techniques.
- Try dimensionality reduction techniques (optional)
- Experiment with various classification models.
- Think about dealing with imbalanced data.

Detailed Description:
Develop predictive models that can determine, given an antibacterial peptide, whether it is also an antibiofilm peptide.

"Proteins are large biomolecules, or macromolecules, consisting of one or more long chains of amino acid residues. Proteins perform a vast array of functions within organisms, including catalyzing metabolic reactions, DNA replication, responding to stimuli, providing structure to cells, and organisms, and transporting molecules from one location to another. Proteins differ from one another primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes, and which usually results in protein folding into a specific three-dimensional structure that determines its activity.

A linear chain of amino acid residues is called a polypeptide. A protein contains at least one long polypeptide. Short polypeptides, containing less than 20-30 residues, are rarely considered to be proteins and are commonly called peptides. [...] The sequence of amino acid residues in a protein is defined by the sequence of a gene, which is encoded in the genetic code. In general, the genetic code specifies 20 standard amino acids; [...] Proteins can also work together to achieve a particular function, and they often associate to form stable protein complexes." [Wikipedia, Accessed 2020-02-07, https://en.wikipedia.org/wiki/Protein]

Biofilms are tightly-connected multicellular communities of microorganisms encased in self-secreted extra-cellular matrices. They are currently one of the major causes of disease for two main reasons. First, roughly 75% of all human infections are caused by biofilms. Second, due to the robust multicellular cellular matrix structure, they are resistant both to the host defense mechanisms and to traditional antimicrobial compounds (antibiotics). Thus, it is important to identify peptide sequences that are not only antimicrobial (can destroy or render inert the invading microorganism), but also antibiofilm (can penetrate the extra-cellular matrix so it can get to the microorganism in the first place).

You have been provided with a training set (train.dat) and a test set (test.dat) consisting of peptide sequences, one per line in the file. Peptides are encoded as strings with characters from an alphabet of 20 characters, each representing an amino-acid residue. The training set also includes the label for each sequence as 1 (antibiofilm) or -1 (not antibiofilm) as the first character in each line of the training file, separated from the sequence by a tab (\t) character.

The input to your classifiers will not be the peptides themselves, but rather features extracted from the peptides. Two simple approaches for feature extraction are the bag-of-words and the k-mer models we have learned in class, where a word is one of the amino-acids in the peptide. In addition, you may ***optionally*** use external data you may find online that provides additional information about the given peptide sequences. If you do, you must include the source in your report and provide a version of the dataset with the additional data to the Professor and the TA for verification. Once you have created your dataset, you may ***optionally*** apply feature selection or dimensionality reduction techniques to improve classification results. Describe those techniques in your report.

Note that the dataset is imbalanced. We will use a different evaluation metric for this assignment, Matthews's correlation coefficient (MCC), which, similar to the F-1 score, combines aspects of the result's sensitivity and specificity. Given the normal confusion matrix resulting from comparing the predicted and true classes of the test samples, MCC is defined as,

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Caveats:

+ Remember that not all features will be good for predicting the object class. Think of feature selection, engineering, reduction (anything that works).
+ Use the data mining/machine learning knowledge you have gained until now, wisely, to optimize your results.

Data Description:

The training dataset consists of 1566 records and the test dataset consists of 392 records. We provide you with the training class labels and the test labels are held out. Your task is to predict those labels for the peptides in the test set and create a test.txt file containing those labels, which you will submit to CLP. Note that CLP only accepts files with extensions .txt or .dat.

Rules:

- This is an individual assignment. Discussion of broad level strategies are allowed but any copying of prediction files and source codes will result in an honor code violation.
- You are allowed 5 submissions per day.
- After the submission deadline, only your chosen or last submission is considered for the leaderboard.

Deliverables:

- Valid submissions to the Leader Board website: https://clp.engr.scu.edu (username is your SCU username and your password is your SCU password).

Canvas Submission for the report:

- Include a 2-page, single-spaced report describing details regarding the steps you followed for feature extraction, feature selection, and classifier model development. The report should be in PDF format and the file should be called **<SCU_ID>.pdf**. Be sure to include the following in the report:
    1. Name and SCU ID.
    2. Rank & MCC-score for your submission (at the time of writing the report). If you chose not to see the leaderboard, state so.
    3. Your approach.
    4. Your methodology of choosing the approach and associated parameters.
- Ensure you submitted the correct code on CLP that matches your output. Code does not need to be submitted on Canvas.

Grading:

Grading for the Assignment will be split on your implementation (70%), report (20%) and ranking submissions (10%).Extra credit (1% of final grade) will be awarded to the top-2 performing algorithms (both partners from each study team). Note that extra credit throughout the semester will be tallied outside of Canvas and will be added to the final grade at the end of the semester.

Files: available on Canvas.