Universite de Sherbrooke - Techniques d’Apprentissages (IFT712) 12/12/2019
Sebastien Leblanc
Jeremie Beliveau

Machine learning techniques applied to the discovery of novel protein interactors within high throughput mass spectrometry data.

Abstract

The network of physical interactions between proteins is particularly relevant to the investigation of pathological processes at the cellular level. Here we explore a dataset aimed at uncovering such interactions and employ machine learning techniques to extract information from data with incomplete labels. The task requires training a one-class classifier that predicts whether a pair of proteins are high confidence interaction proteins (HCIP) or if the identification is part of experimental background (B). Mass spectral data acquired from the purification of over 5000 proteins and their interactors is used to derive a vector of 9 features that characterises each putative interactions and labels for HCIPs are obtained from the literature. We explore the usage of a basis function to project the data in higher dimensional space followed by linear classification. We also attempt kernel methods and neural network approaches with hyperparameter searches. Finally bootstrapping aggregation (bagging) of model ensemble yielded the best results

Introduction

Description of the problem - The BioPlex protein-protein interaction network is the largest dataset concerning the physical interaction network of human proteins.[1] It consists of over 5000 tagged protein purification followed by mass spectrometry analysis where each peptide spectrum association is a putative indication of physical interaction with the tagged protein. The high throughput nature of the data acquisition process however introduced a large amount of noise so that a majority of identifications are considered part of the background. The authors of the original study designed the features used in the present project and trained a naive bayes classifier, but did not publish performance metrics. The original method proved to be difficult to reproduce and yielded poor results (~60% overlap with the final published network). In addition, it is unclear whether the selected features are discriminatory enough for the task.

A major hindrance is the lack of labeled protein-protein interaction data in the literature to properly train a model. Notably a lack of true negative examples since no database exists confirming the non-interaction of any pair or groups of proteins. Of 6.7M putative pairwise interactions, 76K have been confirmed in the laboratory and are reported by two major sources.[2,3] The approach developed here models the data using positive example (P) as high confidence interacting proteins (HCIP) and unlabeled (U) as the rest of the putative interactions. The U group contains both HCIP and background (B) detections, but the underlying ratio between the two is unknown. What is known in the literature is that the degree of connectivity of proteins in interaction networks follows a power law—in other words, the distribution of the number of interactions per protein is such that a large number of proteins have a very small number of interactions and a few proteins have a large number of interactions. This distribution of degree of connectivity could eventually be used as a validation to support the results of our analysis but falls outside the scope of the current project.

One additional challenge is the evaluation of model performance. Since the exact rate of true negative is unknown in both the training and test sets many of the commonly used metrics become out of reach. We must find the proper metric that will allow us to compare the models and strike a balance in the bias-variance tradeoff.

In summary, our task is a form of retrieval problem where a few examples of a positive class are given and the task consists in retrieving similar points from a set of unlabeled points. We approach it using various machine learning algorithms and methods including linear classification, support vector machines, multi-layer perceptrons and ensemble methods.

Dataset description

Preprocessing - The complete dataset comprises more than 6M points over 9 dimensions. Since a data set of this scale is untenable for the current preliminary exploration with reduced it to include a random sampling of 325K (3%) of the interactions from the unlabeled group and 45K (60%) from the labeled group for a total of 370K points. We note that the P-U balance of the random sample set is different from the original dataset but we opted to capitalize on the available labels to ensure proper training. The data was scaled around 0 and the distribution for each feature is shown in figure 1.

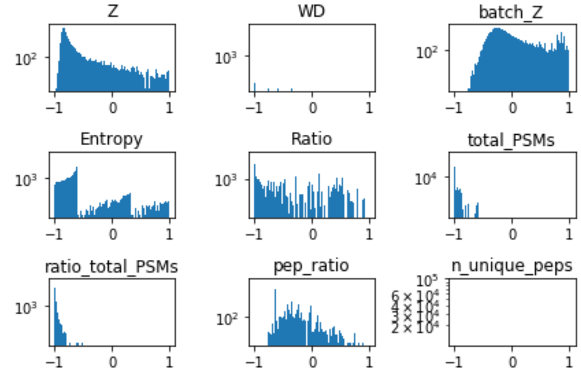

Figure 1. Data distribution of each feature in the 9 dimensional sample data set.

Dimensionality reduction - T-distributed stochastic neighbor embedding (T-SNE) was used to project the 9 dimensional data into 2D space to visualize the discriminative potential of the features. From the feature distribution histograms we noted that “WD” and “n_unique_peps” seemed to contain little information and were removed in two subsequent T_SNE visualizations. Removal of these two dimensions seemed to have negligible effect but although they probably could be removed, we opted to continue including them in the following analyses. In general, the HCIP points seemed to cluster rather tightly while also present diffusely over the larger clusters of unlabeled points which suggest some class overlap. Data processing was accelerated using multithreading with the MulticoreTSNE package. Multiple combinations of parameters were attempted “by hand” and figure 2 shows the best results.

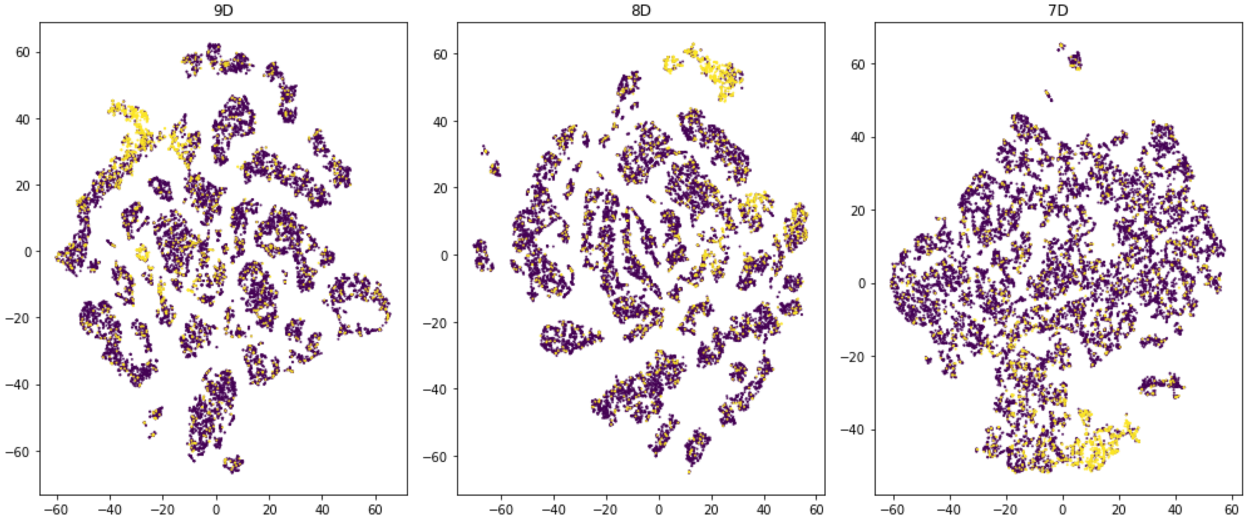

Figure 2. T-SNE visualisation of (9D) the sample data, (8D) the sample data with feature “WD” removed, and (7D) the sample data with features “WD” and “n_unique_peps” removed. Yellow: HCIP, purple: Unlabeled.

Methods

Metrics - The lack of negative labels in our dataset excludes many of the commonly used machine learning metrics. Accuracy and specificity both cannot be measured because they require the rate of true negatives. The rate of false positives is also excluded since the true class of the unlabeled data is unknown. We must then rely on recall—or the rate of true positive—to assess the performance of our classifier. However, recall alone is insufficient for proper characterisation since a constant prediction of the positive class would easily maximize this metric.

The F-score is commonly used to evaluate performance on retrieval tasks but as mentioned by Lee & Lui, it is inappropriate in the case of positive and unlabeled data sets because precision cannot be computed from the validation set.[4] They instead propose to use an adapted version where the harmonized mean of precision and recall (F-score) is transformed to obtain the following:

$$LL_{score} = \ recall^2 / P(f(x) = 1)$$

This score will allow to assess performance while detecting when the model starts to systematically predict the positive class irrespective of the data presented to it. Lee & Lui also note that their score behaves the same as the F-score; that is, it will be high if both precision and recall are high and low if either precision or recall are low.

Cross validation & hyperparameter search - All classification algorithms used in this project comprise a number of hyperparameters that significantly impact performance. To select appropriate combinations of each parameter we implemented a grid search algorithm that scores each combination via a 3 fold cross-validation scheme. The mean of the three LL scores obtained through a 3-tier stratification training and testing was computed and the combination of hyperparameters with the highest mean score was considered the most appropriate for that model.

Linear Classifier - The first attempt at classification consisted in the simple projection of the training points into a higher dimensional space using a polynomial basis function. A linear classifier was then trained on the high dimensional data and tested using the test dataset also projected using the same basis function.

An advantage of using such algorithm is its ease of implementation and relatively few hyperparameters. While convenient and versatile, it can sometimes be too sensitive. It may have the tendency to overfit or underfit, which means it may require tight adjustment of hyperparameters.[5]

Support Vector Machines - Kernel methods can be difficult to apply to large datasets because they require computing a pairwise distance between each training points. Some of their advantages include their effectiveness in high dimensional spaces and allow non-linear classification. Once trained prediction is fast and memory efficient.

The radial basis function (RBF) kernel uses exponentials which make it especially expensive computationally. The sigmoid kernel employs a different measure of distance between the training points.

Multi Layer Perceptron - The versatility of artificial neural networks (ANN) have made them popular for tackling a wide range of problems. Here we thought that it might be appropriate for the modeling of some non linear features within the dataset. Although it is easy to increase model capacity by increasing the number of neurons or layers, it does make for a rather large hyperparameter search space. However, overfitting and computational time increases as model size increases so it must be considered in the tradeoff. For this reason, we opted to test a model with less capacity than the best result from the hyperparameter grid search. We used stochastic gradient descent with adaptive learning rate to train the network.

Model Ensembles - Increasing model capacity eventually leads to overfitting of the positive class. To mitigate this problem in the specific case of a one-class classifier Mordelet *et al.* suggest to employ bootstrap aggregation (Bagging) of multiple models trained on different random samples of the data.[4] Here we attempt bagging with a SVM with the sigmoid kernel and MLP models.

Results

Hyperparameter search

We used Compute Canada in order to parallelize some hyperparameter searches. Figure 3 shows the evaluation of MLPs with varying numbers of neurons and layers. Although the model with 13 layers with each 105 neurons was most performant, we opted to use a model with layers and 45 neurons each for practical reasons.

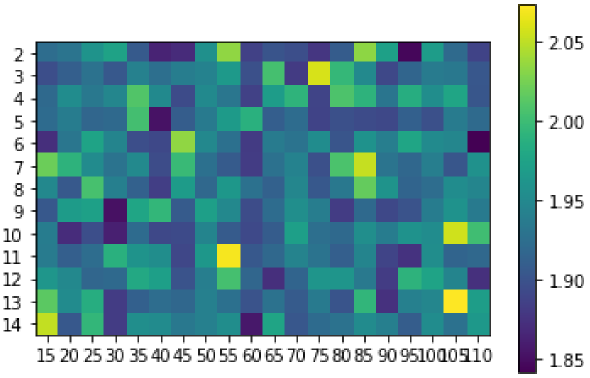

Figure 3. Hyperparameter search for number of neurons (x) and number of layers (y). Color represents the LL score.

The ranges for hyperparameter searches were tightened especially for the SVM with RBF since even one test took considerable time.

Performance

The usual graphical representation of model performance depicts a receiver operator characteristics curve (ROC). However, as mentioned before, these cannot be drawn here since we cannot compute the rate of false positives. Instead we plot the LL score against recall, where the best models will have both a high recall and high LL score.

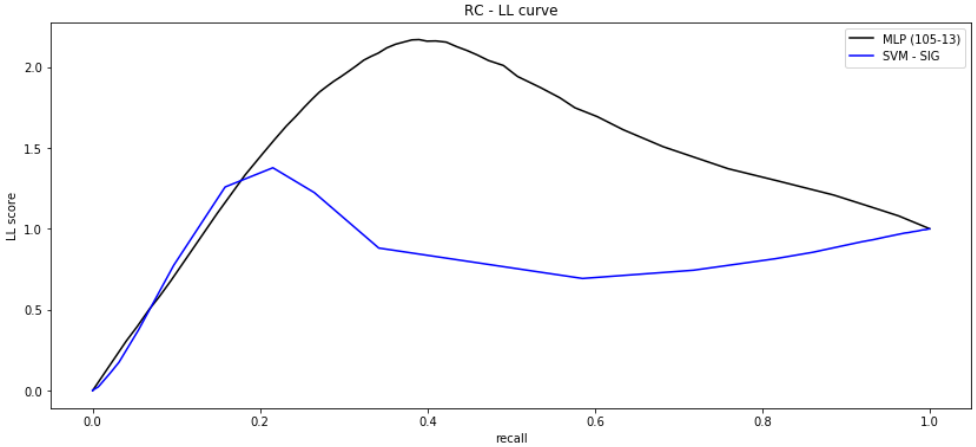

Figure 4. Performance comparison between SVM with sigmoid kernel and MLP.

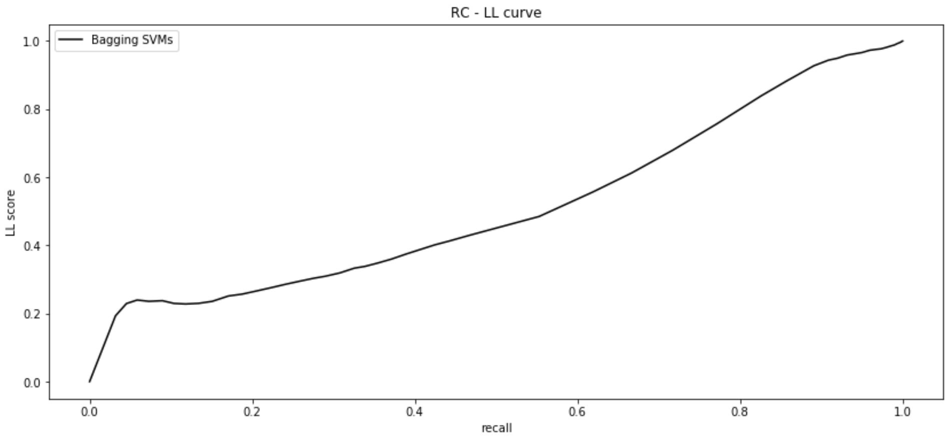

Figure 5. Performance of bagging SVM with sigmoid kernel.

Model Comparison

10 fold cross validation runs of each model on the same data revealed that the polynomial basis function and linear classifier have the best performance. It is somewhat surprising to see the bootstrap aggregation of MLP lagging so far behind. This poor performance, albeit with a large variance, may be due to the small sample size shown to each of the models in the array. The bagging algorithm seems to have worked to increase the performance of the SVM with sigmoid kernel (SVM-SIG) by a small margin, also with a larger variation in scores. The SVM model with RBF kernel was the longest to compute but outperformed both SVM-SIG and its bagging variety.

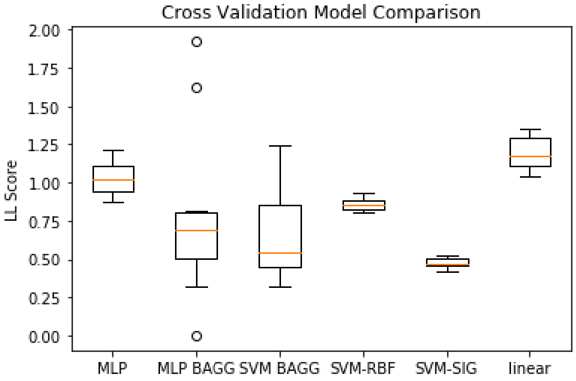

Figure 5. Model comparisons with cross-validation.

Discussion

The BioPlex protein-protein interaction data set presents many challenges. It is large, but lacks labels for the majority of the data and it is unclear whether the designed features are informative enough of the physical interaction between the proteins under study. The high cost of labeling this type of data motivated an in depth exploration of the readily available data in search for a way to identify the most likely interacting candidate. While the dataset was generated faster than any previous mass spectrometry proteome wide investigation of physical interactions, the quality of the resulting data seems to have suffered.

While we attempted many different sophisticated methods including artificial neural networks and model aggregation methods, the most performant turned out to be a simple polynomial projection followed by linear classification. This could be because we had to cut short some of the hyper parameter searches for other models.

The model evaluation metric by Lee & Lui proved useful if somewhat hard to interpret. Further, evaluating the distribution of degree of connectivity of the network at different recall threshold might offer some insight.

References

1. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425 (2015).
2. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
3. Franz, M. *et al.* GeneMANIA update 2018. *Nucleic Acids Res.* **46**, W60–W64 (2018).
4. Mordelet, F. & Vert, J.-P. A bagging SVM to learn from positive and unlabeled examples. *ArXiv10100772 Stat* (2010).