Classification of Protein Ubiquitination Sites using **MLP and GA**

The Pitch, April 3rd 2017

Carleton University

Group: 11

Jinny Lee MASc Student under the supervision of Dr. Andy Adler, Dr. Eran Ukwatta

George Hanna 4th year Biomedical and Electrical Engineering Student.

Review our Method/Implementation

"Hidden Markov Models cannot be used"

Review our Method/Implementation

Method

- Classifier : Multilayer Perceptron
- Feature selection : Genetic search

Implementation

Work environment: Matlab 2017a NN Toolbox
Feature selection by WEKA

Processing Data

Pre-processing data

1. **Normalization**: Multilayer network creation functions such as “patternnet” include default processing functions such as “removeconstantrows” and “mapminmax”

Processing Data

Pre-processing data

1. **Normalization**: Multilayer network creation functions such as "patternnet" include default processing functions such as "removeconstantrows" and "mapminmax"

2. **Missing Data**: -9999 values are replaced with column medians.

Processing Data

Pre-processing data

1. **Normalization**: Multilayer network creation functions such as "patternnet" include default processing functions such as "removeconstantrows" and "mapminmax"

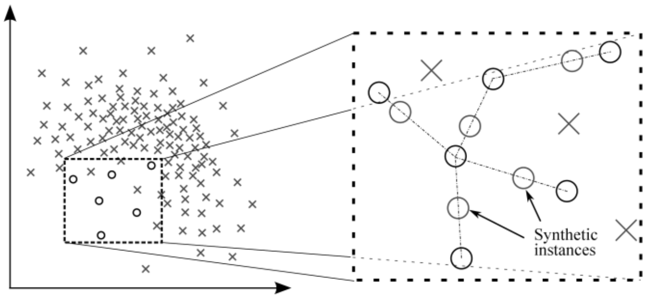
2. **Missing Data**: -9999 values are replaced with column medians.

3. **Outlier Detection:** Mean Absolute Deviation (MAD) used to replace values that are >10 deviations away with the median value.

Processing Data

Mitigating Class Imbalance

- **Synthetic Sampling (SMOTE)**
- Adaptive SMOTE (AdaSyn)
- Cost-sensitive classification
- Undersampling

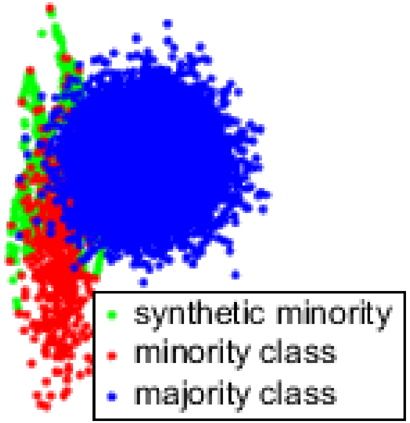


Sampling of Minority Class using K Neighbors

Processing Data

Mitigating Class Imbalance

- Synthetic Sampling (SMOTE)
- **Adaptive SMOTE (AdaSyn)**
- Cost-sensitive classification
- Undersampling



Sampling of Minority Class Near Border

Processing Data

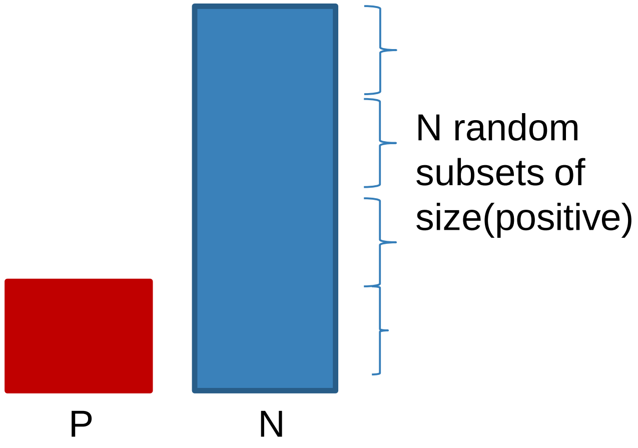
Mitigating Class Imbalance

- Synthetic Sampling (SMOTE)
- Adaptive SMOTE (AdaSyn)
- **Cost-sensitive classification**
- Undersampling

Processing Data

Mitigating Class Imbalance

- Synthetic Sampling (SMOTE)
- Adaptive SMOTE (AdaSyn)
- Cost-sensitive classification
- **Undersampling**

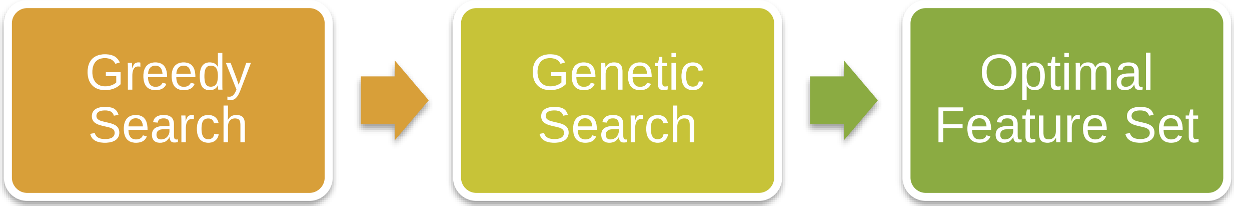


Experiment Design

Datasets : Split 60% for training and 40% for test set

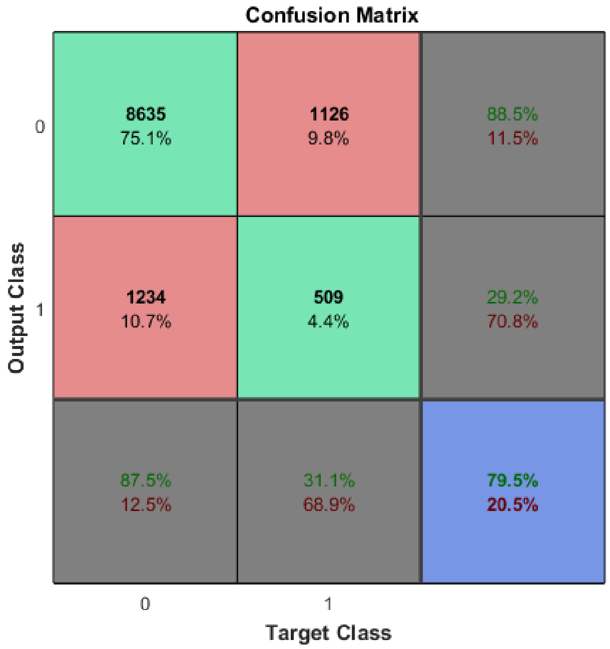
Feature Selection Using WEKA:

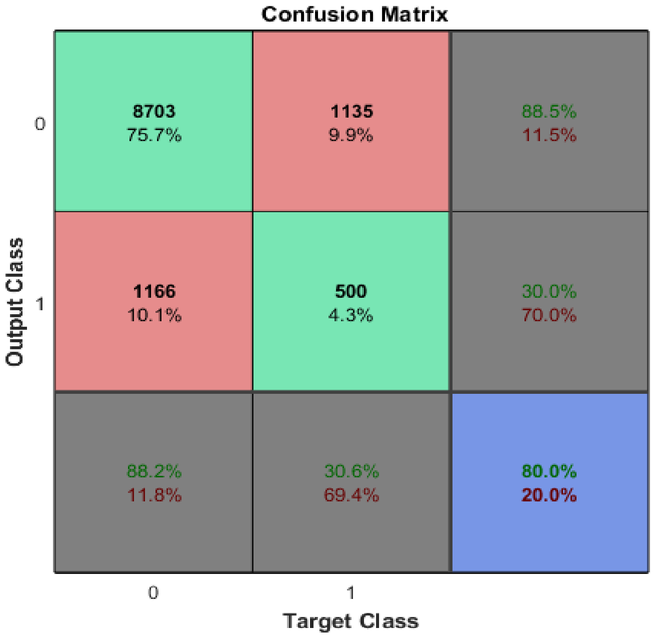
- CfsBestSubsetEval - Greedy search 10 - 60 features
- Results used as initial state for Genetic Search, giving 10 - 60 features



Experiment Design

Feature Selection Using WEKA:





Experiment Design

Training and Testing Protocol:

Train X NNs on X Bootstrap samples of training data

- Samples have equal proportion of positives and negatives
- Training uses Early Stopping, Bayesian Regularization to lower risk of overfitting.

NNs applied to hold-out test-set

- Outputs of each NN are summed, averaged, thresholded.

Parameter Tuning

- Learning Rate, Momentum, Hidden Layers

Experiment **Design**

Meta Learning

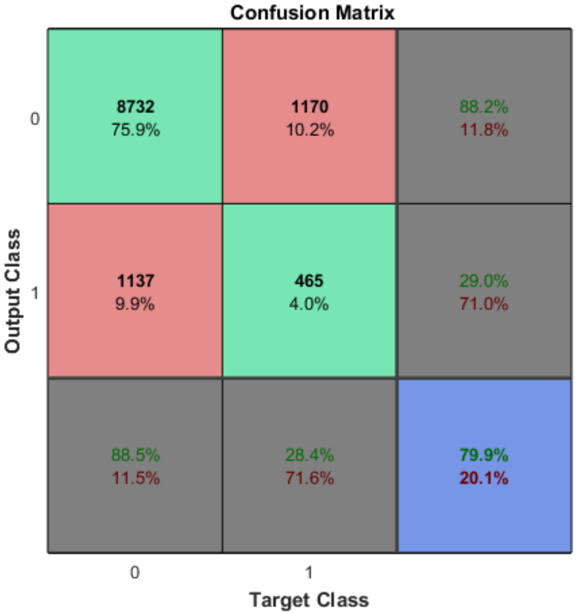
Hybrid Learners:
Combine weak NN with decision trees and KNN classifiers. Weak improvement in classification precision/recall

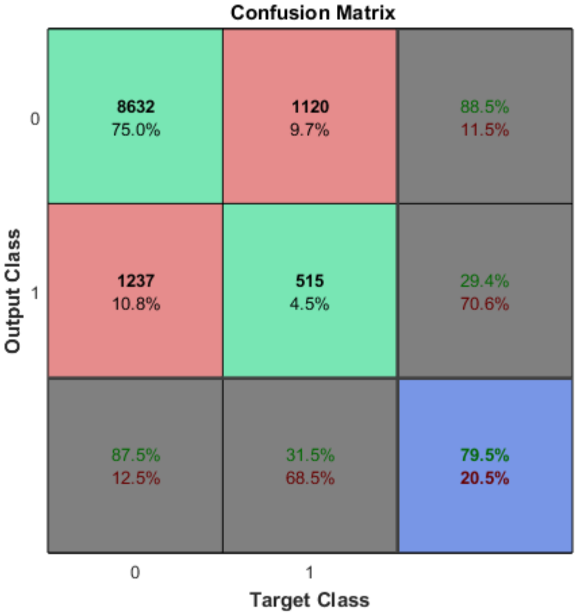
Bagging:
Several tens of NNs are trained on bootstrap samples of training set. Could be combined with Hybrid Scheme to improve stability

GA-based Selective NN Ensemble:
Assign voting weights to NNs using GA.

Experiment Design

Meta Learning





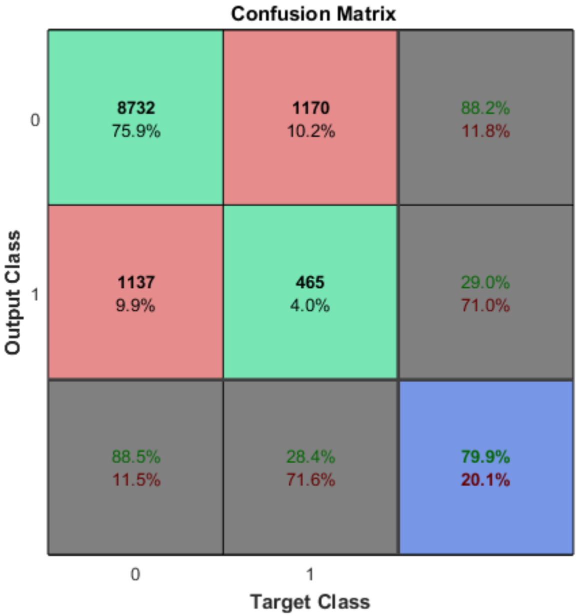
Experiment Design

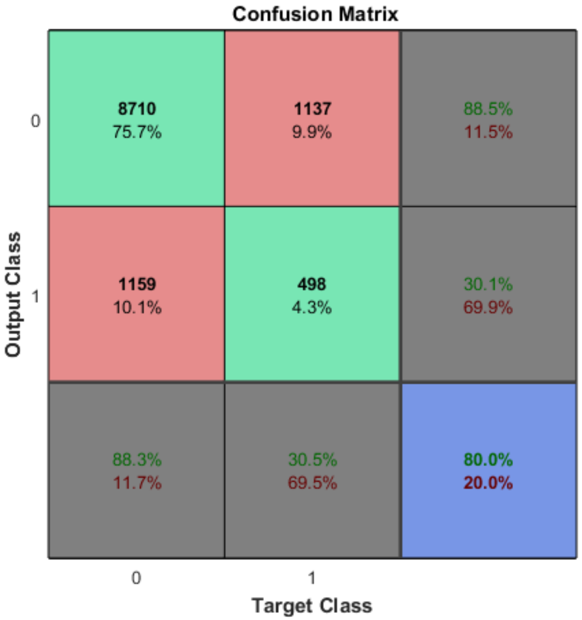
Active Learning

Voting-based Query Selection
1000 points are chosen where vote among X classifiers is closest. These points are deemed uncertain and should help in defining a more accurate classifier.

Experiment Design

Active Learning

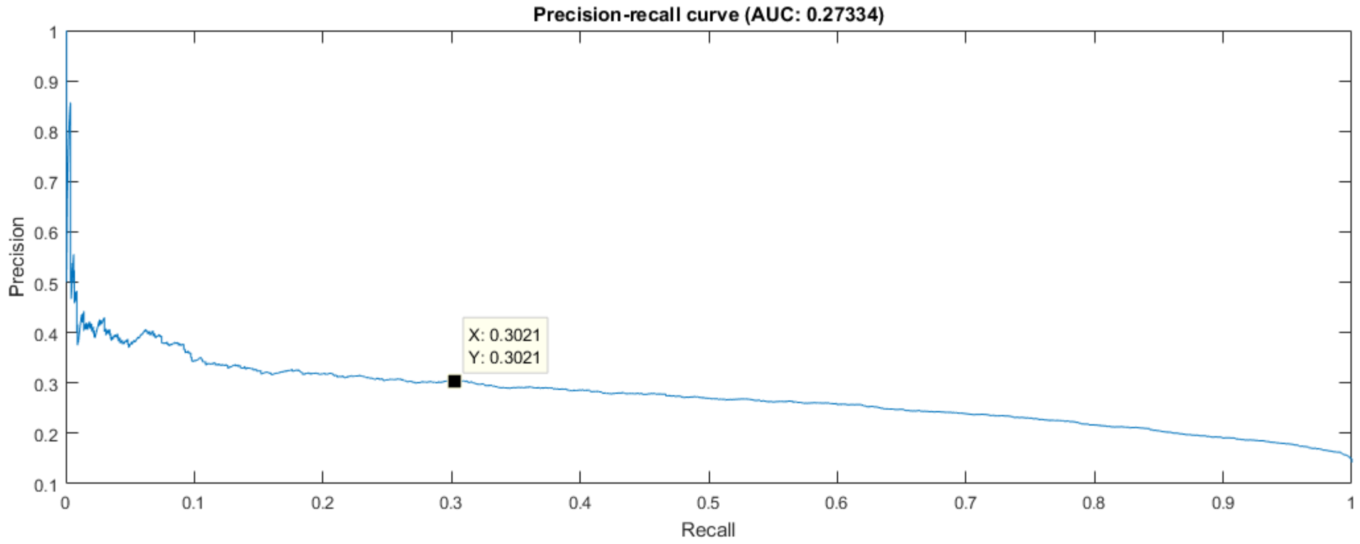




Estimated Results

Min Sensitivity: 30%

Min Precision: 30%



Estimated **Results**

Accuracy: 0.15 +/- 0.05