1.1.1 Variable Description

We hypothesize that semi-supervised method outperforms both unsupervised methods at any training set size and supervised methods at low training set sizes especially when positive labels are contaminated. The analysis combines features from two sources to predict essentiality of the genes in yeast Saccharomyces cerevisiae. Fourteen sequence-derived features from the Seringhaus, et al., 2006 study, and eight additional non-sequenced-derived features from the Ensembl website (ref). Feature definitions are listed in Table 1 (ref) along with the univariate performance in predicting essentiality in Table 2 (ref). Variables include predicted subcellular localization, GC content, hydrophobicity, number of interaction partners, etc. Five of 22 variables (mitochondria, vacuole in\_how\_many\_of\_5\_proks\_blast, intron, er) were removed from the analysis due to low content (less than 20% of non-zero values). At lower training sizes, their low content resulted in deterministic models because the randomization employed multiple seeds until the design matrix had no columns of zeroes. Genes which had values for this low-content variables would have been selected more frequently than genes without information in these features.

1.1.2 Cross-Validation Strategy

The cross-validation strategy incorporates an unbalanced strategy to the test set (Figure 1) along with a contamination rate. For an unbalanced design, test sets utilize the remaining genes not used in the training sets rather than a balanced strategy which matches set sizes. The unbalanced strategy was chosen because, in practice, an investigator would typically want to test all the remaining genes for essentiality rather than just a subset of genes. Another concept interwoven into the analysis is contamination. Contamination considers the dilemma from falsely assigned genes in the training sets. However, semi-supervise and unsupervised methods do consider negative labels in their computations and, thus, are unaffected by contamination. For supervised methods, positive labels in the training set are mixed with negative labeled genes for analysis when varying percentages of contamination are introduced.

In the unsupervised simulation, semi-supervised was compared against the unsupervised method described in section 2.4. We train our two methods from one of two sets of essential genes (positive labels) - known essential genes (769) and a subset of pre-2002 known genes (64). Additionally, we contrasted all 22 features against a subset of 14 sequenced-derived features as predictors of essentiality. Training set sizes were based on increments of 5 with minimum set sizes above the number of features to prevent rank deficiency in training sets. AUC from 30 iterations at each training set size summarized the performances of the two methods.

In the supervised simulation, semi-supervised was compared against three supervised methods (LASSO, SVM, and Random Forest) at low training set sizes. AUC performance of these four methods was compared across training set sizes between 1 and 10%. Unique initial seeds were chosen based on the iteration number, training set size, and contamination. Genes randomly chosen for the supervised training sets reflect the same ratio of positive and negative labels as seen in the full data set. Among the 3500 yeast genes, there are 769 essential genes resulting in a 21% ratio. For example, at 1% training size, 35 randomly chosen genes contained 7 positive labels (21% of 35) and 28 negative labels for supervised methods while semi-supervised methods analyzed 35 positively labeled genes. In order to mimic contamination, negative labels were reassigned a positive label at rates of 0%, 20%, and 50%.

1.1.3 Algorithms

All simulations were performed in R version 3.3.3. The semi-supervised and unsupervised analysis utilized functions from the *lcmix* package. The *lcmix* package developed and implemented in the previous paper by Dvorkin, Biehl, and Kechris A Graphical Model Method for Integrating Multiple Sources of Genome-scale Data and can be downloaded from [http://r-forge.r-project.org/projects/lcmix/](http://r-forge.r-project.org/projects/lcmix/" \t "pmc_ext). LASSO was performed using the *glmnet* command in the *glmnet* package (ref Hastie and Qian). Using *cv.glmnet*, k-fold cross validation optimized the minimum lambda for the LASSO function. SVM analysis used the *svm* command under the *e1071* package (ref David Meyer). Various runs using different criteria revealed a radial kernel density and C-classification optimized AUC performance. Random Forest was performed with the *randomForest* command under the *randomForest* package (ref Breiman, L.) All supervised predictions used the *predict* command in the *stats* package.

1.1.4 Performance

The AUC mean, variance, and CV (median absolute deviation) of the three supervised methods were contrasted against semi-supervised method. Because LASSO outperformed the other supervised models in AUC across all training set sizes and contamination rates, a closer evaluation of its performance was compared with semi-supervised method. In order to fairly contract LASSO performance to semi-supervised, the prediction scores were rescaled to (0,1). Precision, recall, and f-measure further discriminated the two methods with four rescaled prediction score cutoffs including the median and prediction scores of 0.5, 0.8, and 0.95. The median cutoff is a relative measure based on the data while the other three cutoffs are absolute. The f-measure was calculated from the average precision and recall at each training set size from 1% to 5%. The number of predicted positive values at each training size are presented in Supplemental Table 1.

1.2 Results

Semi-supervised method (red) tends to have smaller variance and more stable AUC across various training percentages. LASSO (blue) outperforms the other two supervised methods - SVM (aquamarine) and Random Forest (light blue). At low training sizes (≤ 2%), semi-supervised method is more consistent in AUC performance than the three supervised methods. Semi-supervised method outperforms LASSO when contamination rate is 50% and training size is ≤ 2%. (b) Semi-supervised method has a smaller variance compared to LASSO at training sizes less than ≤ 3% regardless of contamination levels. (c) Semi-supervised method has a lower CV compared to LASSO across all training sizes and contamination levels.

(a) At 1% training level, LASSO kernel densities of posterior probabilities exhibit an uni-modal distribution while semi-supervised methods exhibit multi-modal behaviors. Note that semi-supervised method does not change as the contamination rate increases. Semi-supervised method does not utilize negative labels, thus, alleviating the turbulence caused by contamination. (b) At 5% training level, LASSO kernel densities of posterior probabilities continue to exhibit uni-modal distributions while semi-supervised methods maintain their multi-modal behaviors. Supervised methods such as LASSO with uni-modal distributions do not intrinsically show a clear optimal cutoff compared to the multi-modality of semi-supervised predicted probabilities.

Performance of semi-supervised and LASSO methods with median and mid-range cutoffs. Posterior probabilities from 100 iterations were split at the median (a) and mid-range (b) for a cutoff to predict essentially of remaining yeast genes. Semi-supervised method is shown in red while LASSO is shown in blue. Precision, recall, and f-measures are represented by dotted, dashed, and solid lines, respectively. LASSO has slightly higher precision and f-measures across the training sets of 1% to 5% but loses this advantage as contamination levels increase. Semi-supervised method outperforms LASSO in recall for training sets ≤ 2.5%. As contamination increases, semi-supervised method outperforms LASSO at higher training set sizes and have indistinguishable recall measures thereafter.

Semi-supervise outperforms LASSO in recall and f-measure while LASSO bested semi-supervised in precision. Naturally, LASSO decreases in all performance markers as contamination increases. (a,b) With posterior probabilities ranging from near 0 and 1, the mid-range for both methods is near 0.5. Because of the heavily, skewed low probabilities in the uni-modal distribution from LASSO (Figure 4), the median would divide the set somewhere on the backside of the slope, greater than the maximum but less than 0.5. The median and mid-range for semi- supervise tend to fall near a local minimum, a useful indicator for separating distributional behaviors and cannot be evaluated in uni-modal distributions. The most dynamic difference between the two methods is the behavior of the recall performance. Recall measures how many positive labels were predicted out of the true number of positive labels. In the mid-range cutoff for LASSO, recall would naturally be lower than the median cutoff due to the small area in the right tail greater than 0.5. The expected increase in precision wasn’t strong enough to outperform semi-supervise in the combined f-measure.

Discussion points

Due to the inherent capacity of supervised methods to utilize both positive and negative labels, they have natural advantages over semi-supervised methods which only handles positive labels.

The contamination is a strategy to emulate a real-world scenario that a researcher may know a certain number of positive labels for genes in their experiment but are unsure if the remaining genes are truly negative.

Lasso may have an advantage over the other methods because it can reduce the effect of poorly predicting variables by collapsing their betas to 0.