Gene Function Prediction Using Labeled and Unlabeled Data

Paper by Xing-Ming Zhao, Yong Wang, Luonan Chen and Kazuyuki Aihara

Presented by Colin Kern

Gene Function Prediction

- Gene function prediction can be viewed as a classification due problem due to some assumptions:
  - Genes with similar expression patterns are assumed to have similar functions.
  - Interacting proteins have the same or similar functions.
- Gene expression and protein-protein interaction data can thus be used to train a classifier.

Training a Classifier

- Optimally, training is done with both positive and negative samples.
- Existing gene function data is only about positive samples.
    - i.e., we know which gene belongs to which functional class, but we are not sure which gene does not belong to the class.

Negative Samples

- It is inappropriate to simply use all the genes outside the target functional class as negative samples.
    - A gene may belong to more than one class.
    - It may belong to the class but is not known yet.
    - An imbalance problem may occur because there will be many more negative samples than positive ones.

AGPS

- This paper introduces a new technique called Annotating Genes with Positive Samples (AGPS) for defining negative samples in a training set. In particular:
  - A functional linkage graph is constructed to integrate heterogeneous information sources.
  - Singular value decomposition (SVD) is used to reduce the dimensionality and remove noise.
  - AGPS is presented to define negative samples and predict the function of unknown genes.

AGPS (cont)

- AGPS is a technique for defining negative samples in unlabeled data, so is independent from the learning algorithm.
- In this paper, SVMs were used for the learning algorithm.

Data Sources

- In this paper, three data sources were integrated into a functional linkage graph of S. cerevisiae genes.
  - BioGRID: Protein interaction
  - Stanford Gene expression Database
  - MIPS: Protein complexes
- 13 general functional classes were selected from the FunCat 2.0 database.

Functional Classes

	Functional Categories	Number of genes
1	metabolism	967
2	energy	241
10	cell cycle and DNA processing	727
11	transcription	829
12	protein synthesis	364
14	protein fate	680
20	cellular transport	726
30	cellular communication	86
32	cell rescue, defense and virulence	307
34	interaction with the environment	332
40	cell fate	201
42	biogenesis of cellular components	471
43	cell type differentiation	354

AGPS Input

- Positive training data P1
- Validation set P2
- Unlabeled data Ku
- Unknown gene Ug

AGPS Stage 1: Learning

- **U** = **Ku** + **P2**
- Stage 1.1: Initial negative set generation
  - Construct classifier f_1 based on **P1** and **U** with one-class SVMs
  - Classify **U** using f_1. The predicted negative set $\mathbf{N_1}$ is used as the initial negative training set in Stage 1.2
  - **U** = **U** - $\mathbf{N_1}$

AGPS Stage 1.2: Negative set expansion

- Classifier set FC = [], negative set NS = [], i = 1
- Repeat
    - $i = i + 1$
    - Construct classifier f_i based on **P1** and $\mathbf{N_1}$ with two-class SVMs
    - FC(i – 1) = fi, NS(i – 1) = **N1**
    - Classify **U** by f_i, $\mathbf{N_2}$ is the predicted negative set, where $|\mathbf{N_2}| <= k|\mathbf{P1}|$
    - $\mathbf{N_1} = [\mathbf{N_2}; \mathbf{N_{sv}}]$, where $\mathbf{N_{sv}}$ is the negative SVs of f_i in the previous step.
    - $\mathbf{U} = \mathbf{U} – \mathbf{N}_2$
- Until $|\mathbf{U}| < k|\mathbf{P1}|$

AGPS Stage 1.3
Classifier and negative set selection

- Classify **U** with classifiers from FC, and select the classifier FC(i) with the best prediction accuracy
- Return negative set **TN** ← NS(i)

Stage 2: Classification

- Classify **Ug** with **P** and **TN**, where **P** = **P1** + **P2**

Results

- AGPS was compared to four other methods
  - Conventional two-class SVMs
  - One-class SVMs
  - PSoL
  - Kernel integration
- SVD used to reduce dimensionality.
- Radial Basis Function (RBF) kernel was used for all the methods.

AGPS Method

- 10-fold cross-validation to find optimal parameters for kernel
- Validation genes and genes outside of the target functional family were considered unlabeled data
- In each stage of cross-validation, the best classifier and corresponding negative sample set were returned
- The most frequent samples appearing in the returned negative sample sets were used for the final negative samples
- The size of the final negative sample set was controlled to be nearly equal to the positive sample set size

PSoL Method

- 10-fold cross-validation to determine optimal kernel function parameters
- Unlabeled data set was defined as the genes outside the target class, unknown genes and the validation genes.

One-class SVMs Method

- Classifier trained only on the positive sample set
- 10-fold cross-validation used to find optimal kernel function parameters
  - 9/10 of the positive set was used as training set, the rest was used as a validation set.
- Genes not in the target class were used for negative test samples.

Two-class SVMs Method

- Negative samples consist of genes outside the target class
- 10-fold cross-validation used to find optimal kernel parameters
- Balanced training set used, where the number of positive and negative samples were equal.

Kernel Integration Method

- Diffusion kernel applied to protein-protein interaction and complexes
- RBF kernel applied to gene expression profiles
- Balanced training set was used.

Results of 10-fold cross-validation

Methods	precision(%)	recall(%)	F1(%)
AGPS	68	61	61
PsoL	68	37	47
Two-class SVMs	45	24	33
Two-class SVMs, balanced	61	70	69
One-class SVMs	50	21	31
Kernel integration	58	28	37
Kernel integration, balanced	64	47	52

Further Testing

- 386 previously unknown yeast genes have been annotated since March 2004, and so were not included in the training in the previous section.
- These genes were used as a test set

Prediction Results

Methods	precision(%)	recall(%)	F1(%)	ROC score
AGPS	15	66	22	0.61
Psol	20	18	19	0.55
Two-class SVMs	28	10	16	0.53
Two-class SVMs, balanced	18	36	29	0.57
One-class SVMs	10	42	15	0.53
kernel integration	39	16	23	0.56
kernel integration, balanced	11	32	24	0.59

Observations

- AGPS outperforms all other methods using ROC score.
  - The randomly selected negative training sets used for other methods cannot capture the true distribution of negative samples.
- One-class SVMs do poorly because of the low number of positive samples (underfitting)
- Although two-class SVMs and kernel integration have higher F1 scores, they have lower recall rates than AGPS.

Conclusion

- AGPS is shown to increase performance by selecting negative samples from unlabeled data.
- The advantage of having a balanced training set is shown.
- AGPS is shown to be superior at generating a negative training set than random selection.