COSNet: a Cost Sensitive Neural Network for Semi-supervised Learning in Graphs

Alberto Bertoni
Marco Frasca
Giorgio Valentini

Università degli Studi di Milano

Dipartimento di Scienze dell'Informazione

Outline

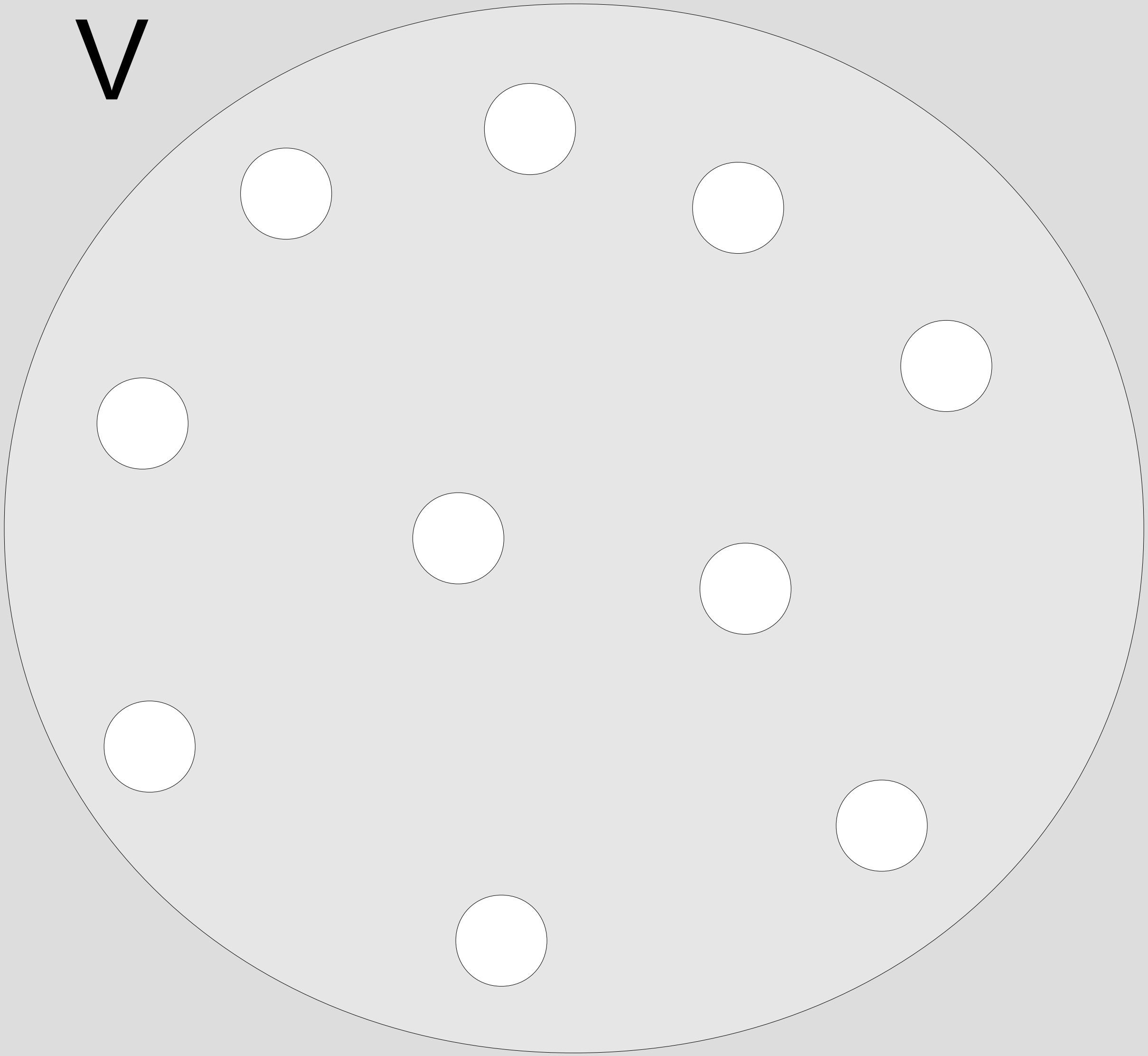
- The Gene Function Prediction (GFP) problem
- Related approach for GFP
- Our proposal for GFP
- Results in yeast
- Conclusions

Gene Function Prediction

Genome sequencing

- Main problem: understanding biological functions of new genes
- Taxonomy: hierarchical definition of gene properties
  - Gene Ontology(GO), FunCat
- Annotation: established involvement of a gene in the biological mechanism represented by a functional class (term)

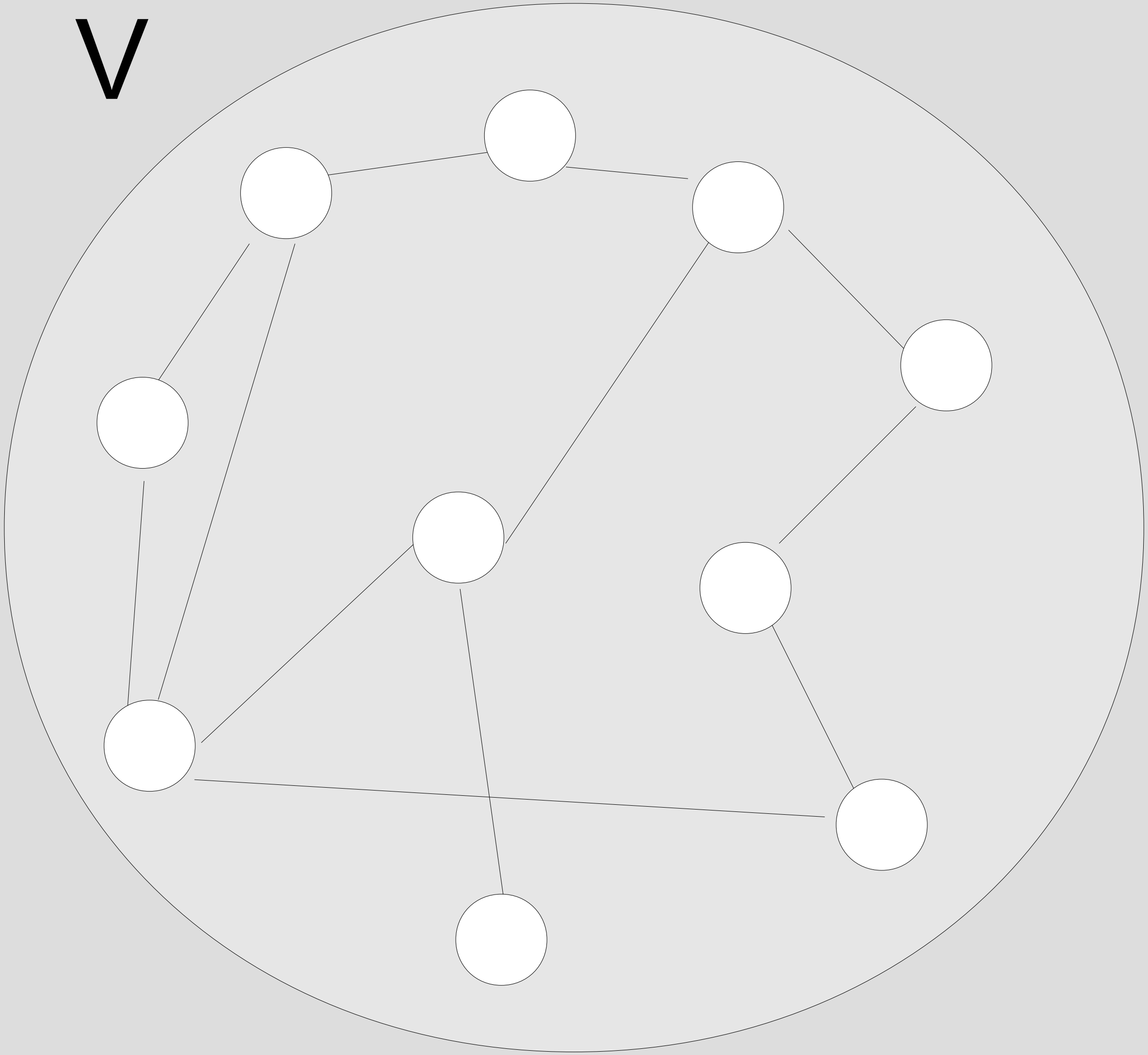
Gene Function Prediction Problem



Input:

- V genes

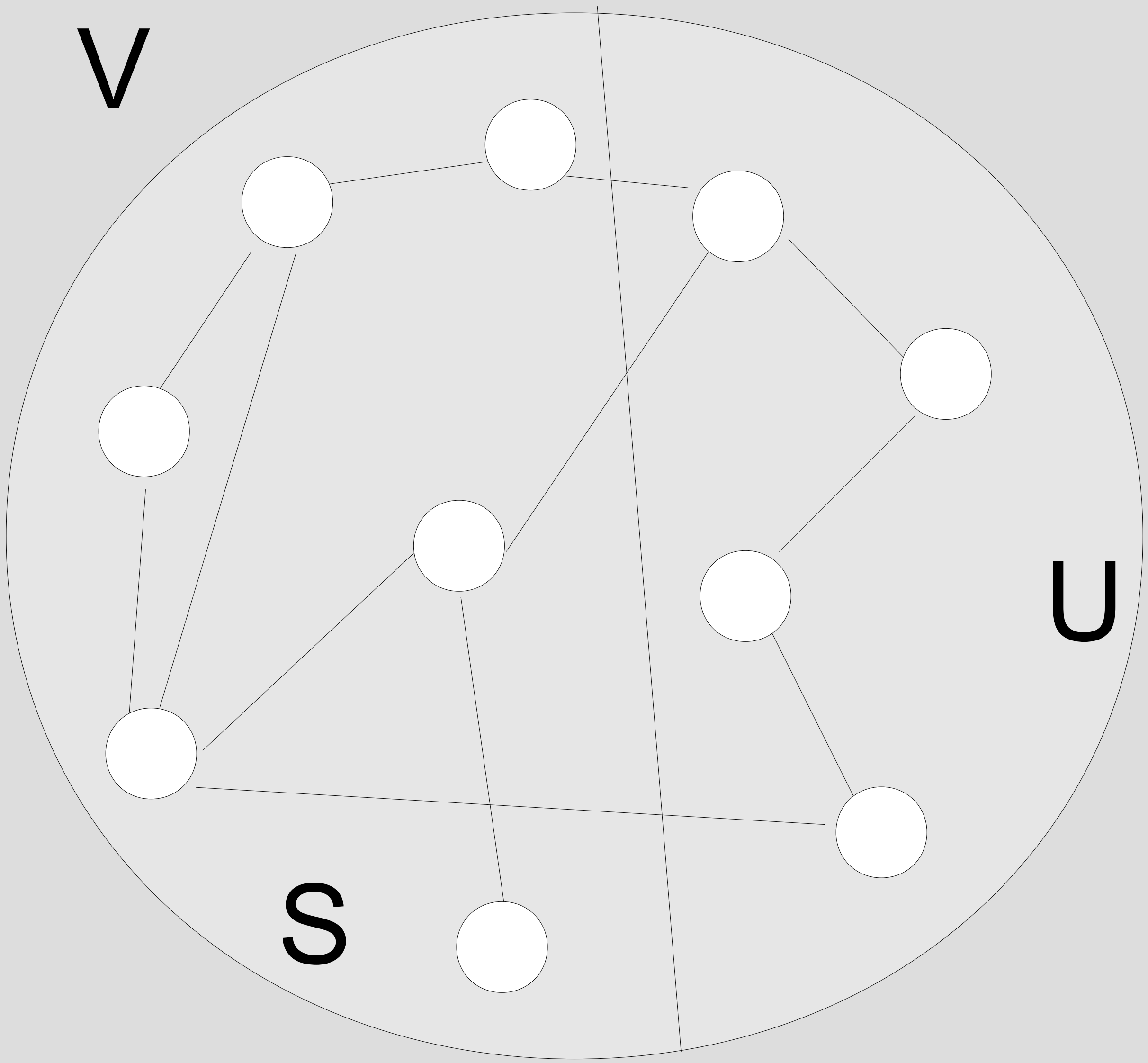
Gene Function Prediction Problem



Input:

- *V* genes
- *W* symmetric matrix

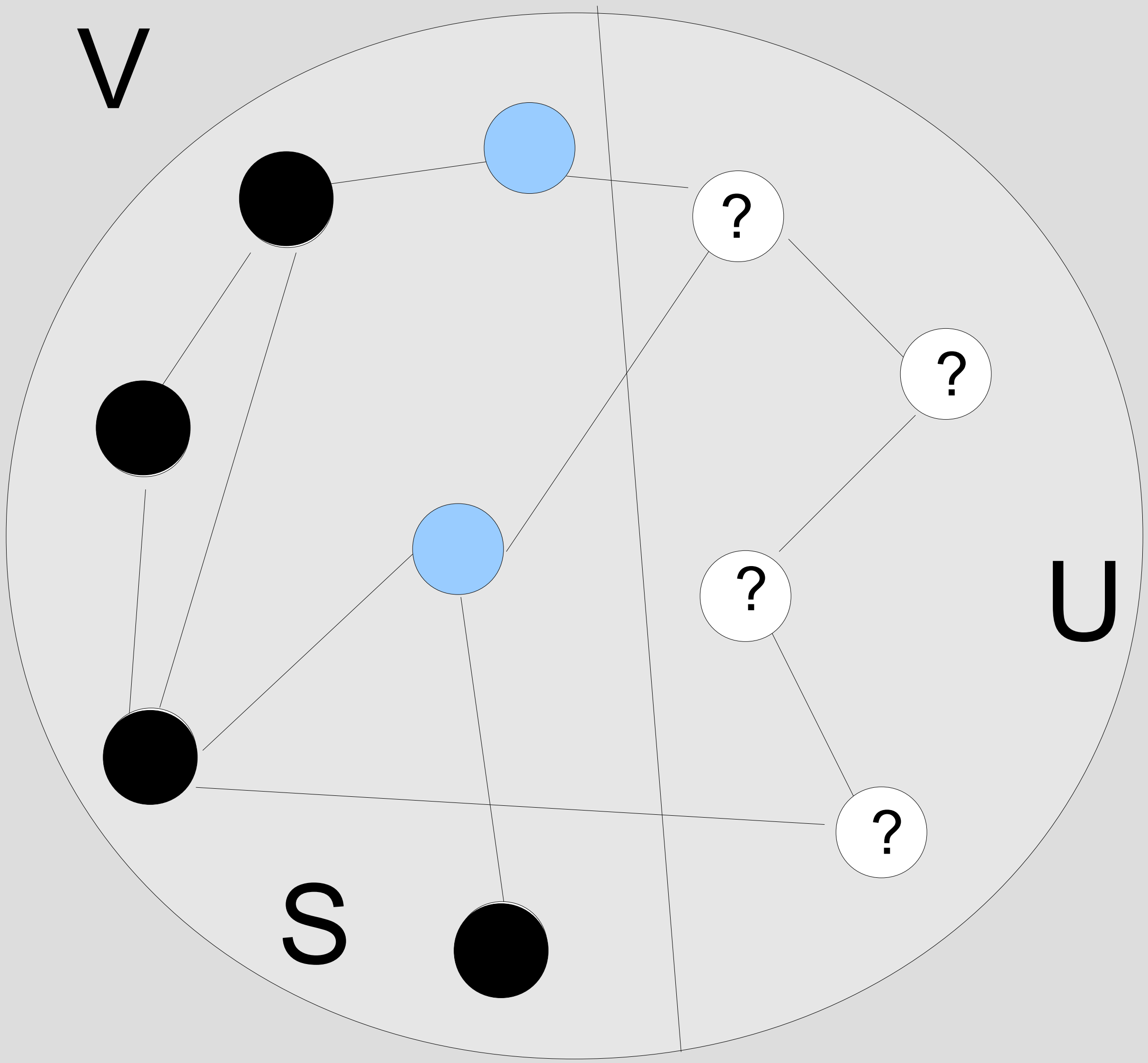
Gene Function Prediction Problem



Input:

- *V* genes
- *W* symmetric matrix
- *S*, *U* bipartition of *V*
  - *S* labeled genes
  - *U* unlabeled genes

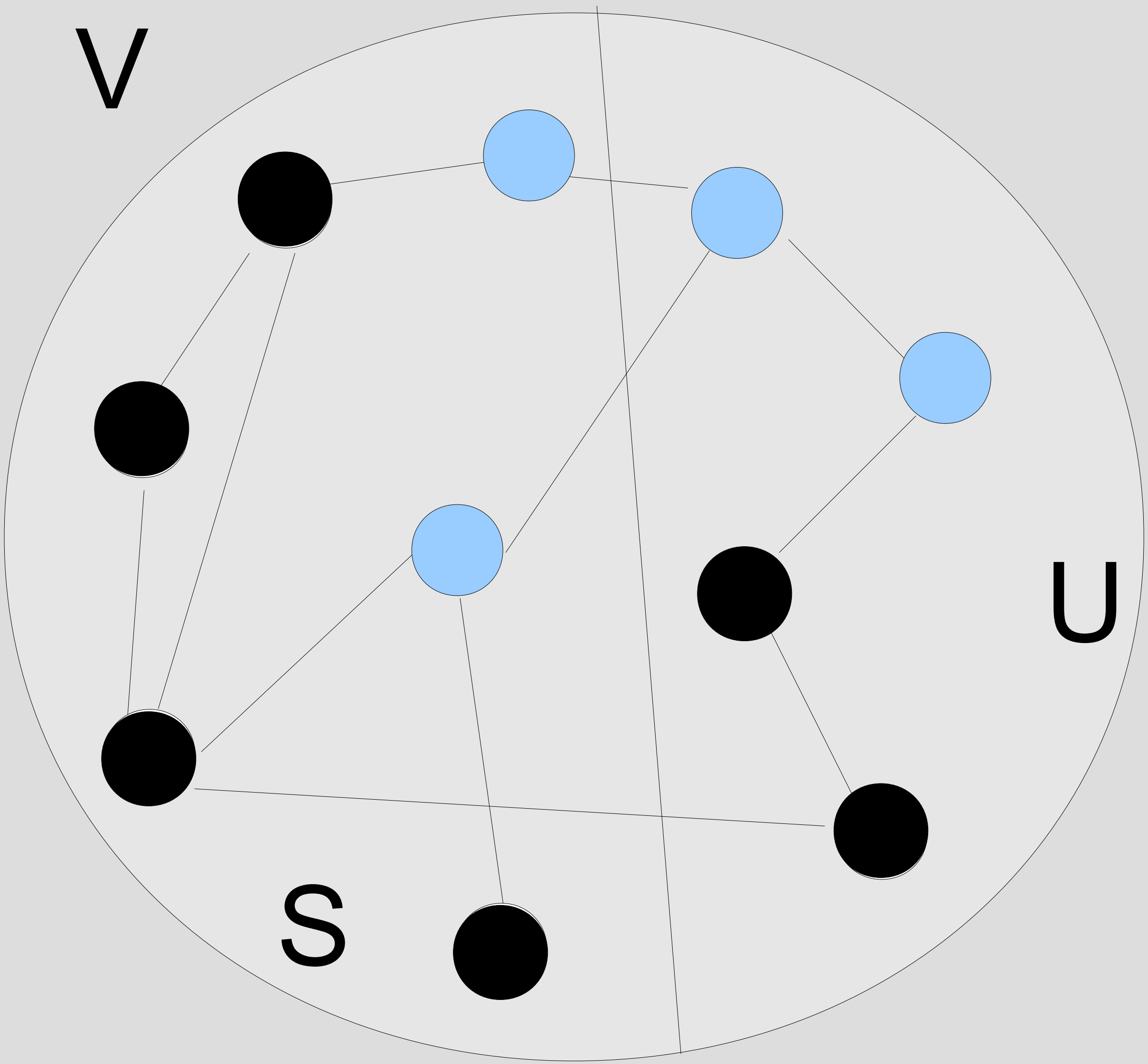
Gene Function Prediction Problem



Input:

- V genes
- W symmetric matrix
- S, U bipartition of V
  - S labeled genes
  - U unlabeled genes
- S^p, S^n bipartition of S

Gene Function Prediction Problem



Input:

- V genes
- W symmetric matrix
- S, U bipartition of V
  - S labeled genes
  - U unlabeled genes
- S^p, S^n bipartition of S

Output:

- U^p,U^n bipartition of U

Machine learning methods for GFP

- Inductive methods
  - Learn a model to infer functions for all genes
    - Support Vector Machines [Lanckriet et al 2004]
- Transductive methods
  - Infer functional predictions only for genes in test set
    - MRF [Deng et al 2002],
    - Neural networks [Karaoz et al 2003],
    - Functional Linkage Networks [Marcotte 1999]
    - Label propagation [Zhu et al 2003, Mostafavi 2008-2010].

Gene Annotation using Integrated Networks (GAIN)

- Karaoz et al. (2003)
- Discrete Hopfield network

Genes → Neurons

Similarity → Weights

(thresholds = 0)

Bipartitions → Initial state

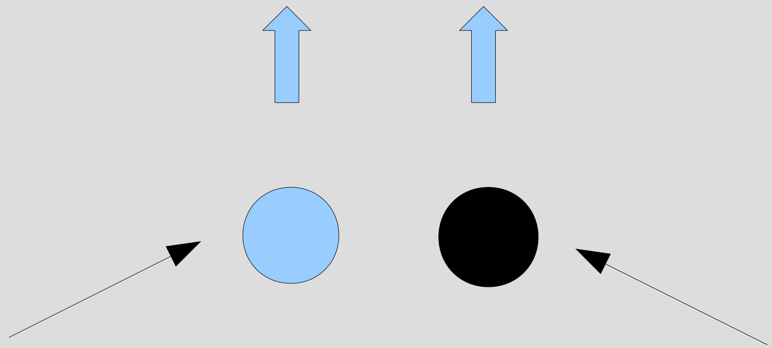
↓ Dynamics

Inference ← Equilibrium state

GAIN

- Initial state for each neuron *i* :

$x_i(0) = \quad 1, \quad -1, \quad 0 \leftarrow$?, Unlabeled



positive label negative label

$$x_i(t+1) = Sgn\left(\sum_{j=1}^{i-1} w_{ij}x_j(t+1) + \sum_{k=i+1}^{n} w_{ik}x_k(t)\right)$$

GAIN

- Energy function

$$E(x) = -\frac{1}{2} \cdot \sum_{i=1}^{n} x_i \left(\sum_{j \in 1}^{n} x_j w_{ij} \right)$$

- *E* is a monotonic decreasing function:

$$E(x(0)) \geq E(x(1)) \geq E(x(2)) \geq \ldots \geq \mathrm{E}(x(t)) \geq \ldots$$

- The equilibrium state $\tilde{x} = (\tilde{s}, \tilde{u})$ characterizes the bipartition of *U*

$$U^p = \{i \in U \mid u_i = 1\}$$

$$U^n = \{i \in U \mid u_i = -1\}$$

Drawbacks of GAIN

- All nodes may be updated
  - prior knowledge might not be preserved
- ”Same relevance” for positive and negative examples
  - Data imbalance not managed

Drawbacks of GAIN

- GAIN tries to find a global minimum $\tilde{x}$ of E assuming that the initial state $\bar{s}$ of labeled nodes is a part of $\tilde{x}$, *i.e.*

$$\tilde{x} = (\bar{s}, \tilde{u})$$

- In many cases $\bar{s}$ is not a part of a minimum
  - No coherence with the prior knowledge

Parametrized Hopfield network

GAIN:

- *Positive labels* := *1*
- *Nagative labels* := *-1*
- *Thresholds* := $\underline{0}$

Our approach:

- *Positive labels* := $\sin\alpha$
- *Nagative labels* := $-\cos\alpha$
- *Thresholds* := $\underline{\gamma}$

parameters to be learned!

Parametrized DHN: $< W, \underline{\gamma}, \alpha >$

Parametrized Hopfield network

- $H = \langle W, \underline{\gamma}, \alpha \rangle$ DHN on nodes in V
  - W connection matrix
  - $\underline{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_n)$ vector of activation thresholds
  - $\alpha \in \,]0, \pi/2[$, neuron values are $\sin\alpha$, $-\cos\alpha$
- In GAIN
  - $\underline{\gamma} = \underline{0}$
  - $\alpha = \pi/4$

Sub Network

Two Subnetworks: $H|_{S,U^p}$ and $H|_{U,S^p}$

Deal with Data Imbalance and prior knowledge ”coherence”

Preserve prior knowledge

Sub-network property

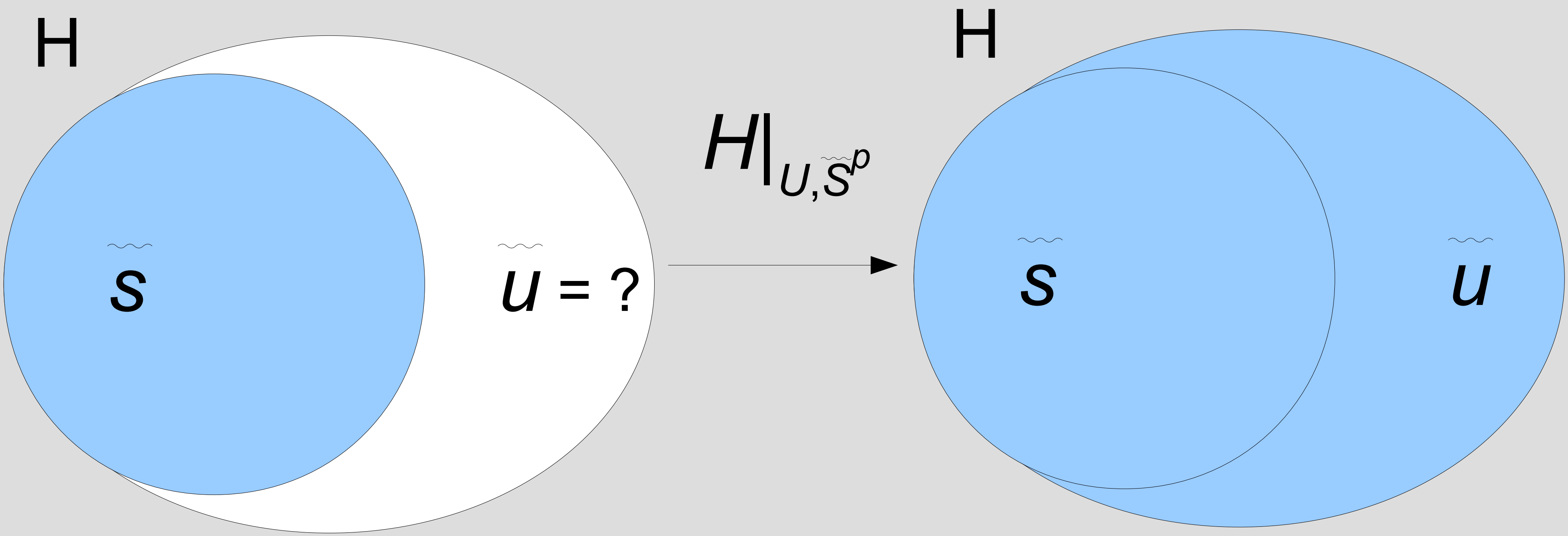
- Given
  - DHN $H < W, \underline{\gamma}, \alpha >$ with neurons V
  - S, U bipartition of V
  - S^p, S^n bipartition of S
  - U^p, U^n bipartition of U

It holds: if $\tilde{x} = (\tilde{s}, \tilde{u})$ is an energy global minimum H,

then $\tilde{u}$ is an energy global minimum of $H|_{U,S^p}$

Sub-network property

- Having a part $\tilde{s}$ of a minimum of energy of H, it's possible to discover the hidden part $\tilde{u}$ by minimizing the energy of $H|_{U,\tilde{S}}^{p}$

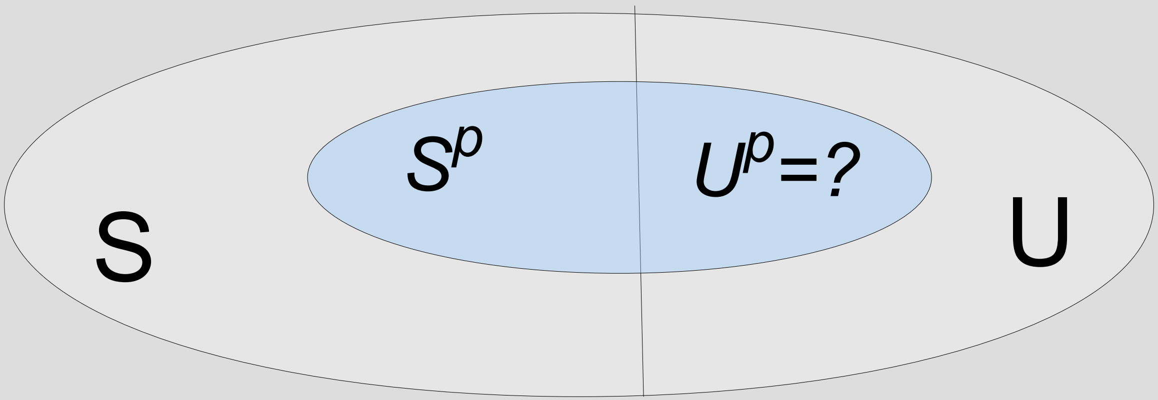


Sketch of COSNet

- INPUT: W similarity matrix; S, U bipartiton of V; S^p, S^n bipartiton of S
- OUTPUT: U^p, U^n bipartition of U

1. Generate a temporary solution U^p, U^n
2. Find the couple (α, γ) such that the initial state of the network $H|_{S,U^p}$ is as close as possible to an equilibrium state
3. Extend the parameters (α, γ) to the network $H|_{U,S^p}$
4. Run the network $H|_{U,S^p}$

Generating a temporary solution



p_s positive rate in S

p_u positive rate in U

Procedure:

- Generate k according to binomial distribution B($|U|, \frac{|S^p|}{|S|}$)
- U^p := k elements randomly chosen in U
- $U^n := U \setminus U^p$

FACT:

$$\frac{|S^p|}{|S|} = \underset{x}{argmax}\, Prob\left\{ p_u = x \,\middle|\, p_s = \frac{|S^p|}{|S|} \right\}$$

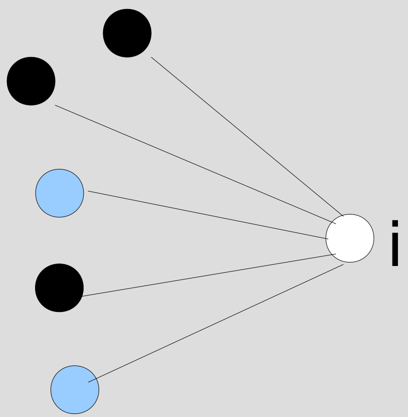
Finding the optimal parameters

ECML PKDD 2011

S $\longrightarrow$ Set of labeled points in R^2

L^p : positive points

L^n : negative points

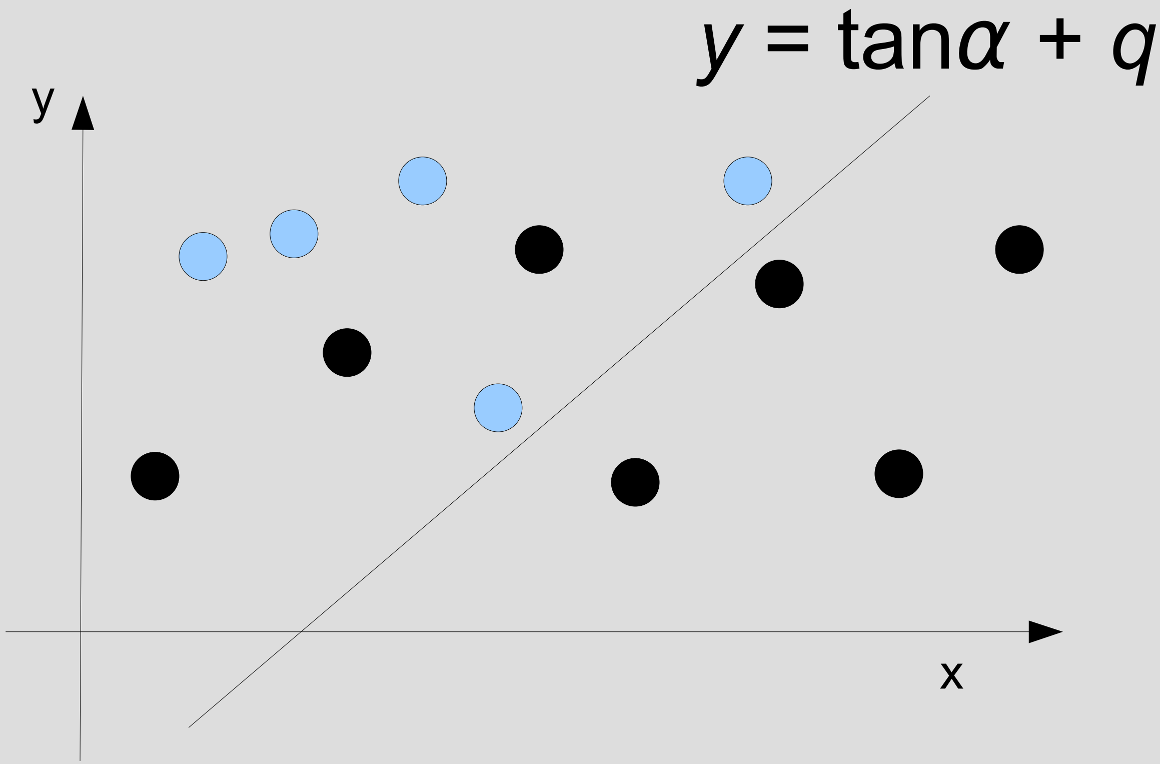


$$P_i \equiv \left(\sum_{j\in S^p \cup U^p} w_{ij} \,,\; \sum_{j\in S^n \cup U^n} w_{ij} \right)$$

Label of P_i = label of i

AIM: "optimal" separation of L^p from L^n by a straight line $y = \tan\alpha + q$ according to the Fscore criterion

Finding the optimal parameters

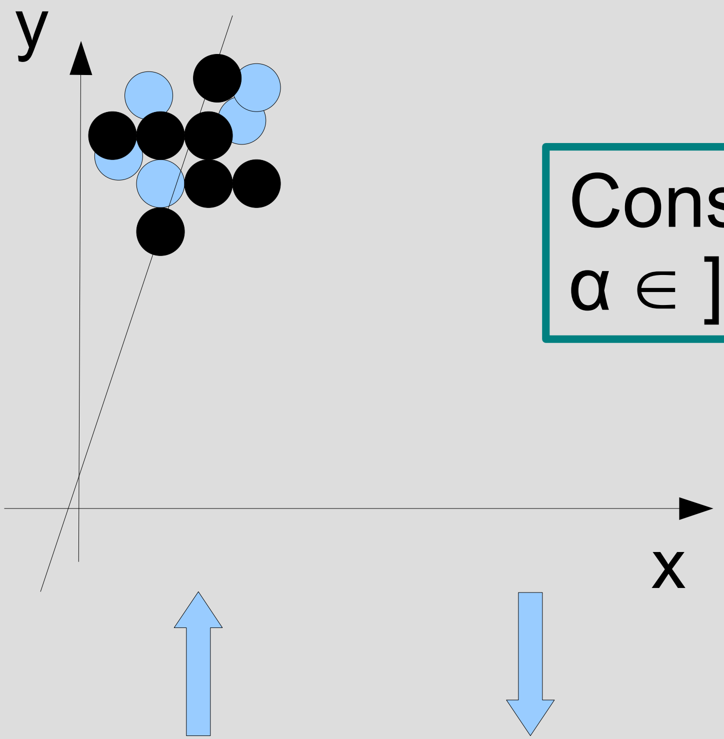


$\gamma = -q \cos\alpha$

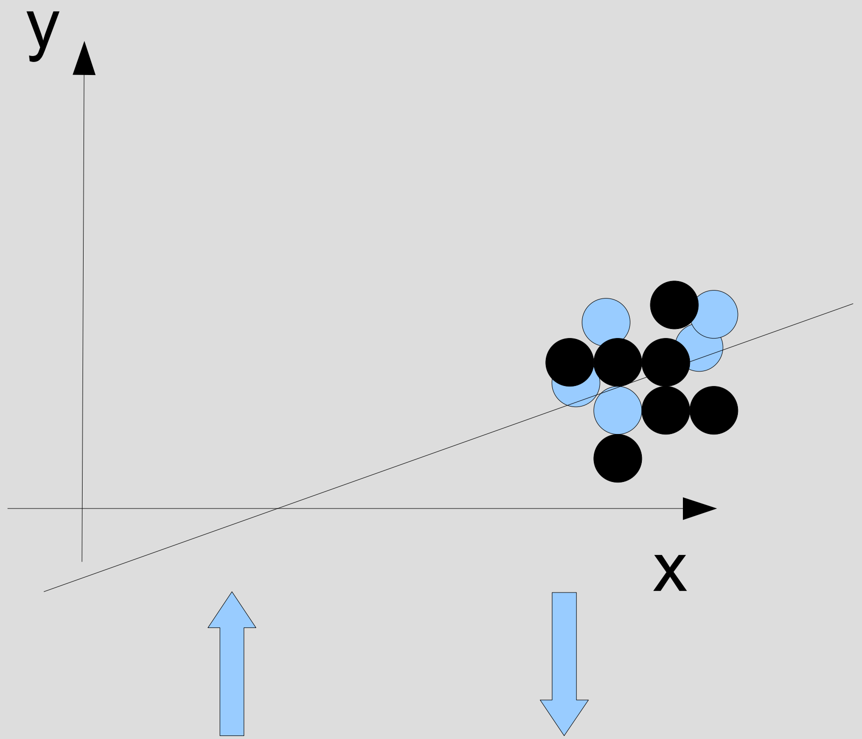
as first approach we chose the same γ for each neuron

Fact: Fscore (opt) = 1 ⟺ the corresponding state of $H|_{S,U^p}$ is an *equilibrium point*

Data imbalance management



Constraint:
$\alpha \in]0, \pi/2[$



$|S^p| << |S^n|, \quad \alpha > \pi/4$

$|\sin\alpha| > |\cos\alpha|$

$|S^p| >> |S^n|, \quad \alpha < \pi/4$

$|\sin\alpha| < |\cos\alpha|$

Finding the final solution

- Dynamics of the sub-network $H|_{U,S^p}$ with the found parameters until fixed point $\tilde{u}$ is reached
- Infer bipartition of U as follows:
  - $U^p = \{i \in U \mid \tilde{u}_i = \sin\alpha\}$
  - $U^n = \{i \in U \mid \tilde{u}_i = -\cos\alpha\}$

Results

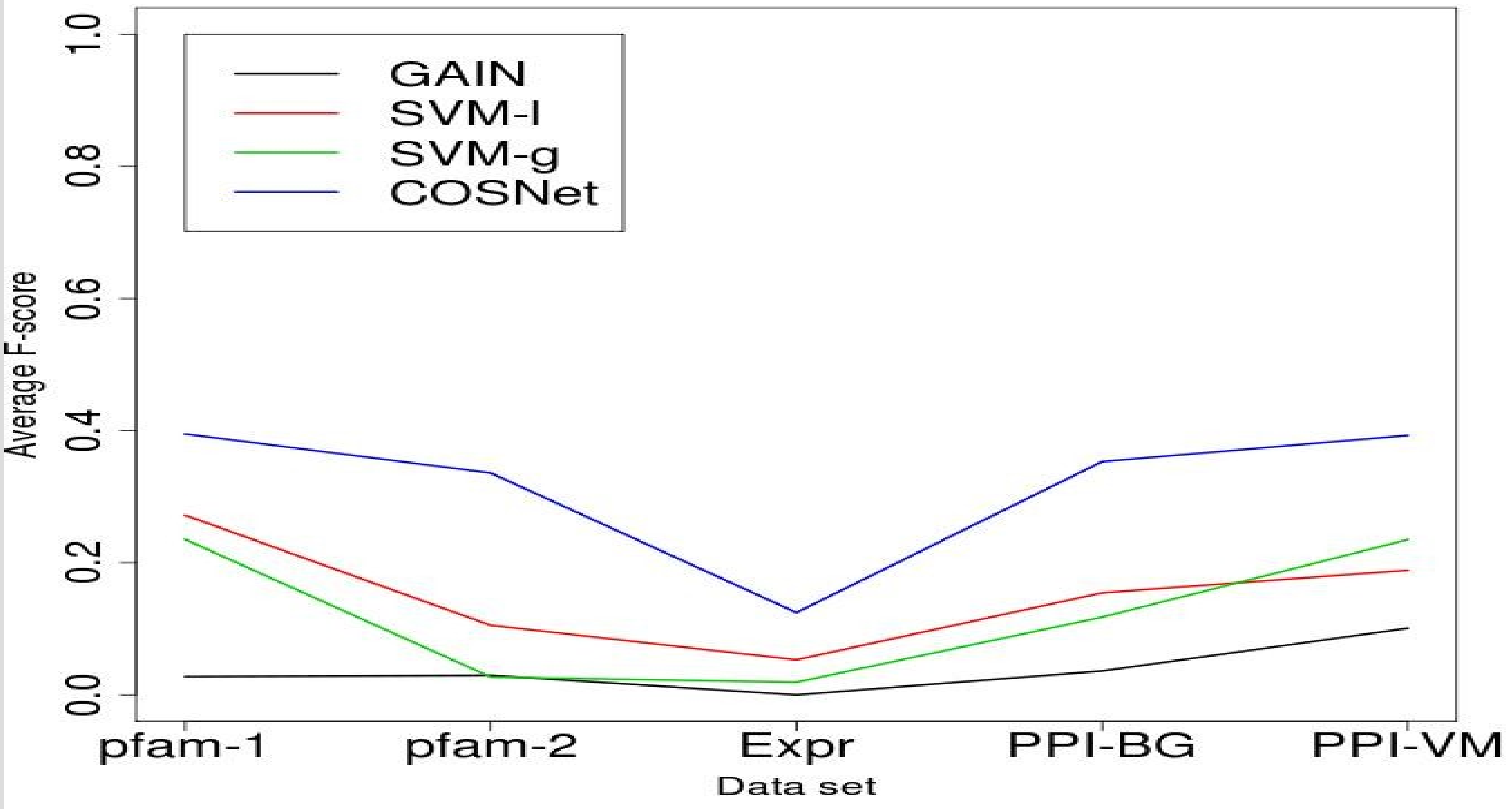
- Experimental setup:
  - 5 data yeast data sets
  - **Genome wide**
  - FunCat taxonomy, ontology wide
  - 10-folds Cross Validation
  - FLAT approach (no hierachical information)
  - F-score performance measure

Results

- Average F-score genome wide - ontology wide

Data set	*GAIN*	*SVM - l*	*SVM - g*	*COSNet*
Pfam-1	0.0277	0.2722	0.2355	0.3892
Pfam-2	0.0296	0.1054	0.0270	0.3233
Expr	0	0.0531	0.0192	0.0957
PPI-BG	0.0362	0.1546	0.1178	0.3486
PPI-VM	0.1009	0.1888	0.2351	0.3844

Results



GAIN
SVM-l
SVM-g
COSNet
Average F-score
0.0
0.2
0.4
0.6
0.8
1.0
pfam-1
pfam-2
Expr
PPI-BG
PPI-VM
Data set

Conclusions

- We developed a Cost-Sensitive method based on neural network for predicting labels in graph
- Effective in managing high imbalanced data
- Better performance w.r.t. the state-of-the-art methods
- The time complexity $\mathcal{O}(|S|*\log|S| + |W|)$ allows the application to large scale data

Ongoing developments

- Increase the number of parameters
  - Different threshlolds for neurons or different slopes
- Multi task extension
  - Use hierarchical relationship between terms
- Data Source integration