PREDICTION OF PROTEIN-PROTEIN INTERACTIONS USING PATTERN RECOGNITION TECHNIQUES

Ömer Nebil YAVEROĞLU

Middle East Technical University

Computer Engineering Department

June 2009

Outline

- Problem Definition
- Dataset Construction
- Previous Work Presented in HIBIT'2009
- Application and Results of …
  - Using Different Scoring Mechanisms during the Dataset Construction
  - SVM Classification after Relieff Feature Selection
  - Naive Bayes Classification
  - K-Nearest Neighborhood Classification
  - Decision Trees
  - Random Forest Classification
- Conclusion

Problem Definition

Performing protein-protein interaction prediction …

- based on phylogenetic profiles of proteins
- of the organism *Saccharomyces Cerevisiae* (baker's yeast)
- applying pattern recognition techniques

Dataset Construction

- Protein-protein interaction (PPI) data collected from *Saccharomyces Cerevisiae*
- 11698 PPI pairs in this data.
- 5849 of these represent interactions and the other 5849 of these represent non-interactions.
- These interaction information is extracted from **Database of Interacting Proteins** (DIP)

Dataset Construction

- The homolog's of each protein is searched in a set of 450 fully sequenced genomes in other words 450 different organisms
- Two phylogenetic profiles are combined to get a combined profile
- The combining process is performed by checking the existence of homolog's of the proteins in the PPI pairs for each of the 450 fully sequenced organisms

Dataset Construction

- The scoring mechanism used for the construction of the first dataset is as follows:
  - If homolog's of both proteins exist in the organism → 2
  - If homolog's of both proteins don't exist in the organism → 0
  - If homolog of only one of the proteins exists in the organism → -1
- The combining process is performed with four different scoring mechanisms which will be discussed later.
- The dataset used in all methods mentioned in this study is a 11698 by 450 matrix consisting of the three scores which comes from the scoring mechanism selected (2 , 0 and -1 for the above example).

Previous Work Presented in HIBIT'2009

- Principle Component Analysis
- Multidimensional Scaling
- K-means Clustering
- Support Vector Machines

… are applied on the dataset formed from the phylogenetic profiles of proteins.

Previous Work Presented in HIBIT'2009

- Protein-protein interaction prediction is performed with 64.0026% accuracy with the application of Support Vector Machines
- The results show that it is possible to perform PPI prediction from the phylogenetic profiles of proteins
- The future work is to find the most suitable classification technique for the prediction of the interaction

Applied Pattern Recognition Methods

- The pattern recognition techniques applied for determining protein interaction are …
  - Relieff Feature Extraction
  - Naive Bayes Classification
  - K-Nearest Neighborhood Classification
  - Decision Trees
  - Random Forest Classification
- For testing the performance of the results, accuracy of classification is used.

$$accuracy = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}$$

Application of different scoring mechanisms

- The four scoring mechanisms used for combining phylogenetic profiles are:

	Both Exist	None Exist	One exist
Scoring 1	2	0	-1
Scoring 2	4	1	-2
Scoring 3	2	0	0
Scoring 4	8	0	-4

Application of different scoring mechanisms

- SVM Classification is performed on the datasets constructed this way.
- Radial Basis Function is used as the kernel of SVM.
- The results of the SVM classification are:

Training Dataset Size	Scoring 1	Scoring 2	Scoring 3	Scoring 4
500	62.0736 %	62.9041 %	58.3318 %	62.0736 %
1000	64.0026 %	64.2363 %	59.1045 %	64.0026 %
2000	61.3529 %	61.5075 %	58.1254 %	61.3529 %
4000	62.0681 %	62.9124 %	60.2884 %	62.0681 %

Application of Relieff Feature Extraction

- Evolution results with dependencies between the features of the dataset.
- Similar organisms cause similar feature information giving an emphasis a group of species that are close to each other in the evolution tree.
- SVM Classification is applied on the feature selected dataset producing the following results:

	Scoring 1	Scoring 2	Scoring 3	Scoring 4
Best Features	58.6100 %	59.9918 %	58.7750 %	60.1155 %
First15 Features	58.3316 %	58.6513 %	58.1254 %	54.6401 %

Application of Naive Bayes Classification

- Aim was to determine the true classes of protein pairs by a probabilistic classification method.
- Prior information is computed from the dataset by a frequentist approach.
- The accuracy results acquired by training different sizes of datasets with different scoring mechanisms are:

Training Size	Scoring 1	Scoring 2	Scoring 3	Scoring 4
1000	58.63 %	58.52 %	59.11 %	58.63 %
1500	58.46 %	57.88 %	56.65 %	58.46 %
2000	58.39 %	57.99 %	59.43 %	58.39 %
4000	57.44 %	57.33 %	60.15 %	57.44 %

Application of K-Nearest Neighborhood Classification

- Considering the protein pairs locally and performing the classification depending on their in between distances was another option to find the classification pattern in data.

- The accuracy results acquired by training different sizes of datasets with different scoring mechanisms are:

Training Size	Scoring 1	Scoring 2	Scoring 3	Scoring 4
1000	58.63 %	60.91 %	57.52 %	60.05 %
2000	59.67 %	59.89 %	55.90 %	59.57 %

Application of Decision Trees for Classification

- Aim was to see whether a graph-based predictive model would produce better results or not

- The accuracy results acquired by 10-fold cross validation with different scoring mechanisms are:

Scoring 1	Scoring 2	Scoring 3	Scoring 4
67.7894 %	67.7894 %	65.1308 %	67.5586 %

- Overall improvement in accuracy compared to SVM application is about 3%.

Application of Random Forest Classification

- The improvement acquired by decision trees showed that graph based classification methods work better for our aim.
- The accuracy results acquired by 10-fold cross validation with different scoring mechanisms are:

Scoring 1	Scoring 2	Scoring 3	Scoring 4
76.7567%	76.5174 %	Insufficient Memory	76.9362 %

- Good predictive performance achieved.
- Disadvantage is the huge memory requirements of the method

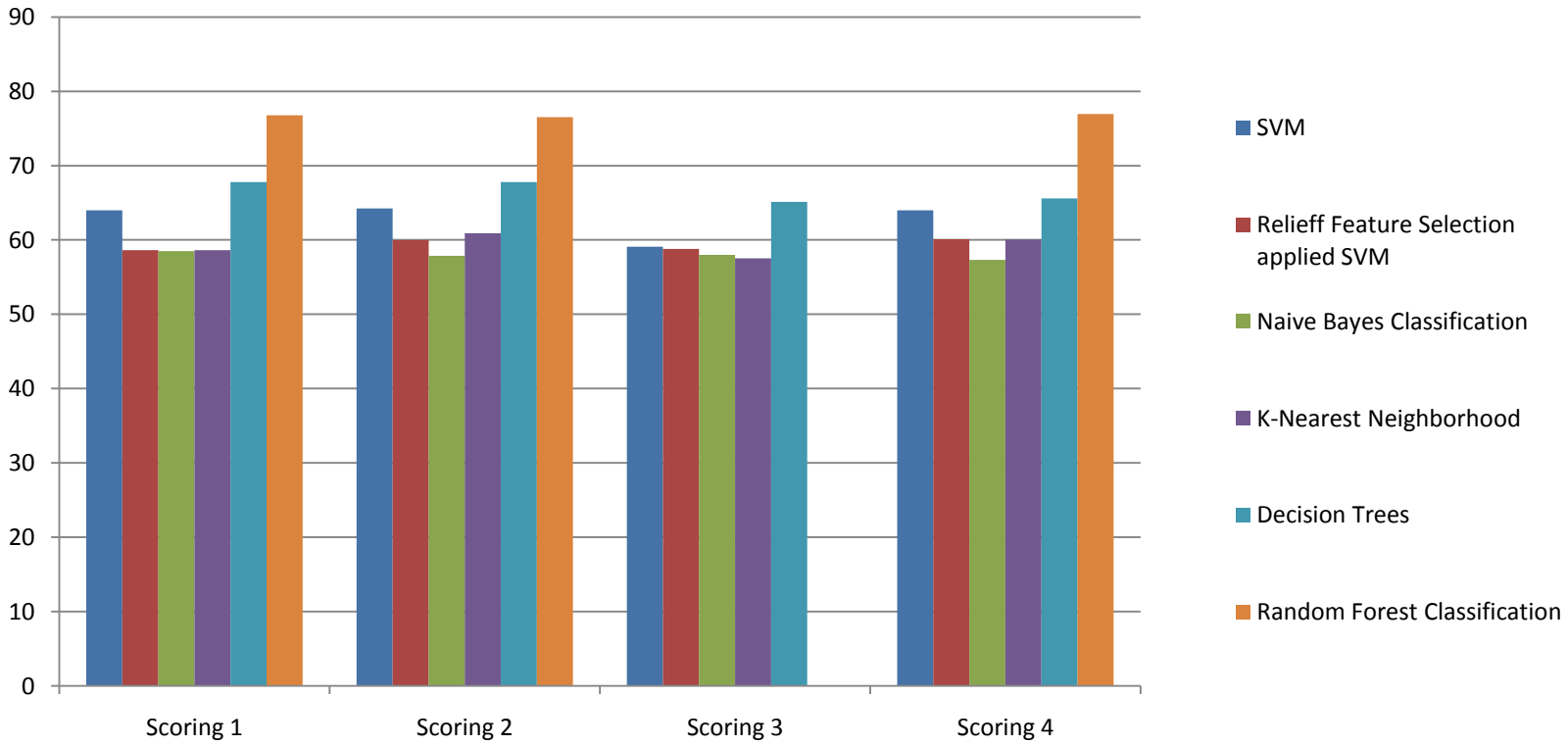
Conclusion

- Relieff Feature Extraction
- Naive Bayes Classification
- K-Nearest Neighborhood Classification
- Decision Trees
- Random Forest Classification

… are applied on the dataset formed from the phylogenetic profiles of proteins.

Conclusion

- A performance comparison of these methods can be seen from the bar chart below



Conclusion

- Protein-protein interaction prediction is performed with 76.9362% accuracy with the application of Random Forest Classification.
- It is shown that performing high accuracy protein interaction prediction is possible by just using primary sequences of proteins.
- In the future,
  - changes can be made on construction of the negative (non-interacting) samples
  - some other classification methods can be considered