Dimensionality Reduction for Protein Function Prediction

Zehra Cataltepe[1,*], Eser Aygun[1], Asli Filiz[1], Ozlem Keskin[2], Caner Komurlu[1],Yucel Altunbasak[3]
[1]Istanbul Technical University, Computer Engineering Department
Ayazaga, Sariyer, TR-34469, Istanbul, Turkey
[2]Koc University, College of Engineering, Chemical and Biological Engineering
Rumeli Feneri Yolu, Sariyer, TR-34450, Istanbul, Turkey
[3]Georgia Institute of Technology, School of Electrical and Computer Engineering
Atlanta, GA 30332-0250
*To whom correspondence should be addressed: cataltepe@itu.edu.tr

1. INTRODUCTION

Dimensionality reduction methods (see [1], for example) reduce the input feature dimensionality and result in faster classification algorithms due to smaller number of inputs. If noisy, irrelevant or redundant features are eliminated, then dimensionality reduction could also lead to better classification accuracy.

Previously, [2] compared a number of feature selection methods, based on entropy, t-statistics and chi^2 statistics and found out that dimensionality reduction helped with distinguishing genes for Acute Lymphoblastic Leukemia and ovarian cancer. [3] used manual or pairwise Fisher's Linear Discriminant Analysis to select features for identifying marker genes on a number of cancer data sets. [4] and [5] used sequence data together with Support Vector Machines for both feature and instance selection for protein function prediction.

In this study, we evaluate the effects of dimensionality reduction on protein function prediction. We consider the GO (Gene Ontology) Molecular Function first level categories and H. Pylori as the organism to collect the amino acid sequence data. We evaluate three different classifiers (Naïve Bayes, kNN, SVC) and three different dimensionality reduction methods (PCA, Fisher's LDA and FCBF (Fast Correlation-Based Filter) algorithm [6]).

2. DATA AND METHODS

We downloaded the fasta sequences for H. Pylori from http://expasy.org/sprot/hamap/HELPY.html There were a total of 564 amino acid sequences. The distribution of these amino acid sequences according to the GO Molecular Function categories are shown in Table 1. We eliminated GO Molecular Function categories with less than 10 sequences.

For features, we obtain physiochemical properties of the amino acid sequences using the PROFEAT software [7]. We also use the ClustalW [8] alignment scores between a sequence and all the training sequences. The dimensionality reduction algorithms that we consider are, PCA (Principal Component Analysis), Fisher's LDA (Linear Discriminant Analysis) and FCBF (Fast Correlation-Based Filter) algorithm. We measure the 10-fold cross validation accuracy of Naive Bayes (NB), kNN (k-Nearest Neighbor) and Support Vector Classifiers (SVC) to compare accuracy of classifiers that use features selected by different feature selection methods.

3. RESULTS

Features selected by FCBF: In our experiments, out of the 1447 PROFEAT features, 5 were selected by FCBF (see Table 2). In Table 2, AURM and AURV are features related to the Normalized Moreau-Broto autocorrelation. Out of 564 ClustalW features (i.e. sequences), 14 were selected by FCBF.

Classification Accuracies: Table 3 shows the average 10-fold cross validation accuracies when H. Pylori and PROFEAT features and H. Pylori and ClustalW features are used respectively. As seen in the tables the best accuracies for a single classifier are achieved when FCBF is used for all four cases. FCBF seems to work especially well for the Naive Bayes classifier. Although its performance is also quite good for kNN and SVC classifiers also. The reason why FCBF works very well with NB classifier may be the fact that NB assumes that each feature is independent from each each other and after elimination of redundant features, the remaining features are actually not correlated with each other. We think that FCBF works better in general for each classifier, because it uses an entropy based measure for redundancy and relevance. PCA or Fisher's LDA, on the other hand, use linear correlations. Entropy based measure is able to capture more general correlations between features and data and in between features.

Classification Time: On the average, FCBF takes less time to execute than the PCA and Fisher's LDA.

4. CONCLUSIONS

While feature selection could result in drastic falls in accuracy for a certain classifier when PCA or Fisher's LDA are used, there is either a very good improvement or very little drop in accuracy when FCBF is used. Therefore, when classification accuracies are averaged over all three classifiers (NB, KNN and SVC), FCBF performs better. Furthermore, in all the experiments, FCBF (usually with Naive Bayes as the classifier) results in the best accuracy. FCBF is also faster than PCA or Fisher's LDA. We did not observe a significant difference in classifier accuracies when either ClustalW alignment scores or PROFEAT features were used.

In the future, we would like to soften the FCBF algorithm, select more features and see the overlap between features when different organisms are used. It would be interesting to see if function is determined by different or similar features for each organism.

5. REFERENCES

1. Yang, Y. and Pedersen, J. O. 1997. A Comparative Study on Feature Selection in Text Categorization, *Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.

2. Liu, H., Li, J. and Wong, L. 2002. A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns, *Genome Informatics*, 13, 51-60.

3. Wang, J., Hellem, T., Jonassen, I. et. al. 2003. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data, *BMC Bioinformatics*, 4:60.

4. Al-Shahib, A., Breitling, R. and Gilbert, D. 2005. Feature Selection and the Class Imbalance Problem in Predicting Protein Function from Sequence, *Applied Bioinformatics*, 4:3, 195-203.

5. Al-Shahib, A., Breitling, R. and Gilbert, D. 2005. FrankSum: new feature selection method for protein function prediction, *Int J Neural Syst.*, 15:4, 259-275.

6. Yu, L. and Liu, H. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research* , 5, 1205-1224.

7. Li, Z. R., Lin, H. H., Han, L. Y. et al. 2006. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, *Nucleic Acids Research,* 34, W32-W37.

8. Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research,* 22:22, 4673-4680.

TABLE1: No of Sequences in Each Class.

GOID	**Molecular Function**	**#seq.**
0003774	motor activity	12
0030528	transcr. reg. act.	15
0004871	signal transd. act.	23
0003674	molec. func. (others)	26
0005215	transporter activity	72
0005488	binding	280
0003824	catalytic activity	334

TABLE2. PROFEAT features selected by FCBF

PROFEAT ID	**Feature name**
F5.1.1.1	Composition of Hydrophobicity(1)
F6.1.2.26	Seq. Order Coupling Number2(26)
F5.1.6.3	Composi. of Secondary structure(3)
F1.2.1.43	DD Dipeptide composition(%)
F2.1.8.15	AURM(15)
F2.1.6.6	AURV(6)

TABLE3. Classification Accuracies Using PROFEAT and ClustalW Features.

	Naïve Bayes		kNN		SVC	
Features	PROFEAT	ClustalW	PROFEAT	ClustalW	PROFEAT	ClustalW
all features	3.40 ± 0.34	2.05 ± 0.20	46.07 ± 2.11	47.18 ± 1.82	45.13 ± 1.17	45.13 ± 1.17
FCBF	**52.75** ± 1.85	**50.95** ± 2.20	41.79 ± 1.98	44.81 ± 2.21	43.28 ± 2.55	45.13 ± 1.17
PCA	44.57 ± 1.54	35.87 ± 1.55	34.72 ± 2.28	25.77 ± 1.30	25.67 ± 5.19	45.13 ± 1.17
FISHER	9.62 ± 1.02	3.25 ± 0.83	31.92 ± 1.78	7.99 ± 1.63	45.13 ± 1.17	45.13 ± 1.17