Textual Information for Predicting Functional Properties of the Genes

Oana Frunza and Diana Inkpen
School of Information Technology and Engineering
University of Ottawa Ottawa, ON, Canada, K1N 6N5
{ofrunza,diana}@site.uottawa.ca

1 Overview

This paper is focused on determining which proteins affect the activity of *Aryl Hydrocarbon Receptor (AHR*) system when learning a model that can accurately predict its activity when single genes are knocked out. Experiments with results are presented when models are trained on a single source of information: abstracts from Medline (http://medline.cos.com/) that talk about the genes involved in the experiments. The results suggest that AdaBoost classifier with a binary bag-of-words representation obtains significantly better results.

2 Task Description and Data Sets

The task that we address is a biology-specific task considered a competition track for KDDCup2002 (http://www.biostat.wisc.edu/~craven/kddcup/winners.html).

The organizers of the KDD Cup competition provided data obtained from experiments performed on a set of yeast strains in which each strain contains a single gene that is knocked out (a gene sequence in which a single gene is inoperative). Each experiment had associated a discretized value of the activity of the AHR system when a single gene was knocked out. 3 possible classes describe the systems' response. The "**nc**" label indicates that the activity of the hidden system was not significantly different than the baseline (the wild-type yeast); the "**control**" label indicates that the activity was significantly different than the baseline for the given instance, and that the activity of another hidden system (the control) was also significantly changed compared to its baseline; the "**change**" label shows that the activity of the hidden system was significantly changed, but the activity of the control system was not significantly changed.

The organizers of the KDD Cup evaluate the task as a two-class problem with focus on the positive class. The first definition is called the "**narrow**" definition of the positive class and it is specific to the knocked-out genes that had an AHR-specific effect. In this case the positive class is defined by the experiments in which the label of the system is "*change"* and the negative examples are the experiments that consist of those genes with either the "*nc*" or the "*control*" label. The second definition consists of those genes labeled with either the "*change*" or the "*control*" label. The negative class consists of those genes labeled with the "*nc*" label. The second partitioning corresponds to the "**broad"** characterization of the positive class genes that affect the hidden system.

The area under the Receiver Operating Characteristic (ROC) - AUC curve is chosen as an evaluation measure. The global score for the task will be the summed AUC values for both the "narrow" and the "broad" partition of the data.

The sources of information provided by the organizers of the task contain: hierarchical information about the function and localization of the genes; relational information describing the protein-protein interactions; and textual information in abstracts from Medline that talk about the genes. Some characteristics of the data need to be taken into consideration in order to make suitable decisions for choosing the trainable system/classifier, the representation of the data, etc. Missing information is a characteristic of the data set. Not all genes had the location and function annotation, the protein-protein interaction information, or abstracts associated with the gene name. Besides the missing information, the high class imbalance is another fact that needs to be taken into account.

From the data that was released for the KDD competition we run experiments only with the genes that had associated abstracts. Table 1 presents a summary of the data sets used in our experiments after considering only the genes that had abstracts associated with them. The majority of the genes had one abstract, while others had as many as 22 abstracts.

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 110–111,
Columbus, Ohio, USA, June 2008.

Table 1. Summary of the data for our experiments with the two definitions of the positive class. In brackets are the original sizes of the data sets.

Data set	Narrow		Broad	
	Pos	Neg	Pos	Neg
Training	24 (37)	1,435 (2,980)	51 (83)	1,408 (2,934)
Test	11 (19)	715 (1,469)	30 (43)	696 (1,445)

3 Related Work

Previous research on the task was done by the teams that participated in the KDD Cup 2002. The textual information available in the task was considered as an auxiliary source of information and not the primary one, as in this article.

The winners of the task, Kowalczyk and Raskutti (2002) used the textual information as additional features to the ones extracted from other available information for the genes. They used a "bag-of-words" representation, removed stop words and words with low frequency. They used Support Vector machine (SVM) as a classifier.

Krogel *et. al*. (2002) used the textual information with an information extraction system in order to extract missing information (function, localization, protein class) for the genes in the released data set.

Vogel and Axelrod (2002) used the Medline abstracts to extract predictive keywords, and added them to their global system.

Our study investigates and suggests a textual representation and a trainable model suitable for this task and similar tasks in the biomedical domain.

4 Method

The method that we propose to solve the biology task is using Machine Learning (ML) classifiers suitable for a text classification task and various feature representations that are known to work well for data sets with high class imbalance. The task becomes a two-class classification: "**Positive**" versus "**Negative**", with a "**narrow**" and "**broad**" definition for the positive class. As classification algorithms we used: Complement Naive Bayes (CNB), AdaBoost, and SVM all from the Weka toolkit (http://www.cs.waikato.ac.nz/ml/weka/). Similar to the evaluation done for the KDD Cup, we consider the sum of the 2 AUC measures for the definitions of the positive class as an evaluation score. The random classifier with an AUC measure of 0.5 is considered as a baseline.

As a representation technique we used binary and frequency values for features that are: words extracted from the abstracts (bag-of-words (BOW) representation), UMLS concepts and UMLS phrases identified using the MetaMap system (http://mmtx.nlm.nih.gov/), and UMLS relations extracted from the UMLS metathesaurus. We also ran experiments with feature selection techniques.

Table 2 presents our best results using AdaBoost classifier for BOW, UMLS concepts, and UMLS relations representation techniques. "B" stands for binary and "Freq" stands for frequency counts.

Table 2. Sum of the AUC results for the two classes without feature selection.

Representation	**AdaBoost (AUC) Narrow**	**AdaBoost (AUC) Broad**	**Sumed AUC**
BOW_B	**0.613**	**0.598**	**1.211**
BOW_Freq	0.592	0.557	1.149
UMLS_B	0.571	0.607	1.178
UMLS_Freq	0.5	0.606	1.106
UMLS_Rel_B	0.505	0.547	1.052
UMLS_Rel_Freq	0.5	0.5	1

5 Discussion and Conclusion

Looking at the obtained results, a general conclusion can be made: textual information is useful for biology-specific tasks. Not only that it can improve the results but can also be considered a stand-alone source of knowledge in this domain. Without any additional knowledge, our result of 1.21 AUC sum is comparable with the sum of 1.23 AUC obtained by the winners of the KDD competition.

References

Adam Kowalczyk and Bhavani Raskutti, 2002. *One Class SVM for Yeast Regulation Prediction,* ACM SIGKDD Explorations Newsletter, Volume 4, Issue 2, pp. 99-100.

Mark A Krogel, Marcus Denecke, Marco Landwehr, and Tobias Scheffer. 2002. *Combining data and text mining techniques for yeast gene regulation prediction: a case study,* ACM SIGKDD Explorations Newsletter, Volume 4, Issue 2, pp. 104-105.

David S. Vogel and Randy C. Axelrod. 2002. *Predicting the Effects of Gene Deletion,* ACM SIGKDD Explorations Newsletter, Volume 4, Issue 2, pp. 101-103.