In *Proceedings of the IEEE International Conference on Data Mining.* 2003.

Effectiveness of Information Extraction, Multi-Relational, and Semi-Supervised Learning for Predicting Functional Properties of Genes

Mark-A. Krogel
University of Magdeburg
FIN/IWS, PO Box 4120
39016 Magdeburg, Germany
krogel@iws.cs.uni-magdeburg.de

Tobias Scheffer
Humboldt University, Berlin
Department of Computer Science,
10099 Berlin, Germany
scheffer@informatik.hu-berlin.de

Abstract

We focus on the problem of predicting functional properties of the proteins corresponding to genes in the yeast genome. Our goal is to study the effectiveness of approaches that utilize all data sources that are available in this problem setting, including unlabeled and relational data, and abstracts of research papers. We study transduction and co-training for using unlabeled data. We investigate a propositionalization approach which uses relational gene interaction data. We study the benefit of information extraction for utilizing a collection of scientific abstracts. The studied tasks are KDD Cup tasks of 2001 and 2002. The solutions which we describe achieved the highest score for task 2 in 2001, the fourth rank for task 3 in 2001, the highest score for one of the two subtasks and the third place for the overall task 2 in 2002.

1 Introduction

A principle challenge of bioinformatics is to generate models which describe the relation between genetic information and corresponding cellular processes. Such models have to explain – and can be derived from – available experimental data. We focus on a set of related problems of functional genomics. We aim at predicting the high-level function and the localization of the protein corresponding to a given yeast gene, and at predicting whether a given gene is involved in the regulation of the aryl hydrocarbon receptor (AhR) signaling pathway (for the available data, this has been determined in a gene deletion experiment).

The data which we use have been provided for the KDD Cups 2001 [3] and 2002 [4]. Besides attributes such as function, localization, and protein class for each gene, the data include relational gene interaction information and abstracts of relevant research papers in MEDLINE. Focusing on the goal of building as accurate a model of the biological system as possible, we explore the effectiveness of several approaches that allow us to utilize these available sources of unlabeled, multi-relational, and textual data.

Approaches to utilizing relational data (in our case, gene interactions) in machine learning algorithms include inductive logic programming and *propositionalization* (originating from ILP) – *i.e.,* casting of a controlled amount of relational information into attributes (*e.g.,* [7, 9]). Because of scalability issues, we follow the latter approach. For the focused problem, unlabeled data is inexpensive and readily available. Approaches to semi-supervised learning include transduction [6] and the co-training algorithm [1]. Abstracts of scientific papers in the MEDLINE collection contain information that can be helpful for model building. Algorithms have been studied that extract information from literature [10, 5], based on dictionary-based extractors (*e.g.,* [5]), rule learners, or hidden Markov models [10].

This paper is organized as follows. We discuss the application and experimental setting in Section 2, and our propositionalization approach in Section 3. Section 4 focuses on our studies on text mining, while Section 5 presents results on semi-supervised learning. A discussion of our competition results and lessons learned is sketched in Section 6.

2 Problem Description

Our task is to predict properties of the proteins corresponding to a given yeast gene. These properties are (1) one (or several) of 15 categories of protein functions, (2) the localization (one of 15 different parts of the cell), and (3) the involvement in the regulation of the AhR signaling pathway. The AhR is a basic helix-loop-helix transcription factor with the ability to bind both synthetic chemicals such as dioxins and naturally-occurring phytochemicals, sterols and heme breakdown products. This receptor plays an important developmental and physiological role. Problems (1) and (2) have been addressed in KDD Cup 2001 whereas problem (3) is one of the tasks of KDD Cup 2002.

The available training data for problem (1) and (2) contains 862 training and 381 test instances [3]. Besides attributes that characterize the individual gene, a relation specifies which genes interact with one another. We have to limit our comparative studies to the three most frequent classes as the minority classes contain too few instances to obtain performance estimates.

The data for problem (3) has been obtained in experiments with yeast strains using a gene deletion array [4]. Each instance in the data set represents a trial in which a single gene is knocked out and the activity of a target system (AhR signaling) is measured. We distinguish genes whose deletion affects the target system (class “change”), affects the entire cell (“control”), or does not have an effect (“no change”). We learn two discriminators: *change* vs. *control* and *no change* (“narrow positive class” problem) and *change* and *control* vs. *no change* (“broad positive class”).

The data contains 3,018 training and 1,489 test examples. 2,934 fall into the class *no change*, 38 into *change* and 45 into *control*. The attributes (with many missing values) describe function, localization and protein class of each gene (hierarchical attributes with four to five levels). Again, a relation describes gene interactions. A table relates genes to 15,235 relevant abstracts in the MEDLINE repository.

We use the area under the *Receiver Operating Characteristic* (ROC) curve to assess hypotheses. [2] This area (the *AUC performance*) is equal to the probability that, when we draw one positive and one negative example at random, the decision function assigns a higher value to the positive than to the negative example. We estimate the standard deviation of the AUC performance, using the Wilcoxon statistics [2].

We selected the Support Vector Machine SVM^{light} as core machine learning algorithm. In order to study the benefit of some attribute x generated by one of the discussed approaches, we compare the performance of the attribute configuration with highest cross validation performance with and without the focused attribute.

3 Propositionalization

The gene interaction data contains pairs of gene names. In order to integrate this data into our solutions, we have to generate attributes from this relation. We use the RELAGGS algorithm that extends the usual framework of propositionalization [7] by first computing user specified joins, and then aggregating the result into a table with a single line per instance [9].

The following example illustrates this process. We have a table *train-class* with attributes *gene-id* and *class*, a table *interaction* with *gene-id1* and *gene-id2* and, slightly simplified, a table *localization* with attribute *gene-id* and an additional attribute for each possible value, including *mitochondria* and *cytoplasm* (this easily allows us to handle set-valued attributes). Gene “1” interacts with genes “2” and “3”, where “2” has value 1 for attribute *mitochondria* and 0 for *cytoplasm*, and “3” has value 1 both for *mitochondria* and *cytoplasm*. After joining the three tables, we obtain two lines for gene “1”; the first has value 1 for *mitochondria* and 0 for *cytoplasm*, the second has value 1 for both.

Table 1. Function and localization prediction with and without relational information.

Class	without	with
function growth	0.872 ± 0.01	0.882 ± 0.014
function transcription	0.886 ± 0.005	0.899 ± 0.011
function transport	0.893 ± 0.01	0.918 ± 0.013
localization cytoplasm	0.861 ± 0.008	0.865 ± 0.014
localization mitochondria	0.909 ± 0.013	0.948 ± 0.01
localization nucleus	0.941 ± 0.005	0.944 ± 0.011

Table 2. AhR prediction with and without relational information.

class	without	first level	second level	third level
narrow	0.62 ± 0.06	0.71 ± 0.05	0.69 ± 0.05	0.65 ± 0.06
broad	0.60 ± 0.04	0.60 ± 0.05	0.63 ± 0.04	0.60 ± 0.04

We now collapse these two lines into one by applying aggregation functions such as *min, max, avg*, or *sum*; in this case, *sum* is appropriate. This leads to one line with values 2 and 1 for attributes *mitochondria* and *cytoplasm*, respectively, indicating that gene “1” interacts with two genes localizing in the mitochondria and one in the cytoplasm. The RELAGGS outputs consisted for all problems of single tables with about 1,000 columns each. This high number is caused by the number of different values for functions, localizations, and protein classes.

For problems (1) and (2), we compare the decision functions with and without the interaction attributes, generated by the RELAGGS algorithm in Table 1 (using 10-fold cross validation). The observed AUC performance obtained when using the interaction is in every single case higher; the improvement exceeds two standard deviations in two cases and one standard deviation in two more cases. These data rule out the null hypothesis that the interaction information does not influence recognition performance. For problem (3), we compare the performance without and with attributes that reflect first, second, and third level interactions in Table 2. We see that first level interactions perform best for the narrow, and second level interactions are best for the broad positive class. A significant improvement ($p \approx 0.05$) is achieved for the narrow class, we see a smaller, insignificant improvement for the broad class.

Table 3. Additional information from information extraction.

	without	with	IE only
narrow	0.590 ± 0.061	0.685 ± 0.052	0.654 ± 0.055
broad	0.597 ± 0.040	0.630 ± 0.039	0.510 ± 0.044

Table 4. Transduction results for (1) and (2).

Class	SVM	TSVM
function growth, 150 positives	0.84 ± 0.006	0.82 ± 0.008
location cytoplasm, all data	0.83 ± 0.010	$0.71 pm 0.016$
function growth, 5 positives	0.67 ± 0.005	0.55 ± 0.007
location cytoplasm, 5 positives	0.62 ± 0.007	0.55 ± 0.011

4 Information Extraction

The attributes of the original data set contain very many missing values. We therefore want to study whether an information extraction algorithm can effectively be used to find missing values in the 15,000 MEDLINE abstracts. We follow a dictionary-based approach [5]. From the hierarchical text files that contain possible values for the attributes function, localization, and protein class, we manually define a thesaurus that lists, for each of the possible values of these attributes, a number of plausible terms that can be used to refer to this value. Terms are constructed by adding synonyms, paraphrased variants, and plural forms, and splitting compound phrases (see [8] for more details).

Table 3 shows that the extractor yields a substantial performance improvement for problem (3). Surprisingly, the problem can even be solved to some degree using *only* the information extracted from the abstracts ("IE only"). Applying the information extractor to problems (1) and (2) raises an interesting problem. The extractor identifies both, function and localization of genes in the scientific abstracts. Hence, it solves the functional genomics problem. However, this solution would not be practically useful because it could not possibly predict function and localization values that have not previously published.

5 Semi-Supervised Learning

The transductive SVM [6] maximizes the margin between hyper-plane and both, labeled and unlabeled data; but only for the labeled data, it is required that they lie on a specific side. For problems (1) and (2), Table 4 compares 10-fold cross validation results of the "vanilla SVM" to the TSVM. In both cases, transduction *significantly decreases* performance although it has additional information (the unlabeled hold-out instances) available. Given our previous positive experience with TSVM, we hypothesized that transduction is only beneficial if only few labeled data are available. We averaged 10 iterations in which we drew only 5 labeled positive examples (and 12 and 20 negatives, respectively) and used all remaining instances as unlabeled data. Table 4 shows that transduction still dramatically decreases classifier performance, refuting our hypothesis.

For problem (3), the transductive SVM decreased the AUC for the broad class from 0.63 (± 0.04) to 0.60 (± 0.04) and increased AUC for the narrow class from 0.685 (± 0.05) to 0.695 (± 0.05). Both differences are well below the standard deviations. The transductive SVM dramatically increases computation time.

Blum and Mitchell [1] have proposed the co-training algorithm which splits the available attributes into disjoint subsets; a labeled example (x, a) is then viewed as (x_1, x_2, a). The co-training algorithm learns two classifiers $f_1(x_1)$ and $f_2(x_2)$ which bootstrap each other by providing labels for the unlabeled data. When the views are *compatible – i.e.*, $\exists f_1, f_2$ such that $f_1(x_1) = f_2(x_2) = f(x)$ – and *independent* given the class labels – $P(x_1|f(x), x_2) = P(x_1|f(x))$ – then the co-training algorithm labels unlabeled examples in a way that is essentially equivalent to drawing labeled data at random [1].

For problems (1) and (2), we split the available attributes randomly. In each experiment, we averaged ten co-training curves for distinct attribute splits. Figure 1 shows that the performance of the decision functions remains unchanged when we use all available labeled examples, (topmost curve), or 150 labeled examples (second curve, plot for "function") whereas performance *decreases* over the co-training iterations when we use only 5 positive examples.

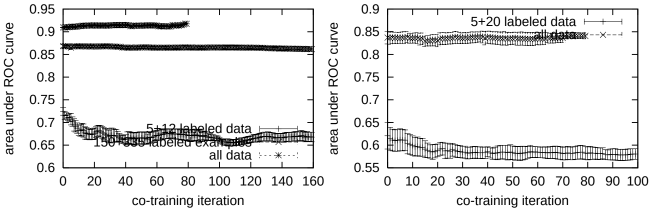

Figure 1. Co-training results for (1) and (2).

For problem (3) we tried to minimize dependency between the two attribute sets by splitting into attributes extracted from the abstracts together with the relational attributes, and all other attributes (the "natural" split). As control strategy, we randomly partition the attributes.

Figure 2 (left) shows the AUC over 200 iterations of co-training using the "natural" split. The performance does not improve; the standard deviations are around 0.05. The combined decision function (the average of two decision functions) is significantly worse than one single decision func-

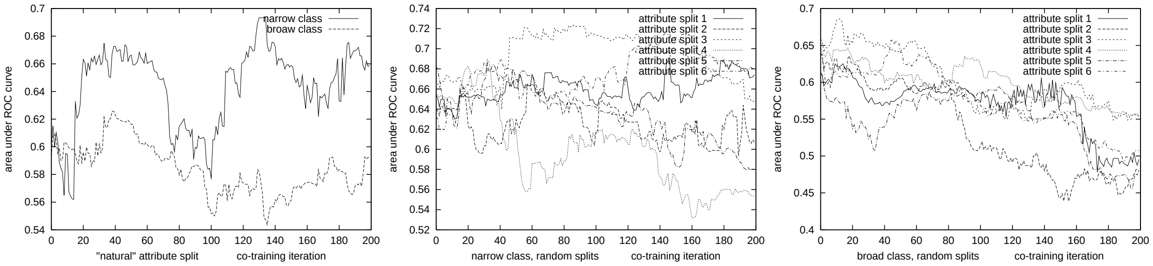

Figure 2. Co-training results for (3).

tion. For random attribute partitioning (Figure 2, center and right), the average AUC decreases significantly over the co-training iterations ($p < 0.05$) for the broad and seems to decrease for the narrow class.

6 Competition Results and Lessons Learned

For KDD Cup 2001, tasks 2 and 3, (here, problems 1 and 2) Krogel [3] submitted a solution that used the interaction attributes. Since the gene names were anonymized, we could not use text mining results; we did not use transduction or co-training either. The submission achieved the highest score for function prediction, and the fourth highest score for localization prediction [3]. Our solution for KDD Cup 2002, task 2 (here, problem 3) used second-level interactions, and entries won by information extraction. We did not include transduction or co-training. We achieved the third highest result (the highest for the narrow positive class); retrospectively, we can now obtain better performances than any team could within the tight competition time frame.

From our experience in the KDD Cup and retrospective studies, we draw a number of lessons learned. (i) In functional genomics, the interactions between genes play a crucial role. Effective utilization of the interaction information was the key success factor for both competitions. Using RELAGGS to propositionalize the data proved to be effective and scalable. (ii) Semi-supervised learning techniques such as the transductive SVM and co-training are less generally applicable than – at least we – expected. Our expectation was that taking unlabeled data into account should not decrease performance, and should at least be beneficial when only few labeled examples are available. For functional gene classification, neither of these assumptions is true. (iii) MEDLINE abstracts contain important knowledge that can help to build better models and thus to perform better on classification tasks. Our data shows that even simple, dictionary-based extractors can generate attributes that substantially improve classification performance.

Acknowledgements

We would like to thank David Page and Mark Craven. Thanks to Marco Landwehr and Marcus Denecke for their contributions to the KDD Cup 2002 entry. Tobias Scheffer is supported by Emmy Noether Fellowship SCHE540/10-1 of the German Science Foundation (DFG).

References

[1] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*, 1998.

[2] A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.

[3] J. Cheng, C. Hatzis, H. Hayashi, M.-A. Krogel, S. Morishita, D. Page, and J. Sese. KDD Cup 2001 Report. *SIGKDD Explorations*, 3(2):47–64, 2002.

[4] M. Craven. The 2002 KDD Cup competition results for gene regulation prediction. *SIGKDD Explorations*, 4(2), 2003.

[5] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Towards information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing*, 1998.

[6] T. Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the International Conference on Machine Learning*, 1999.

[7] S. Kramer, N. Lavrač, and P. A. Flach. Propositionalization Approaches to Relational Data Mining. In *Relational Data Mining*. Springer, 2001.

[8] M.-A. Krogel and T. Scheffer. Effectiveness of information extraction, multi-relational and multi-view learning for predicting gene deletion experiments. In *BIOKDD*, 2003.

[9] M.-A. Krogel and S. Wrobel. Transformation-Based Learning Using Multirelational Aggregation. In *Proceedings of the Eleventh International Conference on Inductive Logic Programming (ILP)*. Springer, 2001.

[10] T. Leek. Information extraction using hidden Markov models. Master's thesis, UCSD, 1997.