One Class SVM for Yeast Regulation Prediction

Adam.Kowalczyk
Telstra Research Laboratories
770 Blackburn Clayton, Victoria 3168, Australia
Adam.Kowalczyk @team.telstra.com

Bhavani.Raskutti
Telstra Research Laboratories
770 Blackburn Clayton, Victoria 3168, Australia
Bhavani.Raskutti @team.telstra.com

ABSTRACT

In this paper, we outline the main steps leading to the development of the winning solution for Task 2 of KDD Cup 2002 (Yeast Gene Regulation Prediction). Our unusual solution was a pair of linear classifiers in high dimensional space (~14,000), developed with just 38 and 84 training examples, respectively, all belonging to the target class only. The classifiers were built using the support vector machine approach outlined in the paper.

Keywords

Support Vector Machines, One Class Learning, SVM, yeast gene.

1. INTRODUCTION

The Yeast Gene Regulation Prediction data, Task 2 of KDD Cup 2002, was heavily unbalanced, with 38 and 84 'target' class examples only out of the total of 3018 examples in the training set. Most machine learning procedures for developing a discrimination in such a data will require some sort of re-balancing of the priors, i.e., boosting the impact of examples from the minority class combined with diminishing the impact of examples from the majority class. Using cross-validation on the training set we have found that the optimal solution for our setting is obtained in the extreme case, when the majority class is completely eliminated.

In this short paper, we present some details of our submission, including specifics of data representation and classification procedure as well as some results of cross-validation tests.

2. DATA REPRESENTATION

Each training and test gene was represented by a vector of binary attributes extracted from the data sources provided. Attributes were extracted by using only the entries from the data sources corresponding to the training genes.

- Hierarchical information about function, protein classes and localization was converted to a vector per gene. For instance, the following two entries in the file function.txt

  YGR072W cytoplasm | SUBCELLULAR LOCALISATION
  YGR072W nucleus | SUBCELLULAR LOCALISATION

  yielded three function attributes: "*cytoplasm*", "*subcellular localization*" and "*nucleus*" each with a value of 1 for the gene "YGR072W". This processing created 409 attributes: 42 for localization, 213 for gene function and 154 for protein classes.
- Textual information from all abstracts associated with a gene was converted to 'word token' presence vectors ('a bag of words'). A 'word token', in this context, is understood as any string of alphanumeric characters, which may and may not correspond to an ordinary word. Word tokens corresponding to words in a standard list of stop words, such as "the", "a" and "in", have been excluded. All ordinary words were stemmed using a standard Porter stemmer. This abstract processing resulted in 48,829 word token attributes. Around 3/4th of these attributes were subsequently eliminated by discarding all those that occurred in only one training gene, and by discarding all those which had a total frequency that was greater than one standard deviation from norm. After this processing, we were left with 12,480 word token attributes for the abstracts.
- The gene-gene interaction file is symmetric. Hence, each entry in the file interaction.txt creates two attributes. For instance, the entry "YFL039C YMR092C" creates the interaction attributes "*YFL039C*" and "*YMR092C*", and the attribute "*YFL039C*" is set to 1 for the gene "YMR092C" and vice-versa. Processing of the gene interactions file yielded a total of 1,447 attributes.

Thus, the total number of binary attributes used by the learning algorithm was 14,336 (= 409 + 12,480 + 1,447).

3. MODEL SELECTION

We have used a linear support vector machine [4] with quadratic penalty. This is a classifier allocating to each data sample $x \in R^n = R^{14336}$ the score $f(x) = x \cdot w + b$, where the solution vector $w \in R^n$ and the bias $b \in R$ are defined as minimisers of

$$\Phi(w,b) = \|w\|^2 + b^2 + \sum_{i=1}^{3018} C_i \left[\min(0,\ 1 - y_i x \cdot w - y_i b)\right]^2 .$$

Here $x_i \in R^n$ are feature vectors and $y_i \in \{\pm 1\}$ are bipolar labels of the training examples, $i = 1,\ldots,3018$. The individual regularisation constants, $C_i \geq 0$, are defined as $C_i = CB / n_-$ if $y_i = -1$ (the background class) and $C_i = C(1-B)/n_+$ if $y_i = +1$ (the target class) with *the balance factor* $0 \leq B \leq 1$ and *the regularization constant* $C \geq 0$ being free parameters, and $n_+ \in \{38, 84\}$ and $n_- \in \{2934,\ 2980\}$ denoting the numbers of target and the background class examples, respectively. Thus the smaller the balance factor B, the smaller the impact of the background class and the more promoted are examples from the minority (target) class. In particular, $B = 0.5$ represents the case of both classes with even balance of priors (ordinary 2-*class learning*); $B = 0$ is the extreme case of learning from the target class examples only (1-*class learning*); $B = 1$ is the opposite extreme, the case of learning from majority class examples only.

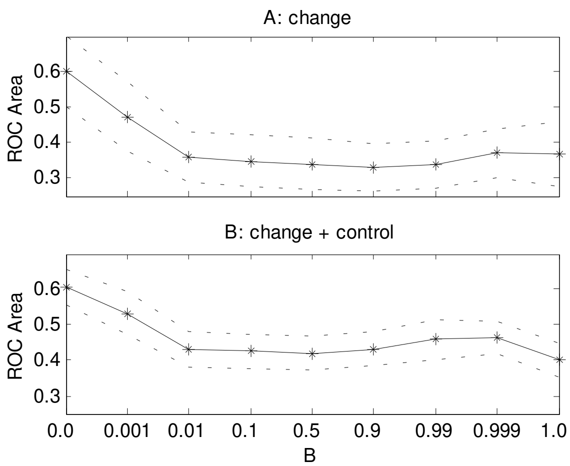

Figure 1: Mean AROC +/- std as a function of balance factor *B* (C = 5000).

Some results of cross-validation experiments aimed at 'optimization' of the balance factor B and the regularization constant C are shown in Figures 1 and 2. The figures show the mean AROC (area under ROC curve) with standard deviation as an envelope, where the means are computed on the validation set over 20 random splits of data into 70% : 30%, learning : validation.

Figure 1 shows the impact of the balance factor B on accuracy. We have used $C = 5000$, and the cross-validation tests are performed using splits of the training data only. Figure 2 shows the effect of the regularization constant C on AROC. Results are shown for $B = 0$ and $B = 0.5$, with cross-validation splits of the training data only in Figures 2A and 2B and combined training and test data in Figures 2C and 2D. Based on the results in Figures 1, 2A and 2B, the values $B = 0$ and $C = 5000$ were selected for the competition submission. This selection amounts to training 'hard margin SVM' with examples from a single (target) class only in the 14,336 dimensional feature space.

An additional point to note is that cross-validation estimates of AROC from the training data and the combined training + test data are very close to each other. Thus, in retrospect, the cross-validation technique for model selection was a justified step.

4. DISCUSSION

Our approach has a number of distinct features.

Automatic pre-processing and large number of features for classification: We have used a minimal domain knowledge and passed a large number of features to the classifier. This follows from our previous experience in practical text categorisation systems where laborious manual interventions without a deep domain insight often produced mediocre, if any, improvements.

One-class learning: A possibility of one class learning (with SVMs) has been explored previously [1,2,3]. In these experiments, while 1-class models performed reasonably, they were systematically outperformed by models developed using data from both classes. To our knowledge, the experiments with Yeast Gene data set reported in this paper, is the only case where the contrary is true.

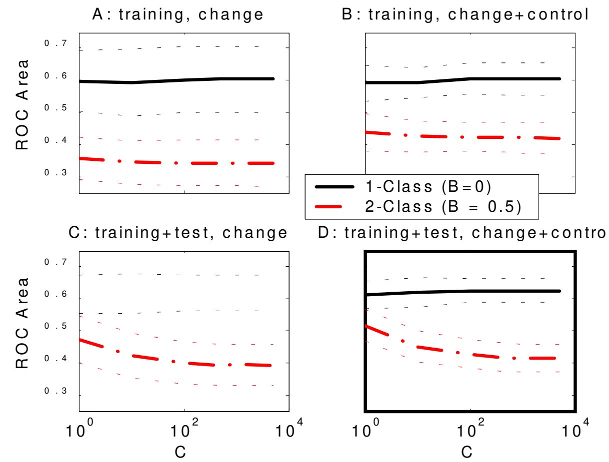

Figure 2: Mean AROC +/- std as a function of the regularization constant *C*.

We have arrived at our 1-class solution through systematic investigation of priors. The open question still is why such a solution works so well on this data set. Our explanation is that this is the effect of a specific 'interaction' of high dimensionality and sparsity of feature space with the noise in the data. Our recent experiments with this data and some other, artificial, data provide evidence that this happens for other data sets and classifiers.

5. ACKNOWLEDGMENTS

The permission of the Managing Director, Telstra Research Laboratories, to publish this paper is gratefully acknowledged

6. REFERENCES

[1] Y. Chan, X. Zhou and T.Huang. One class SVM for learning in image retrieval. In Proceedings of IEEE International Conference on Image Processing, 2001.

[2] L. M. Maneivitz and M. Yousef One-class SVMs for Document Classification Journal of Machine Learning Research}, 2:139--154, 2002. .

[3] B. Scholkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution, Neural Computations, 13 (7), 2001.

[4] V.Vapnik. Statistical Learning Theory, Wiley, New York, 1998.

About the authors:

Adam Kowalczyk works for Telstra Research Laboratories, Australia. His main research focus is theory and applications of machine learning systems. Adam obtained his Ph.D. from Warsaw University of Technology, in mathematics, and worked for a number of years in academia, prior to joining Telstra.

Bhavani Raskutti has been with the Telstra Research Laboratories, Australia for the last 10 years. Her main research focus is text mining and its practical applications in industry. Bhavani received her Ph.D. in Computer Science from Monash University, Australia. Prior to her Ph.D., she has worked in the software industry both in India and Australia.