Adapting Support Vector Machines to Predict Translation Initiation Sites in the Human Genome[1]

Rehan Akbani, Stephen Kwek
Human Genome Lab
Department of Computer Science
University of Texas at San Antonio
E-mail:{akbani,kwek}@HuGeLab.org

Abstract

This study is concerned with predicting Translation Initiation Sites (TIS) in the human genome that start with the nucleotide sequence ATG. This sequence occurs 104 million times in the entire genome. However, current estimates predict that there are only about 30,000 or so TIS in the human genome, giving an imbalance ratio of about 1:3500 for TIS ATG vs. non-TIS ATG sites. Algorithms that are designed using datasets that have low imbalance ratio may not be well suited to predict TIS at the genomic level. In this paper, we modified the SVM algorithm that can handle moderately high imbalance ratio. The F-measures for other approaches were: Linear Discriminant 0%, SVM with under-sampling 4.1%, SVM with over-sampling 8.2%, Neural Network 13.3%, Decision Tree 20%, our approach 44%. This shows how poorly standard approaches perform at the genomic level due to the high imbalance ratio. Our approach improves the performance significantly.

1. Background

This study is concerned with predicting Translation Initiation Sites (TIS) in the human genome that start with the nucleotide sequence ATG. Accurate prediction of TISs will aid in the discovery of new genes in the human genome by providing a starting point where protein translation begins. The vast majority of these sites start with the sequence ATG, while a small number of non-ATG TISs also exist [4]. In this study, we will only focus on ATG TISs. Most previous researchers in this field also restricted themselves to ATG only TIS [2, 9, 10, 11].

Previous work in TIS prediction mostly deals with predicting TIS in Expressed Sequence Tags (EST), mRNA sequences, or cDNA sequences [2, 9, 10, 11, 13, 14]. One of the most popular datasets used for this purpose is the Pedersen and Neilsen [9] dataset that contains data derived from mRNA sequences from vertebrates. In this dataset, the sequence ATG occurs a total of 13375 times, out of which 3312 are TIS ATG while 10063 are non-TIS ATG sites. This gives an imbalance ratio of only about 1:3 for the total number of TIS ATG vs. non-TIS ATG sites.

By contrast, we have found that the sequence ATG occurs about 104 million times in the entire human genome. This figure was obtained by scanning the human genome data available from the National Center for Biotechnology Information (NCBI) [6]. However, current estimates of the number of genes present in the human genome vary from about 30,000 to 40,000 genes [8]. Assuming most of these genes have TISs that start with ATG, this means that we can expect an imbalance ratio of about 1:2600 to 1:3500 in the human genome. Consequently, algorithms that trained on the Pedersen and Nielsen [9] dataset may not be directly applicable to TIS prediction at the genomic level.

Unfortunately, current Machine Learning (ML) algorithms do not perform well if the imbalance ratio exceeds about 1:10 [3] (in the remainder of this paper we will refer to the positive class as the minority class and the negative class as the majority class). Since ML algorithms are usually designed to maximize the overall accuracy on the dataset, most ML algorithms have a tendency of classifying a vast majority of instances as negative. This generates a lot of false negatives deteriorating the recall. Specifically, Support Vector Machines (SVM) usually end up classifying everything as negative and, therefore, have zero recall

[1] This research is supported by NSF grant CCR-0208935.

and precision. In our research we modified the basic SVM algorithm in order to improve its performance on highly imbalanced data. Because an imbalance of over 1:1000 is well beyond the performance capabilities of any ML algorithm, we decided to generate the TIS data from the human genome with an imbalance of 1:100 for our current scheme. Even this ratio causes most ML algorithms to perform very poorly, as our results below indicate. In the future, we hope to enhance our technique even further to handle larger imbalance ratios.

2. Our Method

Our first strategy was to construct a dataset containing sequences from the human genomic data and then use it to generate several candidate features for our algorithm. We then used feature selection algorithms to select the best attributes from among them. This technique was originally proposed by Zeng et al. [13]. To begin with, we randomly chose known ATG TIS sites from the NCBI database for our positive examples. Then we randomly picked ATG sites from the genome that are not known to be TIS sites, for our negative examples. We maintained a ratio of 1:100 for positive to negative examples. A window of 200 nucleotides was chosen for every example, running from 100 bps upstream of the ATG to 100 bps downstream of the ATG. This set constituted our raw dataset.

From this raw dataset, we generated several features. Every position in the raw data was used as a candidate feature. In addition, we generated the frequency of occurrence of all possible monomers, dimers, trimers, all the way up to hexamers that lie upstream of the ATG and also for those that lie downstream of the ATG. This gave us a total of 11120 features. Then we ran several different feature selection algorithms on this large set of attributes to determine the top attributes. We ran the Correlation Feature Selection (CFS) algorithm, which prefers those set of attributes that have a high correlation with the class label, but low correlation among themselves, and also Information Gain, Gain Ratio, and chi-squared test. By observing their results, we were able to choose the top 15 of the 11120 attributes, which were found to be the following (in order of importance): dn-CG, dn-TA, dn-AT, up-AT, up-CG, dn-GC, dn-G, up-TA, dn-CGG, up-CGG, dn-T, dn-ATT, pos -3, pos -1, pos +4, where dn-CG means the frequency of occurrence of CG downstream of the ATG, and up-CG means the frequency of CG upstream of the ATG, pos -3 means the nucleotide at position -3. Although we found pos -3, pos -1 and pos +4 to be the most important positions, the relevance score for these was much lower than the relevance score for the frequency counts, but we included them in our experiments nevertheless. It should also be noted that these positions correspond to the Kozak consensus sequence [4]. Our final dataset consisted of these 15 selected features. We used a similarly generated separate test set for evaluation.

For our algorithm, we modified the basic SVM algorithm by first generating several synthetic minority instances [1]. This was done by repeatedly randomly selecting two neighboring positive instances using the Euclidean distance measure and then generating a new instance that lies somewhere randomly in between these instances. The underlying assumption is that the space between two positive neighboring instances is assumed to be positive. We found this assumption to hold for our dataset. In this way we synthetically over sampled the minority class. We found that over sampling in this way was much more effective than the traditional over sampling technique of simply generating multiple identical copies of existing positive instances.

Our next strategy was to bias the SVM classifier so that it would be more inclined to classify instances as positive. One reason why SVM performs poorly on imbalanced data is that it tends to create a decision boundary that lies well inside the positive "space." By biasing the algorithm, we intended to push the boundary away from the positive space and closer to the ideal boundary that separates the positive from the negative instances. We accomplished this bias by increasing the relative penalty associated with misclassifying a positive instance as compared to a negative instance [12]. By varying the degree of over sampling and also the relative cost factor, one can obtain different values of recall vs. precision. The algorithm can be fine tuned to increase one measure relative to the other.

3. Results

We compared our algorithm with several other standard ML algorithms for predicting TISs in the human genome. We also compared our technique with the common approach of over sampling the minority class or under sampling the majority class in order to reduce the imbalance ratio in the dataset prior to training. For evaluation, we used the F-measure as our metric, which is the harmonic mean of the recall and precision. The results are shown in the following table:

Algorithm	F-measure
Voted Perceptron	0
ZeroR	0
SVM	0
SVM with Under Sampling	0.041
SVM with Over Sampling	0.082
Neural Network	0.133
AdaBoost with C4.5	0.148
3 Nearest Neighbors	0.182
Decision Tree	0.2
Naïve Bayes	0.205
Bagging with C4.5	0.25
Our Algorithm	0.44

As mentioned previously, by varying the parameters of our algorithm we are able to obtain different recall and precision values. Some examples of recall/precision obtained respectively are: 15%/100%, 29%/95%, 85%/4%. Thus, depending on the application the algorithm parameters can be varied to obtain the desired level of recall vs. precision.

4. Conclusion

The results in the table above show how poorly standard ML approaches perform for predicting TISs at the genomic level due to the high imbalance ratio. Our approach improves the performance significantly and in future we hope to enhance our algorithm further to handle even higher degrees of imbalance.

References

[1] Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357, 2002.

[2] Hatzigeorgiou, A. G. Translation initiation start prediction in human cDNAs with high accuracy. Bioinformatics 18, 343-350, 2002.

[3] Japkowicz, N. The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence: Special Track on Inductive Learning, Las Vegas, Nevada, 2000.

[4] Kozak, M. Interpreting cDNA sequences: some insights from studies on translation. Mammalian Genome 7, pp. 563-574 (1996).

[5] Liu, H., Han H., Li J. and Wong, L. Using amino acid patterns to accurately predict translation initiation sites. In Silico Biology 4, Bioinformation Systems e.V., 2004

[6] National Center for Biotechnology Information (NCBI), MD, USA, est. 1988. http://www.ncbi.nlm.nih.gov/

[7] National Human Genome Research Institute, USA, est. 1989. http://www.genome.gov/

[8] National Human Genome Research Institute, USA, est. 1989. http://www.genome.gov/12011238

[9] Pedersen, A. and Nielsen, H. Neural network prediction of translation initiation sites in eukaryotes: perspectives for EST and genome analysis, Proc. Int. Conf. Intell. Syst. Mol. Biol., 5:226--233, 1997.

[10] Salamov, A. A., Nishikawa, T. and Swindells, M. A. Assessing protein coding region integrity in cDNA sequencing projects. Bioinformatics 14, 384-390, 1998.

[11] Stormo, G., Schneider, T., Gold, L. and Ehrenfeucht, A. Use of the 'perceptron' algorithm to distinguish translational initiation sites in E.coli Nucleic Acids Res., 10, 2997–3011, 1982.

[12] Veropoulos, K., Campbell, C., & Cristianini, N. Controlling the sensitivity of support vector machines. Proceedings of the International Joint Conference on AI, 55–60, 1999.

[13] Zeng, F., Yap, H. C. and Wong, L. Using feature generation and feature selection for accurate prediction of translation initiation sites. In: Proceedings of 13th Workshop on Genome Informatics. Universal Academy Press, pp. 192-200, 2002.

[14] Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lemmen, C., Smola, A., Lengauer, T., and Muller, K.R. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 16: 799-807, 2000.