A Discriminative Approach to Structured Biological Data

Stefan Mutter* and Bernhard Pfahringer

Department of Computer Science
The University of Waikato
Hamilton, New Zealand
{mutter,bernhard}@cs.waikato.ac.nz

Abstract. This paper introduces the first author's PhD project which has just got out of its initial stage. Biological sequence data is, on the one hand, highly structured. On the other hand there are large amounts of unlabelled data. Thus we combine probabilistic graphical models and semi-supervised learning. The former to handle structured data and the latter to deal with unlabelled data. We apply our models to genotype-phenotype modelling problems. In particular we predict the set of Single Nucleotide Polymorphisms which underlie a specific phenotypical trait.

Keywords: Bioinformatics, Probabilistic Graphical Models, Semi-Supervised Learning, Single Nucleotide Polymorphism

1 Introduction

Biological data is highly structured in many different aspects. First of all there is an obvious structure in the sequence of DNA bases and amino acids. Different parts of a sequence, even far away from each other, can interact. These interactions may depend upon time and place. Another example is the regulatory network of gene expression which is a complex system.
In recent years there has been a push for methods that are able to deal with this kind of data, because, traditionally, Machine Learning has focused on independent and identically-distributed (iid) data [1]. This is why it is important to extend recent advances in machine learning theory and practice to structured, interdependent data.

2 Structured Data

As complex structured data becomes the focus of research, probabilistic graphical models become more and more important. They are well-founded in probabilistic and graph theory. Consequently a graphical model is a family of probability distributions that factorise according to an underlying graph [2]. A common

* Author to whom correspondence should be addressed.

distinction between probabilistic graphical models is to differentiate between generative and discriminative models.

A generative model models the full joint probability distribution $p(y, x)$ where the variables y stand for the attributes of the entities we wish to predict and x stand for our observed knowledge [2].

In contrast a discriminative model is directly based on the conditional probability $p(y|x)$ [2–4]. Sutton and McCallum [2] point out that the crucial difference between a generative and a discriminative model is that the latter does not include a model for $p(x)$. First of all, for a classification task this model isn't needed anyway, secondly it often contains highly dependent features. Thus it is hard to model. However, if we want to integrate interdependent features in a generative model, Sutton and McCallum [2] offer two possibilities. On the one hand, potentially unwarranted independence assumptions can help, on the other the introduction of additional parameters can solve the problem. The second approach can only be used in a limited way because the model can easily become intractable. In contrast there exist well-known examples for the first approach e.g. Naive Bayes works well in document classification, but the independence assumption can also hurt performance on average across a range of applications where its discriminative counterpart logistic regression outperforms Naive Bayes [5].

We are looking at probabilistic graphical models for sequences. A Hidden Markov Model (HMM) is a well-known generative model for sequences. Its discriminative counterpart is a Conditional Random Field (CRF). This research project will focus on this discriminative technique. A CRF can be seen as an extension of logistic regression to arbitrary graphical structures [2]. Thus, in addition, CRFs relax the independent and identically-distributed assumptions in the sequence itself and between sequences. In Bioinformatics they have been successfully applied to gene prediction, RNA structural alignment, protein structure prediction [2] and finding gene and protein mentions in the literature [6].

3 Unlabelled Data

In many areas including Biology there exists a large amount of unlabelled data, because labelling is often difficult, time-consuming and expensive.

In this context where labelled and unlabelled data exists semi-supervised learning is a new approach in Machine Learning. It uses a potentially large amount of unlabelled data together with a usually small amount of labelled data to build a classifier. Generative models are the oldest semi-supervised learning technique [7].

Usually we can get $p(x)$ from unlabelled data [7]. For discriminative learning it is believed that semi-supervised learning cannot help if $p(x)$ and $p(y|x)$ do not share parameters [7, 8]. But, as a lot of approaches show, semi-supervised learning can outperform supervised learning when it is applied carefully and the underlying assumptions are correct [7]. Current research tries to adapt discriminative techniques to semi-supervised learning [9, 1].

4 Genotype-Phenotype Modelling

We are investigating biological sequences in particular sequences of single nucleotide polymorphisms (SNPs) where each one is a sequence alternation of a single nucleotide in a DNA sequence which occurs in at least one percent of the population.

A fundamental problem in contemporary genetics is the relation between genotype and phenotype known as genotype-phenotype modelling. Examples include identifying superior dairy cows that is identifying genes that are responsible for phenotypical traits which increase economic merit [10].

A lot of SNPs have been identified by high-throughput methods and need now to be analysed. SNPs can be used as genetics markers but they are also a reason for phenotype differences even though most SNPs have no effect on the phenotype. This is why it is important to find the SNPs that are related to a particular trait. This problem is called tagSNP selection [11]. Lee [11] emphasises the need for new, probabilistic methods.

As the number of discovered genes that contribute to a specific phenotype grows, so does the complexity of models describing genotype-phenotype relations [12]. Rodin [12] suggests the use of probabilistic graphical models to represent this kind of structured data.

5 Research Synopsis and Project Status

This PhD project aims for advances in genotype-phenotype modelling by the use of probabilistic graphical models, especially Conditional Random Fields. Due to the fact that in biological domains there is a vast amount of unlabelled data, the incorporation of semi-supervised learning methods is an important aspect. The primary source of data are biological sequences.

From a Machine Learning point of view the underlying hypothesis is that discriminative techniques should outperform generative ones on a classification task [9]. This statement is supported by Vapnik [13], who argues that it is better to solve the classification problem directly than looking at the more general problem of modelling the joint probability distribution as an intermediate step. However the work of Ng and Jordan [14] shows empirical results suggesting that discriminative learners have a lower asymptotic error but generative models approach their (higher) asymptotic error faster. Research in this area will also lead to get some more insights in the differences between generative and discriminative modelling.

We expect that the adaptation of models to perform semi-supervised learning should enhance them. However first results also highlight that there is a lack of appropriate graphical structures for biological problems. This is crucial, because the graph determines how a family of distributions get factorised. Because the problems are highly structured, a good representation of the structure is essential.

This allows two possible ways of optimisation: Using semi-supervised learning

and enhancing the graphical structure. Currently a combination of both seems to lead to promising results. The next step is to define an exact optimisation criterion.

We are going to build models using Conditional Random Fields and apply them to SNP data first. Currently we are setting up an experimental environment and pre-processing the data so that it can be used to solve the tagSNP problem.

References

1. Lafferty, J.D., Zhu, X., Liu, Y.: Kernel conditional random fields: representation and clique selection. In Brodley, C.E., ed.: Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004). (July 2004)
2. Sutton, C., McCallum, A.: An introduction to conditional random fields for relational learning. In Getoor, L., Taskar, B., eds.: Introduction to Statistical Relational Learning. MIT Press (2006) To appear.
3. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2001) 282–289
4. Wallach, H.: Efficient training of conditional random fields. Master's thesis, University of Edinburgh (2002)
5. Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms using different performance metrics. Technical Report TR2005-1973, Cornell University, Ithaca,USA (2005)
6. McDonald, R., Pereira, F.: Identifying gene and protein mentions in text using conditional random fields. BMC Bioinformatics **6** (2005) S6
7. Zhu, X.: Semi-supervised learning literature survey. Technical Report 1530, Computer Science, University of Wisconsin-Madison (2005)
8. Seeger, M.: Learning with labeled and unlabeled data. Technical report, University of Edinburgh (2001)
9. Brefeld, U., Büscher, C., Scheffer, T.: Multi-view discriminative sequential learning. In Gama, J., Camacho, R., Brazdil, P., Jorge, A., Torgo, L., eds.: ECML. Volume 3720 of Lecture Notes in Computer Science., Springer (2005) 60–71
10. Garrick, D., Snell, R.: Emerging technologies for identifying superior dairy cows in new zealand. New Zealand veterinary journal **53(6)** (2005) 390–399
11. Lee, P.H.: Computational haplotype analysis: An overview of computational methods in genetic variation study. Technical Report 2006-512, Queen's University, School of Computing, Kingston,Canada (April 2006)
12. Rodin, A., Boerwinkle, E.: Mining genetic epidemiology data with bayesian networks i: Bayesian networks and example application (plasma apoe levels). Bioinformatics **21(15)** (2005) 3273–3278
13. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons (1998)
14. Ng, A.Y., Jordan, M.I.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press (2002)