Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

The Wisdom of Crowds in Bioinformatics: What Can We Learn (and Gain) from Ensemble Predictions?

Mariana R. Mendoza
PhD Thesis Supervisor: Ana L. C. Bazzan
PPGC, Instituto de Informática, UFRGS, Porto Alegre, Brazil

Abstract

The combination of distinct algorithms expertise to improve prediction accuracy, inspired by the theory of wisdom of crowds, has been increasingly discussed in literature. However, its application to bioinformatics-related tasks is still in its infancy. This thesis aims at investigating the potential and limitations of ensemble-based solutions for two bioinformatics prediction tasks, namely inference of gene regulatory networks and prediction of microRNAs targets, as well as propose new integration methods. We approach this by considering heterogeneity in the contexts of data and methods, and adopting machine learning methods and concepts from multiagent systems, such as social choice functions, for integration purposes.

Introduction

Bioinformatics is a highly active scientific field concerned with the development and application of computational tools to biological data analysis, dealing mainly with tasks of prediction, classification and network inference. There is a wide range of research areas in the field, most of which face major challenges such as sparseness of data sets and limited knowledge about the organisms dynamics. Among these areas, this thesis focus on the prediction of gene regulatory networks (GRNs) and microRNA target genes, both closely related to the analysis of gene expression and regulation.

The inference of GRNs from gene expression data is one of the most complex problems in the field. It aims at uncovering and modelling the mechanisms of gene regulation responsible for cellular development and function, performed by the organisms' underlying GRNs, through the analysis of implicit information embedded on experimental biological data. The main motivation relies in the fact that GRNs represent blueprints of the functional cooperativity among genes. This is particularly interesting, for instance, in drugs and medical treatments development.

However, it is already known that only a small percentage of the genome encodes protein sequences, and that many other elements with (in)direct participation in gene regulation derive from noncoding DNA sequences. For instance, microRNAs (a class of short noncoding RNAs) play an important role as negative regulators of gene expression. Therefore, the identification of microRNAs target genes based on features extracted from the alignment between microRNAs and their true/false targets has become a major classification task in ML applied to bioinformatics. Among other challenges, ML-based tools face the small availability of negative examples of microRNAs targets, an issue that affects the accuracy of classifiers sensitive to class imbalance.

While most works on the above problems focus on the development of new methods or heuristics to address weaknesses of existing methods, a novel trend in several areas of knowledge aims at combining distinct algorithms in order to provide more accurate predictions. For instance, studies on classification methods discuss the improvements yielded by the use of classifier ensembles, given mainly by the complementarity among the information retrieved by different algorithms (Dietterich 2000; Kuncheva 2004). Algorithms tend to have explicit or implicit biases, and as a result, different algorithms applied to the same data set hardly generate the same outcome and no method is inherently superior to any other (*No Free Lunch* theorem). In fact, previous studies on the inference of GRNs have already confirmed this theorem in practical applications (Marbach and et al. 2012).

Goals and preliminary results

Given these facts, this thesis aims at exploring the potential and benefits of the diversity introduced by ensemble predictions, studying its application to the biological problems outlined above. The main motivation comes from the concept of wisdom of crowds: under specific criteria, a group of individuals deciding collectively is likely to perform better than experts (Surowiecki 2005). This raises the hypothesis that by fully understanding the circumstances that lead to this phenomenon and investigating more robust mechanisms to aggregate individual predictions, one could reach an increase in the predictive accuracy as greater as those obtained by improved or newly developed methods.

Diversity is the main criteria for ensemble predictions and in this work I'll explore it in two aspects: i) making inferences or predictions upon heterogeneous data sets, derived from diverse biological evidences, and ii) combining the results of distinct algorithms, or multiple runs of heuristic methods. Previous studies (Marbach and et al. 2012) have already explored diversity among methods in the problem

of predicting GRNs, showing that this is indeed a promising approach. Thus, here I intend to go beyond the aggregation mechanisms used in past studies and import ML methods, as well as concepts from multiagent systems, such as social choice functions, as different and more sophisticated ways of combining classifier decisions. My goal is to seek further insight into the potential, robustness and limitation of ensemble-based approaches, applying it to optimise performance in the selected bioinformatics problems. Nonetheless, I hope the solutions and knowledge drawn in this work concerning ensemble predictions will be broadly applicable in other domains.

As part of my thesis, I've developed a genetic algorithm (GA) solution for the inference of GRNs modelled as Boolean networks. I explored different fitness functions, representations and heuristics for genetic operators (Mendoza and Bazzan 2011; Mendoza, Lopes, and Bazzan 2012), as well as proposed a new mutation operator based on the epsilon-greedy strategy to exploit available prior knowledge (Mendoza, Werhli, and Bazzan 2012). Moreover, while most GA solutions adopt a criterion to select one of the individuals in the last generation as the final network, in these works the concept of wisdom of crowds is applied to build a consensus network given a set of highly scored individuals in the final GA population. Results show that the area under the ROC curve (AUROC) for the final consensus network is higher than the average AUROC score for individual predictions in every case tested, even with simple voting mechanisms, such that results could be even better if more robust integration methods were considered.

Moreover, I performed a comparative study among popular fusion-based combination methods in classifier ensembles, considering the prediction of GRNs upon different algorithms (Mendoza and Bazzan 2012) in order to investigate ensemble effects in this specific problem. Results confirm the hypothesis that even simple and linear combination methods, such as the average of confidence levels given by inference methods to gene interactions, outperform the individual solutions. More sophisticated combination methods, such as voting by Borda count and the statististical method Dempster-Shafer, provide outstanding performance gain.

In what concerns the problem of predicting microRNAs target genes, I've worked in collaboration with biologists to develop a tool for the prediction of human microRNAs targets based on the Random Forest algorithm (Mendoza et al. 2012), which is by itself an ensemble approach based on the aggregation of several predictions provided by a set of decision trees. The collection of features based on which our tool performs classification, combined to the robustness of Random Forest, has yielded a superior predictive accuracy when compared to other widely used classification algorithms.

More recent work consists in the development of a multiagent classification system to cope with classification tasks characterised by diverse methods and vertical data distribution, testing it with the prediction of human microRNAs targets. Multiagent systems (MAS) are a natural choice for such scenario as they are able to deal with distributed data and expertise, as well as heterogeneous information. In the proposed framework, agents encapsulate distinct classifiers (mimicking different expertise) to learn a model independently, upon different features (point of view), and then work towards building a consensus domain for the classification task by means of social choice functions (SCFs). Thus, this work also aims at devising a solution for distributed classification tasks. Based on a multi-fold evolution, I found that the combination of predictions by SCFs has improved classification results in contrast to individual classifiers, increasing accuracy and yielding more consistent predictions across multiple runs.

Future work

The next steps of my work, to be accomplished until August 2013, concern a deeper investigation about SCFs, as well as the development of new combination methods and criteria, seeking inspiration in ML and MAS negotiation methods. Ideally, both data and algorithms diversity will be explored and compared for the two problems outlined in this summary. Experiments will be designed to help in better understanding the power and limitation of ensemble predictions, their potential to overcome intrinsic bias, the impact of properties such as homogeneity and heterogeneity of the group, as well as their robustness to weak classifiers or noisy data. Final goals of this thesis include the development of a framework that automates the generation of consensus predictions based on the principle of wisdom of crowds, yielding more accurate and interpretable (in a biological sense) results.

References

Dietterich, T. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*. Springer. 1–15.

Kuncheva, L. I. 2004. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Marbach, D., and et al. 2012. Wisdom of crowds for robust gene network inference. *Nature Methods* 9(8):796–804.

Mendoza, M. R., and Bazzan, A. L. C. 2011. Evolving Random Boolean Networks with Genetic Algorithms for Regulatory Networks Reconstruction. In *Proceedings of GECCO 2011*, 291–298. New York, NY, USA: ACM.

Mendoza, M. R., and Bazzan, A. L. C. 2012. On the Ensemble Prediction of Gene Regulatory Networks: a Comparative Study. In *Proceedings of SBRN 2012*, 55–60. IEEE Computer Society.

Mendoza, M. R., and et al. 2012. RFMirTarget: a Random Forest Classifier for Human miRNA Target Gene Prediction. In *Proceedings of BSB 2012*, number 7409 in Lecture Notes in Bioinformatics, 97. Berlin Heidelberg: Springer.

Mendoza, M. R.; Lopes, F. M.; and Bazzan, A. L. C. 2012. Reverse Engineering of GRNs: an Evolutionary Approach based on the Tsallis Entropy. In *Proceedings of GECCO 2012*, 185–192. New York, NY, USA: ACM.

Mendoza, M. R.; Werhli, A. V.; and Bazzan, A. L. C. 2012. An Epsilon-Greedy Mutation Operator Based on Prior Knowledge for GA Convergence and Accuracy Improvement: an Application to Networks Inference. In *Proceedings of SBRN 2012*, 67–72. IEEE Computer Society.

Surowiecki, J. 2005. *The Wisdom of Crowds*. New York, NY: Random House.