Protein Structure Prediction as a source of challenges for Machine Learning

Jaume Bacardit, Newcastle University, UK
Jaume.bacardit@newcastle.ac.uk

Introduction. We live in data-rich times. Almost all aspects of science and society generate data at unprecedented rates and of rich variety. This is both a challenge and an opportunity for machine learning methods. Biological data are an invaluable source of challenges for machine learning methods for several reasons: Large sets of records, large dimensionality spaces, diverse data representations (e.g. the discrete alphabet of DNA sequences, quantitative measurements of a molecules and molecular states) or rich meta-data (curated annotations, cross-linking between databases, etc.). Protein structure prediction data is a type of biological data that is particularly interesting because it joins together almost all of these challenges. The ECBDL'14 big data competition took place during 2014 using such data. This abstract describes the dataset, rules of the competition, results and lessons learnt. Finally, it describes another type of protein structure prediction data that would be suitable for other machine learning challenges.

The ECBDL'14 challenge. The Evolutionary Computation for Big Data and Big Learning workshop (ECBDL'14; http://flexgp.csail.mit.edu/GECCO-big-data-workshop/) took place on July 13th, 2014 within the Genetic and Evolutionary Computation conference (GECCO-2014) in Vancouver, Canada. An important part of the workshop was a big data classification competition, in which participants were provided with a training set of ~32million instances, 631 attributes and 2% of positive examples. Participants had to train models using their own resources and then submit prediction for the test set to the competition's server. A period of two months (May-June of 2014) was given to submit predictions, in which participants could see their progress through the competition's leaderboard (http://cruncher.ncl.ac.uk/bdcomp/index.pl?action=ranking). Finally, the results of the competition were discussed during the workshop.

Dataset description. Protein structure prediction (PSP) is the estimation of a complete 3D model of a protein's structure (figure 1, top-right) from the amino acid composition of the protein's chain. This problem remains the holy grail of computational biology despite many decades of progress. The overall PSP problem is generally treated as an optimisation problem, in which an objective function that assesses the 'native-like' properties of a model is used. Many sub-problems of PSP have been defined as machine learning task. The value of predicting these sub-problems is that they can make easier the optimisation process of the overall PSP problem. Among the machine learning sub-problems of PSP, contact map prediction is probably the most challenging of them. Two amino acids of a protein are said to be in contact if their distance in the structure is less than a predefined threshold. A contact map (figure 1, top-left) is a binary matrix where the rows and columns are the amino-acids of a protein and cells indicate whether that pair is in contact or not. This matrix is very sparse as generally only 2% of all possible pairs are in contact. This creates classification datasets with huge class imbalance. Moreover, because training sets contain (almost) any possible pair of amino acids in a protein, just by using a few thousands of proteins with known structure for training, the number of pairs of amino acids easily reaches the tens of millions. Finally, the state-of-the-art knowledge representations for this problem easily reach the hundreds of attributes, and are composed of three large blocs (figure 1, bottom): (1) a detailed representation of the local context (neighbours in the protein chain) around the pair of amino acids for which we are predicting contact/non-contact and optionally also the middle-point between them, (2) statistical profiles of the segment

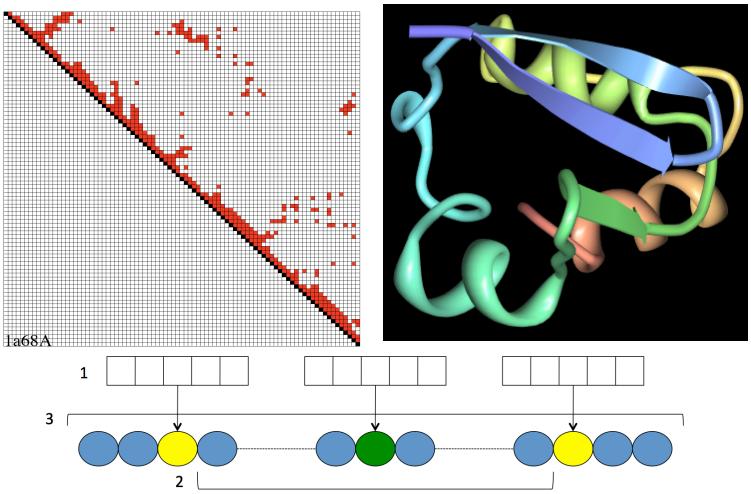

Figure 1. Top-left: Contact Map of protein 1a68A. Top-right: 3D cartoon visualisation of protein 1a68A. Bottom: Diagram showing the different parts of the representation for the contact map prediction dataset

connecting the pair of amino acids and (3) statistical profiles about the overall protein sequence. Depending on the size of this local context (number of neighbours) and what information about them is used, the number of attributes in the dataset can greatly vary. In the particular case of the dataset used in the competition [1], 552 attributes were used for part 1, 38 attributes for part two and 41 attributes for part 3. A diverse set of 3262 proteins with known structures was used. 90% were used for training (32M pairs of amino acids) and 10% for test (2.89M pairs of amino acids).

Rules of competition. Given the huge class imbalance in the dataset (2% of positive examples), the scoring function had to balance the correct predictions for the positive and negative examples. To this aim, the product of the true positive rate (TP/P) and true negative rate (TN/N) on the test instances was used as the overall score for the competition. Participants were told that their accounts could be cancelled if they tried to abuse the system by submitting too many predictions.

The prediction season. Seven teams participated in the competition, with a total of 364 submissions through the two months of competition (figure 2). Among the participants a variety of classification algorithms and computational infrastructures were used. The winning team EFDAMIS from the University of Granada in Spain used a Hadoop-based pipeline including random oversampling of the minority class, evolutionary feature weighting and random forests. Their experiments were run on a 144-core cluster and the wall-clock time of the training process of their winning solution was 39 hours. The second team (ICOS; our own team) used an ensemble of rule sets generated using evolutionary computation. Our experiments were distributed as a series of batch jobs run in a standard HPC cluster using 3000 hours of CPU time for the training of our best solution. The third team in the ranking (UNSW, from the University of New South Wales) used deep learning-style learning algorithm with 8250 hours of CPU time for training in a 24-core cluster using parallel programming. Other teams lower in the ranking used Support Vector Machines, standard neural networks, genetic programming, etc.

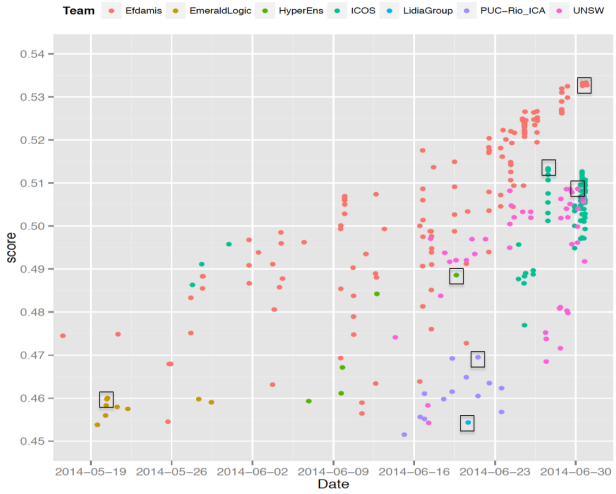

Figure 2. Timeline of the competition, marking the best prediction of each team

Lessons learnt from the competition. The most interesting aspect of the competition is the variety of strategies (random forest, rule-based machine learning, deep learning, support vector machines) and frameworks (Hadoop, batch HPC, parallel programming HPC, GPUs) used by the participants. This diversity, however, makes it difficult to compare the computational effort made by the participants. Moreover, in order to make it as easy as possible for the participants, the rules of the competition diverged considerably from the standard way in which contact map prediction methods are assessed [2], which makes very difficult to assess if these learning strategies are a real contribution to the problem domain. Nevertheless, this was an interesting and challenging machine learning task, and participants (based on verbal feedback) greatly enjoyed it.

Other PSP challenges. The PSP community currently acknowledges model selection as one if its biggest challenges: Given a large set of complete 3D structure models, choose which one of them would be the most similar to the real structure of the protein. The criteria used to decide this selection use a variety of model descriptors that are either derived from physics first principles (energy terms) or from statistics collected from proteins with known structure (knowledge-based potentials). From a machine learning perspective, this problem can be treated in many different ways: (1) construct, through **regression**, scoring functions, (2) rather than scoring individual models, the selection can be tackled as a **learning to rank** task given a set of input models. (3) As a **feature generation** task, by creating descriptors, based on a catalogue of energy terms and knowledge based potentials that then can be used by standard regression/rank algorithms.

[1] Bacardit, J. et al. Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features. Bioinformatics (2012) 28 (19): 2441-2448.
[2] Monastyrskyy B. et al. *Evaluation of residue-residue contact predictions in CASP9. Proteins 2011;79 Suppl. 10:119-125.*