Dynamic Outlier Exclusion Training Algorithm for Sequence Based Predictions in Proteins Using Neural Network

Shandar Ahmad

National Institute of Biomedical Innovation,
Saito Asagi, Ibaraki-shi, Osaka, Japan
shandar@netasa.org

Abstract. Many structural and functional properties of proteins can be described as a one-dimensional one-to-one mapping between residues of protein sequence and target structure or function. These residue level properties (RLPs) have been frequently predicted using neural networks and other machine learning algorithms. Here we present an algorithm to dynamically exclude from the neural network training, examples which are most difficult to separate. This algorithm automatically filters out statistical outliers causing noise and makes training faster without losing network ability to generalize. Different methods of sampling data for neural network training have been tried and their impact on learning has been analyzed.

Keywords: Binding sites, Neural networks, Sequence information, Outliers.

1 Introduction

Sequence-structure-function relationship of proteins has been historically one of the most important issues in bioinformatics for a very long time [1-3]. However, despite an intense effort to predict protein structure from the amino acid sequence, the task has remained difficult and far from complete. Compared to that ambitious goal of predicting everything from sequence or structure, it seems much more plausible to predict the so-called one-dimensional properties of protein structure such as secondary structure, solvent accessibility and coordination number on the one hand and biological functions such as binding with specific ligands or DNA bases on the other. Both one-dimensional structural features of proteins and probability of binding of an amino acid with other molecules have been predicted from the information of amino acid sequence with good success [4-9] and have in many ways led the way for an eventual ab initio structure and function prediction without homology or structure models. One of the most widely used method for mapping sequence information on to functional and structural target properties has been neural network. Neural networks provide a very efficient tool to model almost any non-linear relationship between sequence data and their target properties. These models have been successful in predicting secondary structure, solvent accessibility and binding sites. As larger data sets of binding sites and structural properties become available, their processing with

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 142–147, 2007.

neural networks will become slower albeit more powerful. Faster algorithms and efficient analysis of feature vectors and their relationshop with target properties are needed to address these problems. One of the problems is poor predictability of some of the patterns even when most of the samples are well predicted. We have developed an algorithm to dynamically select training examples for neural network and flag them as prediction outliers. In this algorithm a neural network is not trained on the entire data set, but the error scores are computed for each data example and then the examples contributing the most to the error score are eliminated from the training process. We report the resulting learning curves, amount of excluded data and their impact on the ability of the neural network to generalize prediction.

2 Methods

2.1 Definition of an Outlier

A statistical outlier is generally known to be a pattern with too high or too small value of its corresponding attribute. In the context of feature-based predictions of target properties, we define a statistical outlier to be a pattern in which the relationship between its feature vector and its target property does not follow the same relationship as done by the overall data set. Formally, a pattern will be classified as an outlier if the prediction error (ε_i) in that sample is much more than the overall variance in the data i.e.

$$\varepsilon_i > \varepsilon_{av} + \alpha.\, \sigma(\varepsilon) \quad (1)$$

Where ε_{av} is the average absolute error in the overall data, $\sigma(\varepsilon)$ is the standard deviation in the pattern-wise absolute error and ε_i is the error in the ith sample, to be tested for being an outlier or not and α is an adjustable parameter to determine the strictness of the flagging criterion.

2.2 Treatment of Outliers

Once the training examples have been flagged as outliers, there are at least two methods of treating them. First, instead of assigning them high error values returned by the predictor, their predicted values may be reassigned such that their contribution to error does not exceed the criterion set by (1). Alternatively, the outliers may be totally removed from the data set and they do not contribute at all to the performance scores. Later leaves behind a smaller data and and the calculation of the error gradient becomes faster in the process. We have used the both these criterion to analyze learning behavior but report the results obtained from the second one.

2.3 Dynamic Identification of Outliers

Using the outlier identification criterion given by (2), the identification of outliers has to be done for every epoch as data points move from normal to outlier categories and vice versa as the training progresses. In particular, the random initialization of weights produces large variance in error and therefore very few outliers according to the

above definition. As the training progresses, both mean error and their variance decrease with a clearer picture of outliers emerging. A typical variation in the number of patterns identified as outliers with training (epoch number) has been shown in Figure 2 (see results section).

2.4 Data Sets and RLP Types

Three types of predictions are attempted viz. Solvent accessibility (class-type predictions and real value predictions) [5-6], DNA-binding site predictions [7-8] and Carbohydrate-binding site predictions [9]. Data sets used for these predictions have been explained in the corresponding previous publications. In this work, we have used 512 proteins for analyzing ASA prediction, 40 proteins for analyzing sugar-binding sites and 62 proteins for assessing DNA-binding sites. Similar results have been obtained for these data sample, but the results discussed in this paper are based on solvent accessibility data because its values are distributed in a range from 0 to 1, instead of binary values in the case of binding sites and hence analyzing performance in solvent accessibility prediction is easier.

2.5 Neural Networks

In all our prediction experiments, a layered neural network with single hidden layer containing two units was used. The input layer consists of 60 units representing a tripeptide with a target residue at the center and one sequence neighbors on either side included as context information. Output layer is a single neuron with real valued outputs, transformed into binary values with a simple threshold function. Activation function for the hidden layer is *arctan*, and for the output layer it is a *sigmoidal* function. Neural network is trained using generalized delta-rule and weights are updated in the direction of maximum gradient after presenting all patterns at the end of each epoch according to the following learning rule:

$$\Delta W_{ijk} = \eta \, \partial E/\partial W_{ijk} \tag{2}$$

Learning rate was maintained at 1.0 for all these calculations.

3 Results and Discussion

3.1 Outlier Exclusion Does Not Affect Generalization

Figure 1 shows learning curves of a neural network trained for 200 epochs using generalized delta rule, using different criterion of data exclusion. Mean absolute error of prediction in the test data, used for determining the stopping point for training was used as a measure of generalization. This particular graph shows the learning curves for solvent accessibility and similar curves were also obtained with binding data of DNA and carbohydrates. We observe that the neural network training carried out at α=3 has almost the same prediction error as the one carried out on full data sets. Training performed with a more strict criterion of data inclusion (α=1) suffered from poor training performance as it excluded too many data points. An interesting

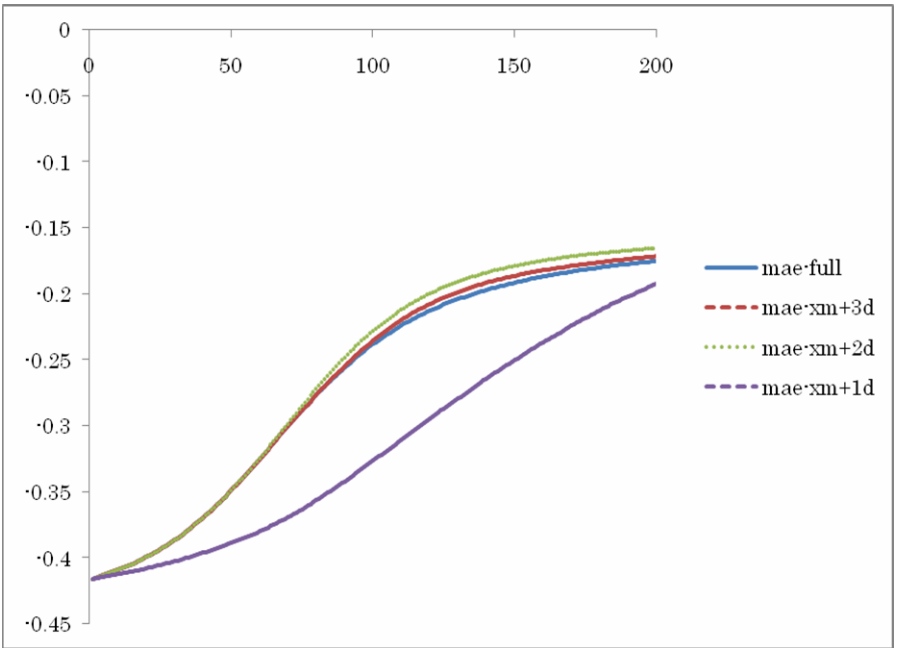

Fig. 1. Learning history of mean absolute error in the test data (generalization). Abbreviation (mae-full: Mean Absolute Error in test data without error exclusion, mae-xm+α.d: MAE in test data when training points with $\varepsilon_i > \varepsilon_{av} + \alpha.\ \sigma\ (\varepsilon)$ were excluded). MAE values are marked as negative to contrast them from correlation and other accuracy scores to indicate that a smaller MAE means better prediction.

observation was made for α=2. There was a small improvement in prediction performance of the neural network at this value, suggesting that a suitably selected value of α may actually improve the generalizing ability of neural network. However, DNA and Carbohydrate-binding sites data did not show a similar improvement, probably because the amount of data available in these categories was not large enough to take advantage of this situation.

3.2 Error Distribution and History of Outlier Frequency

In Figure 2, we show the outlier frequency variations in different stages of neural network training. In the early stages of neural network training errors are randomly distributed leading to a large value of variance and hence no outliers can be identified in the early training. As the neural network learns the variance in prediction error decreases and outliers can be identified. With a strict criterion of outliers (large α), very few outliers are detected and at small values of α, too many patterns are excluded from training. A large number of rejected data for α=1, is clearly responsible for poor generalization of prediction (Figure 1). A value of α=2 is suitable for generalization and also excluding sufficient number of data points to speed up the process of learning.

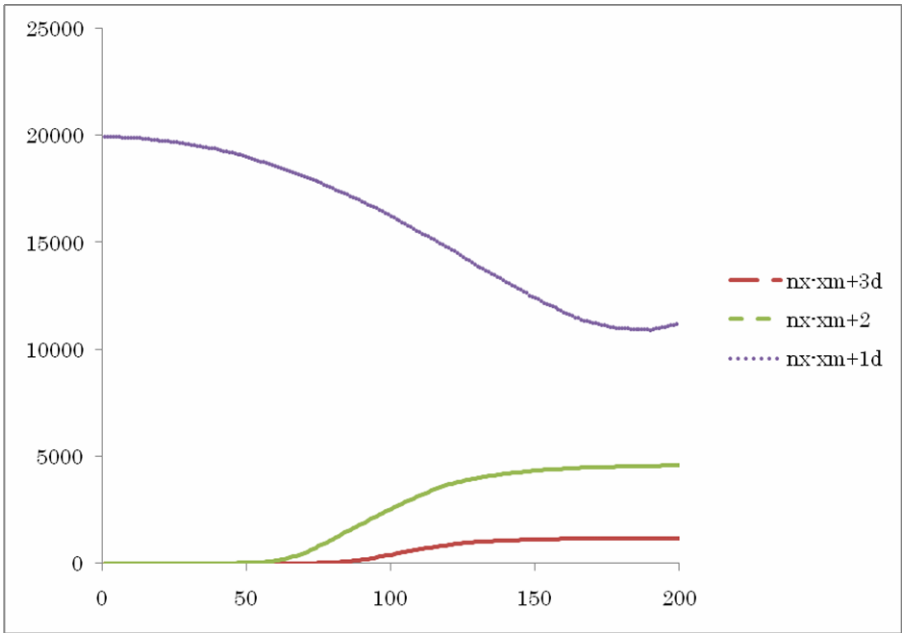

Fig. 2. Learning history and number of excludable outliers. Abbreviations: nx-xm+α.d (number of outliers with) $\varepsilon_i > \varepsilon_{av} + \alpha.\ \sigma\ (\varepsilon)$.

3.3 Role of Data Sets

Solvent accessibility and binding sites employ similar neural networks and hence similar results were obtained by using outlier exclusion criterion. However, target vectors in binding site problem are binary valued, whereas ASA is a real-valued function. Mean absolute error in case of binding sites does not carry much physical meaning like ASA which can take continuous values. Thus the neural network for these problems was also trained to maximize coefficient of correlation between predicted and observed values (data not shown). Variance in the prediction error for these two binary class predictions was found to be smaller than ASA data and no outliers could be detected at α=3. However a value of α=2, was found to be optimum at which significant number of outliers could be removed.

3.4 Biological Basis of Prediction Outliers

Machine learning relies on pattern recognition and a neural network tries to recognize patterns which it has seen during training. Thus if a pattern has not been seen before, the neural network fails to recognize it. Conversely, a pattern which is present in the training data but has no similar patterns in the validation data does not contribute to the performance. A poorly predicted pattern within the training data is just a noise which might tend to over-train the neural network without leading to generalization, thus increasing the unnecessary computational overhead. Furthermore, the nature of relationship between selected features and their target property for some patterns may

not follow a general trend for which neural networks are trained. From a point of view of protein structure, this may be caused by some unusual bonds (e.g. disulfide bond), presence of some ligand in the neighboring region or some features of the biochemical or thermodynamic environment, which is not usually seen by proteins or which cannot be determined from local sequence and evolutionary information.

4 Conclusion

A new algorithm for filtering noisy sequence data from neural network training has been developed which shows promise for applications in RLP predictions. Outlier removal can speed up neural network training without loss of generalization. Different definitions of prediction outliers have been employed and structural basis of the same has been discussed.

References

1. Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., Ofran, Y.: Automatic prediction of protein function. Cell Mol. Life Sci. 60(12), 2637–2650 (2003)
2. Moult, J.: A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr. Opin. Struct. Biol. 15(3), 285–289 (2005)
3. Wolfson, H.J., Shatsky, M., Schneidman-Duhovny, D., Dror, O., Shulman-Peleg, A., Ma, B., Nussinov, R.: From structure to function: methods and applications. Curr. Protein Pept. Sci. 6(2), 171–183 (2005)
4. Schlessinger, A., Rost, B.: Protein flexibility and rigidity predicted from sequence. Proteins 61(1), 115–126 (2005)
5. Nguyen, M.N., Rajapakse, J.C.: Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins 59(1), 30–37 (2005)
6. Ahmad, S., Gromiha, M.M., Sarai, A.: A Real value prediction of solvent accessibility from amino acid sequence. Proteins 50(4), 629–635 (2003)
7. Ahmad, S., Sarai, A.: PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatic 6, 33–35 (2005)
8. Ahmad, S., Gromiha, M., Sarai, A.: Analysis and Prediction of DNA-binding proteins and their binding residues based on Composition, Sequence and Structural Information. Bioinformatics 20, 477–486 (2004)
9. Malik, A., Ahmad, S.: Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Structural B