A Hybrid Machine Learning Algorithm for Complex Sequential Data Classification, Using a Novel Data Representation Method

Antreas Dionysiou, Michalis Agathocleous, Chris Christodoulou, Vasilis Promponas

Abstract— **Trying to extract features from complex sequential data for classification and prediction problems is an extremely difficult task. Deep Machine Learning techniques, such as Convolutional Neural Networks (CNNs), have been exclusively designed to face this class of problems. Support Vector Machines (SVMs) are a powerful technique for general classification problems, regression, and outlier detection. In this paper we present an innovative by design combination of CNNs with SVMs as a solution to the Protein Secondary Structure Prediction problem, with a novel two dimensional (2D) input representation method, where Multiple Sequence Alignment (MSA) profile vectors are placed one under another. This 2D input is used to train the CNNs achieving preliminary results of 80.40% per residue accuracy (Q3), which are expected to increase with the use of larger training datasets and more sophisticated ensembles methods.**

I. PROJECT SUMMARY

Analysis of sequential data, feature extraction and prediction through Machine Learning (ML) algorithms/techniques, has been excessively studied. Nevertheless, the complexity and divergence of the big data that exist nowadays keep this field of research open. When designing ML techniques for complex sequential data prediction, one must take into account, (a) how to capture both short- and long-range sequence correlations [1], and (b) how to focus on the most relevant information in large quantities of data [2].

A CNN is a class of deep, feedforward artificial neural networks that has successfully been applied to analyzing visual imagery [3,4]. Overall, CNNs are in general a good option for feature extraction, immense complexity sequence and pattern recognition problems [3,4,5,6].

SVMs were introduced by Cortes & Vapnik [7], initially for binary classification problems. SVMs are a powerful technique for linearly and non-linearly separable classification problems, regression, and outlier detection, with an intuitive model representation [7].

A challenging task for ML techniques is to make predictions on sequential data that encode high complexity of interdependencies and correlations. Application examples include problems from Bioinformatics such as Protein Secondary Structure Prediction (PSSP) [8]; even though the three dimensional (3D) structure of a protein molecule is determined largely by its amino acid sequence, yet, the understanding of the complex sequence-structure relationship is one of the greatest challenges in computational biology. A ML model designed for such data has to be in position to extract relevant features, and at the same time reveal any long/short range interdependencies in the sequence of data given. In order to maximize the prediction accuracy, the adjacent amino acids have to be considered by the proposed NN architecture.

In this paper we present a hybrid machine learning method based on the application of CNNs in combination with SVMs, and a novel 2D input representation method where the MSA records are placed one under another, for complex sequential data classification and prediction. The implemented model is then tested on the PSSP problem for 3-state secondary structure (SS) prediction.

II. RESULTS AND DISCUSSION

The combination of CNN using the well-known ensembles technique and SVM as a filtering technique, achieves 80.40% Q3 accuracy on CB513 [9] dataset. As a conclusion, we can see that the CNNs can effectively detect and extract features from complex sequential data, by utilizing our proposed “image” like data representation method used to train the CNNs for the PSSP problem. This is due to the fact that our CNN architecture was exclusively designed to face such problems. In addition, SVMs seem to be a good, in terms of filtering, technique to be used for filtering the CNN output.

The combination though, of these two ML algorithms seem to be a great option for complex feature extraction and prediction on sequential data, as we take advantage of the benefits of both techniques. Results are expected to be improved by collecting more experiments for each fold, using larger datasets (e.g., PISCES) and deploying more sophisticated ensembles techniques.

REFERENCES

[1] Graves, A.: arXiv preprint arXiv:1308.0850, 2013.

[2] Blum, A. L., Langley, P.: Artificial Intelligence, 97(1-2), 245-271, 1997.

[3] Krizhevsky, A., Sutskever, I., Hinton, G. E.: In Advances in Neural Information Processing Systems 25: Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada. F. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger, Eds. Red Hook, NY: Curran Associates, 1097-1105, 2012.

[4] Rawat, W., Wang, Z.: Neural Computation, 29(9), 2352-2449, 2017.

[5] LeCun, Y., Bengio, Y.: In The Handbook of Brain Theory and Neural Networks, M. A. Arbib, Ed. Cambridge, MA: MIT Press, 255-258, 1998.

[6] Wang, S., Peng, J., Ma, J., Xu, J.: Scientific Reports, 6, 18962, 2016.

[7] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning, 20(3), 273-297, 1995.

[8] Baldi, P., Brunak, S., Frasconi, P., Soda, G., Pollastri, G.: Bioinformatics, 15(11), 937-946, 1999.

[9] Cuff, JA. & Barton, GJ., Proteins: Structure, Function, and Bioinformatics, 34(4), 508–519, 1999.