Ensemble Deep Learning and Bioinformatics

Literature Review

Bioinformatics is an invaluable hybrid field that brings together the knowledge of biology and computer science. The human mind cannot grasp the large-scale, complex biological data without computational power. In the era of big data, bioinformatics has caught the attention of the academia and industry. To uncover patterns, build models and make predictions from bioinformatics big data, deep learning and other machine learning techniques have been a successful methodology.[1] Deep learning has outstanding achievement in bioinformatics, where the artificial intelligence community had been struggling for many years. However, deep learning has encountered challenges with some bioinformatics core attributes- low sample size and difficulty of getting patient samples for rare diseases, high-dimensionality of data, data noise and heterogeneity etc. [2] To overcome these challenges, researchers are applying machine learning descendant methods named as ensemble deep learning.

Ensemble deep learning, a machine learning technique, has attracted substantial attention from machine learning community and achieved great success in various artificial intelligence applications due to alleviating weights initialization and better generalization. Dietterich in AI magazine listed ensemble learning as the first of the four research directions in machine learning.[3] Since ensemble learning can improve the generalization ability of learning systems, researchers have been working on its theoretical algorithms and applications as early as 1990. The progress of ensemble learning has been remarkable in the last three decades.[2] A couple of recent individual and joint works of deep learning, ensemble learning, and bioinformatics and their achievements and shortcomings are discussed below.

Geddes et al. in “Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis” (2019) paper, proposed an autoencoder-based ensemble clustering and evaluated the method’s validity using different datasets of sc-RNA. They applied random subspace projection due to the higher dimension of the datasets and then applied encoder training. According to the metrics and the datasets used for evaluation, the ensemble of clusters improved performance about 30% on average and the cell type identification is more accurate when ensemble of autoencoder-based clustering framework is being used.[5] However, they failed to make this algorithm optimum across scRNA-seq datasets globally. An interesting insight of the paper is they confirmed using a comparison that autoencoder based cluster ensemble outperforms PCA-based clustering in almost all cases using Wilcoxon Rank Sum test.

P. Mirowski et al. in “Classification of patterns of EEG synchronization for seizure prediction” (2009) paper, compared three types of machine learning classification on publicly available EEG (Electroencephalogram) database. They used logistic regression at first and then, a specific convolutional network architecture which resembles with LeNet5[4]. These two architectures were finally compared with Support Vector Machines (SVM). The classifiers were trained to handle high-dimensional patterns and managed to select subsets of features (channels and frequencies) during the learning process. For predicting correctly, they had to cross-validate each method on EEG recordings which is the main limitation of this study. There was no exact standard indicator for choosing the best mixture of methods. Although they could correctly classify all training samples by predicting all training seizures data, the algorithm predicted some false negatives.[6]

In “A deep learning-based multi-model ensemble method for cancer prediction” (2018) paper by Xiao, Y. et al., they proposed a deep learning-based multi-model ensemble method with a 5-fold stacking and a five-layer neural network. They first applied k-nearest neighbor, support vector machines, decision trees, random forests and gradient boosting decision trees- five classification methods in the first stage individually. To derive the predictions from these methods, they used 5-fold cross-validation technique and averaged the predictions from the earlier stage. They applied a deep learning-based multi-model ensemble method for extracting insights from the datasets and by using the first-stage predictions as features, they had been able to reduce more generalization error than by training them in isolation. To overcome higher computational cost, they carried out a feature selection method in data preprocessing phase, and this resulted in improved prediction accuracy.[7]

Khamparia A. et al. in "A Novel Deep Learning-Based Multi-Model Ensemble Method for The Prediction of Neuromuscular Disorders" (2018) paper, adopted convolutional neural networks (CNN) as the ensemble method to assemble multiple classifiers. The fitness of the method was computed based on the ensemble outputs. They proved using the obtained results that the learning method utilized the clinical sample data more effectively.[8] Similar to other ensemble deep learning models, this increased computational cost. They performed feature selection techniques which reduced the running time and improved prediction accuracy the same as the previous paper mentioned. But because of the huge number of genes, the proposed algorithm could not find a way around the high-dimensional space issue.

Affeldt, S., Laboid, L. & Nadif, M. in "Spectral Clustering via Ensemble Deep Autoencoder Learning (SC-EDAE)" (2019) paper, combined the advantages of deep learning ensemble strategy and proposed a novel clustering method for unsupervised data- Spectral Clustering. They used deep autoencoders (DAE) to reduce the dimension and denoise data, and they rely on the LSC (Linear Spectral Clustering) idea, which depends on the number of landmarks. K-means algorithm is carried upon to initialize LSC landmarks, and this proves to have better accuracy than random initialization. They compared results of SC-EDAE with three popular state-of-the-art methods for using deep-learning algorithms and k-means approach jointly or sequentially: IDEC (Improved Deep Embedded Clustering), DCN (Deep Clustering Network) and DKM (Deep k-means). SC-EDEA outperforms other methods regarding accuracy and normalized mutual information (NMI). The most important feature of SC-EDAE is that it bypasses the need for pretraining of any deep models. They achieved significantly improved performance using a few different encodings, yet more complex datasets would require a larger number of multiple encodings.[9]

In "Deep Unsupervised Clustering Using Mixture of Autoencoders" (2017) paper by Zhang, D. et al., their approach was to use a MIXture of AutoEncoders (MIXAE)- a separate autoencoder to model each data cluster, and thereby the entire dataset as a collection of autoencoders. To model the data cluster, they had to initialize the number of clusters. By training simultaneously with a mixture assignment network via a composite objective function, the autoencoders are proved to have low reconstruction error per manifold and cluster identification error. This kind of joint optimization has been shown to have good performance in other unsupervised architectures as well. Here, for each input data, the concatenated latent features are taken as input by the mixture assignment network. This outputs soft clustering assignments, and the mixture aggregation is done by combining the weighted reconstruction error together with proper regularizations.[10] Although they managed to improve performance on the unbalanced dataset over Deep Embedded Clustering (DEC), the accuracy for different handwritten digit datasets (i.e., MNIST) is lower than that of other methods such SC-EDAE.

Based on this literature research, it can be noted that some ensemble deep learning models are facing challenges with false negative predictions, overall accuracy, an optimal algorithm for global dataset and more. Whereas some models have been able to outperform and achieve better results and predictions in some artificial intelligence applications, but they have not yet been applied to the bioinformatics research area. Again, due to extra processing time and cost overheads compared with traditional ensemble learning, ensemble deep learning algorithms still needs more enhancement in some specific fields, where the time required to development and processing resources are usually restricted or the data to be processed is with large dimensionality such as bioinformatics.

Therefore, we would like to propose a model of mixture of auto-encoders with spectral clustering in bioinformatics. Leveraging a mixture of variational autoencoders, feature selection method, and clustering method may further improve performance on both supervised and unsupervised data and may play a vital role in the discovery of important genes and in entire bioinformatics.

Reference:

1. Min, S., Lee, B. & Yoon, S. "Deep Learning in Bioinformatics" (2017)
2. Cao, Y. et al. "Ensemble deep learning in bioinformatics" Review Paper (2020)
3. Dietterich, TG. “Ensemble Methods in Machine Learning” (2000)
4. LeCun, YA. et al. “Gradient-Based Learning Applied to Document Recognition” (1998)
5. Geddes et al. “Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis” (2019)
6. Mirowski, P. et al. “Classification of patterns of EEG synchronization for seizure prediction” (2009)
7. Xiao, Y. et al. “A deep learning-based multi-model ensemble method for cancer prediction” (2018)
8. Khamparia A. et al. “A Novel Deep Learning-Based Multi-Model Ensemble Method for The Prediction of Neuromuscular Disorders” (2018)
9. Affeldt, S., Laboid, L. & Nadif, M. “Spectral Clustering via Ensemble Deep Autoencoder Learning (SC-EDAE)” (2019)
10. Zhang, D. et al. “Deep Unsupervised Clustering Using Mixture of Autoencoders” (2017)