Applied Soft Computing 62 (2018) 251–258

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Full length article

Deep neural network in QSAR studies using deep belief network

Fahimeh Ghasemi[a], Alireza Mehridehnavi[a], Afshin Fassihi[b], Horacio Pérez-Sánchez[c,*]

[a] Department of Bioinformatic and Systems Biology, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Hezar-Jerib Ave., Isfahan 81746 73461, Islamic Republic of Iran
[b] Department of Medicinal Chemistry, School of Pharmacy and Pharmaceutical Sciences, Isfahan University of Medical Sciences, Hezar-Jerib Ave., Isfahan, Islamic Republic of Iran
[c] Bioinformatics and High Performance Computing Reserch Group (BIO-HPC), Computer Engineering Department, Universidad Católica de Murcia (UCAM), E30107 Murcia, Spain

ARTICLE INFO

Article history:
Received 17 October 2016
Received in revised form 8 August 2017
Accepted 26 September 2017
Available online 18 October 2017

Keywords:
Deep belief network
Biological activity prediction
QSAR
Deep neural network
Drug design

ABSTRACT

There are two major challenges in the current high throughput screening drug design: the large number of descriptors which may also have autocorrelations and, proper parameter initialization in model prediction to avoid over-fitting problem. Deep architecture structures have been recommended to predict the compounds biological activity. Performance of deep neural network is not always acceptable in QSAR studies. This study tries to find a solution to this problem focusing on primary parameter computation. Deep belief network has been getting popular as a deep neural network model generation method in other fields such as image processing. In the current study, deep belief network is exploited to initialize deep neural networks. All fifteen targets of Kaggle data sets containing more than 70 k molecules have been utilized to investigate the model performance. The results revealed that an optimization in parameter initialization will improve the ability of deep neural networks to provide high quality model predictions. The mean and variance of squared correlation for the proposed model and deep neural network are $0.618 \pm 0.407e-4$ and $0.485 \pm 4.82e-4$, respectively. The outputs of this model seem to outperform those of the models obtained from deep neural network.

1. Introduction

Machine learning is a computer programming technique applicable in statistical and mathematical research. It is evolved from the study of pattern recognition and computational learning theory of artificial intelligence. This technique constructs algorithms that they can learn to perform biological activity predictions or data classifications. These objectives are also important in all drug discovery protocols [1].

Over the previous decades, lots of machine learning algorithms have been applied in drug design. Some conventional techniques include: ANN[1] [2,3], KNN[2] [4–6], RF[3] [7], SVM[4] [8–10], MLR[5] [11] one against one [12], Bayes classifier [13] and kernel based methods such as Gaussian process [14,15]. These methods mostly suffer from the same drawbacks, *i.e.* relying on a small number of ligands and a limited selection of descriptors. Therefore, they are called shallow learning techniques. The training of these methods is simple, and they are applicable for few molecules and descriptors. Thanks to several recently developed descriptor-generating softwares, thousands of descriptors are available for a large number of compounds being reported nowadays in literature. The shallow learning techniques are inefficient to model complex relationships between the molecular descriptors. Therefore, deep architecture becomes essential when high amount of data are under process. DL[6] configuration is based on the hierarchical construction in which higher level features are founded on lower level ones. In fact, this approach comprises of multiple levels of linear and nonlinear operations. The number of these operations (depth model) refers to the longest path from an input node to an output one.

In this study, the impact of DNN architecture initialization with DBN technique was considered to predict the biological activities of Kaggle compounds. Huge numbers of element exist in each

* Corresponding author.
[1] ANN: Artificial Neural Networks.
[2] KNN: K-Nearest Neighbors.
[3] RF: Random Forest.
[4] SVM: Support Vector Machine.
[5] MLR: Multiple Linear Regression
[6] DL: Deep Learning.

https://doi.org/10.1016/j.asoc.2017.09.040
1568-4946/

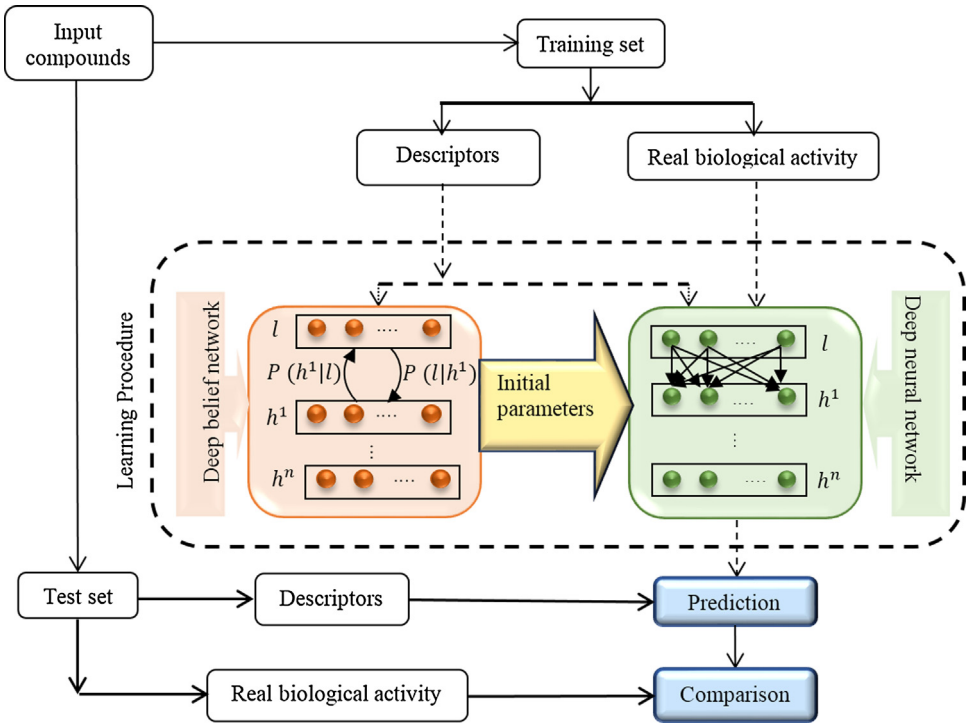

Fig 1. Main stages of proposed Model.

Kaggel dataset. In the smallest data base, there are more than 1500 molecules and 4500 descriptors for each compound. Therefore, neural network with back propagation algorithm is certainly prone to over-fitting. Based on the Hinton suggestion in 2006 [16], DBN was applied to avoid this problem. Thus, first of all, DBN was applied in order to initialize the learning procedure that fine-tunes weights of deep neural networks. These networks are called deep belief network-deep neural network (Fig. 1). In each layer of DBN, restricted Boltzmann machine was utilized. Since the gradient computation of log probability is difficult, contrastive divergence algorithm was performed on the data [26]. In this algorithm, input data should be divided into some batches. The batch size selection can change the results. Furthermore, the number of network layers and nodes directly affects the model construction time and output accuracy. It was expected that more accurate predictions would be made in terms of better correlations in shorter computation time. In addition, the proper selection of the initial parameters may decrease the probability of being stuck in local minima. This means that a more generalized model will be obtained. These advantages are added to the general advantage of DNN which overcomes the over-fitting problem in predictions. This was the reason to decide for a study on the influence of the differences in the number of nodes and layers in the network. The importance of selecting the best batch size in achieving a single set of adjustable parameters that perform well for all data sets was also regarded in the present study.

2. Related works

Recently, the number of biologically active molecules and molecular descriptors has raised exponentially. Parallel to this increment, deep neural network (DNN) which is a multilayer perceptron (MLP) network with many hidden layers and plenty of nodes in each layer could not overcome prone to over-fitting and getting stuck in local minima problems in drug discovery the same as other research area such as image processing and speech processing [1,2]. Hinton et al. [16] introduced a fast and greedy algorithm to improve each layer using RBM.[7] This method was used to initialize a slower learning procedure that fine-tunes the weights using a contrastive version of the wake-sleep algorithm [16]. He showed that this invented algorithm could prevent over-fitting problem. Later, Benjio et al. in 2009 proposed deep architecture, in which single-layer models such as RBM were exploited as unsupervised learning building blocks to construct deeper models such as DBN[8] [17]. After that, Hinton (2012) introduced a practical guide useful to construct RBM algorithm step by step [18]. In 2014, new algorithm was introduced to prevent over-fitting problem by Srivastava named drop-out [19]. Nowadays, DL has been successfully applied in different processing fields such as computer vision, speech processing, image processing and chemo-informatics [20].

In chemo-informatics research, deep learning is applied to capture complex statistical patterns between thousands of descriptors extracted from numerous compounds. Various approaches have been used based on deep architecture. Alessandro Lusci et al. (2013) showed how recursive neural network approaches can be applied to the problem of predicting molecular properties [21]. Restricted Boltzmann machine was used for predicting drug-target interactions by Yuhao Wang and Jianyang Zeng in 2013 [22]. Thomas Unterthiner et al. [23] compared the performance of deep learning approach in seven target prediction methods on ChEMBL database. They found out that deep learning outperformed all other methods with respect to the AUC[9] [23]. Junshui Ma et al.proved that DNN[10] based on the procedure of drop-out can routinely make better prospective predictions than RF on a set of large diverse QSAR[11] data sets [24]. Hughes [25] utilized a database of 702 epoxidation reactions to build a deep machine learning network. Finally, it was

[7] RBM: Restricted Boltzmann Machine.
[8] DBN: Deep Belief Networks.
[9] AUC: Area Under the Curve.
[10] DNN: Deep Neural Networks.
[11] QSAR: Quantitative Structure Activity Relationship.