INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 1)
Available online at www.ijariit.com

DNA Methylation Data Analytics in Cancer Research

Apoorva Patil
Maharashtra Institute of Technology, Pune, Maharashtra
patil.apoorva0211@gmail.com

Rashmi A. Rane
Maharashtra Institute of Technology, Pune, Maharashtra
rashmirane@mitpune.in

Abstract: ***Many studies demonstrated that the DNA methylation, which occurs in the context of a CpG, has a strong correlation with diseases, including cancer. There is a strong interest in analyzing the DNA methylation data to find how to distinguish different subtypes of the tumor [1]. However, the conventional statistical methods are not suitable for analyzing the highly dimensional DNA methylation data with bounded support. DNA methylation is one of the most extensively studied epigenetic marks and is known to be implicated in a wide range of biological processes, including chromosome instability, X-chromosome inactivation, cell differentiation, cancer progression and gene regulation [4]. Identification of cancer subtypes plays an important role in revealing useful insights into disease pathogenesis and advancing personalized therapy. In order to explicitly capture the properties of the data, a deep neural network is used, which composes of several stacked binary restricted Boltzmann machines, to learn the low-dimensional deep features of the DNA methylation data.***

Keywords: DNA Methylation, Deep Neural Networks, Restricted Boltzmann Machine.

I. INTRODUCTION

In life science, the study of DNA is an important factor of understanding about organisms. DNA carries most of the genetic instructions of development, functioning, and reproduction in all organisms. Nowadays, with the development of sequencing technologies, we can easily read a DNA sequence. The amount of data about DNA sequences is also exponentially increasing. For example, the size of Gene Bank, a popular database of DNA sequences, has grown up to more than 2 billion base pairs in December 2015.

It is excellent if we could use these huge data with the power of the modern computer to help us understand about DNA [3]. By using thoroughly understood sequence to train machine learning models, we could use the trained models to predict profile of unknown sequences. In recent years, a new branch of machine learning models called deep learning was introduced. It is a group of models which have multiple non-linear transforming layers used for representing data at successively high-level abstractions. With many layers, these models are expected to be able to solve complicated problems. There are several types of research which applied deep learning models for studying DNA sequences.

DNA methylation is one of the most extensively studied epigenetic marks and is known to be implicated in a wide range of biological processes, including chromo- some instability, X-chromosome inactivation, cell differentiation, cancer progression and gene regulation. Due to the role of methylation patterns in the etiology of complex diseases, DNA methylation analysis becomes a powerful tool in cancer diagnosis, treatment, and prognostication. The high throughput methylation profiling technology, e.g., the Illumina methylation platform, has been developed to survey methylation status of more than 800 cancer-related genes, which makes it is easy to measure genome-wide from limited amounts of DNA and allows measurements in clinical specimens[1]. Currently, there is a strong interest in studying how the methylation profiles can be used to distinguish different subtypes of the tumor. These researches perform unsupervised clustering of large-scale DNA methylation data sets. Formerly, the clustering work focused on the sequence level. Recently, the exact value levels of methylation expression have been fully considered and attract more and more attentions.

II. BACKGROUND

A. Deep Learning

Deep learning is currently at the center of a new revolution in making sense of a large volume of data. Deep learning is an approach to machine learning, aiming at producing end-to-end systems that learn from raw data and perform desired tasks without manual feature engineering. Deep Learning is a class of machine learning algorithm that:

• Use a cascade of many layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
• are based on the (unsupervised) learning of multiple levels of features or representations of the data.
• are part of the broader machine learning field of learning representations of data

• learn multiple levels of representations that correspond to different levels of abstraction.

Deep learning is part of a broader family of machine learning methods based on learning representations of data. An observation can be represented in many ways such as a vector of intensity values per pixel, or in a more abstract way as a set of edges, regions of a particular shape, etc. Some representations are better than others at simplifying the learning task [2]. One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi supervised feature learning and hierarchical feature extraction. Various deep learning architectures such as deep neural networks, convolutional deep neural networks, deep belief networks, and recurrent neural networks have been applied to fields like computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics where they have been shown to produce state-of-the-art results on various tasks.

B. Deep Neural Network (DNN)

A deep neural network (DNN) is an artificial neural network (ANN) with multiple hidden layers of units between the input and output layers. Similar to shallow ANNs, DNNs can model complex non-linear relationships. DNN architectures, e.g., for object detection and parsing, generate compositional models where the object is expressed as a layered composition of image primitives. DNNs are typically designed as feed forward networks, but research has very successfully applied recurrent neural networks, especially LSTM, for applications such as language modeling. Convolutional deep neural networks (CNNs) are used in computer vision where their success is well-documented.

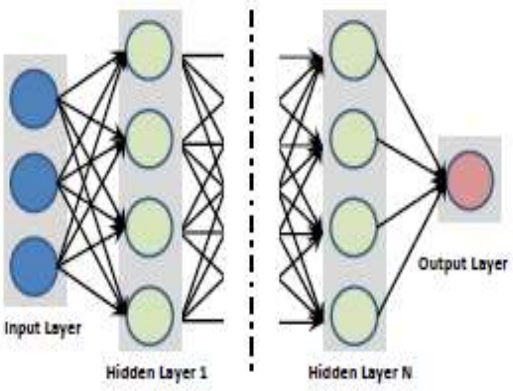

Fig. 1. Deep Neural Network

C. Deep Autoencoder

An autoencoder is an NN designed exactly for extracting features more accurately. Specifically, an autoencoder has the same number of input and output nodes, as in the diagram, and it is trained to recreate the input vector rather than to assign a class label to it. The method is therefore unsupervised. Usually, the number of hidden units is smaller than the input/output layers, which achieve encoding of the data in a lower dimensional space and extract the most discriminative features. If the input data is of high dimensionality, a single hidden layer of an auto encoder may not be sufficient to represent all the data.

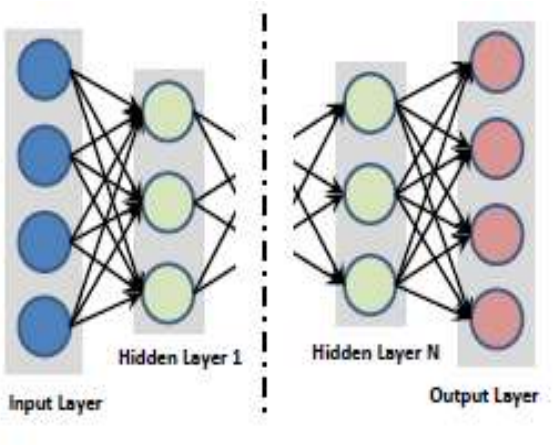

Fig. 2. Deep Auto encoder

D. Deep Belief Network

A DBN can be viewed as a composition of RBMs where each sub-networks hidden layer is connected to the visible layer of the next RBM. DBNs have undirected connections only at the top two layers and direct connections to the lower layers. The initialization of a DBN is obtained through an efficient layer by layer greedy learning strategy using unsupervised learning and is then fine-tuned based on the target outputs.

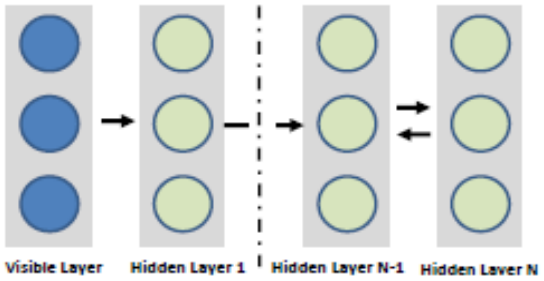

Fig. 3. Deep Belief Network

E. Deep Boltzmann Machine

An RBM was first proposed in and is a variant of the Boltzmann Machine, which is a type of stochastic NN. These networks are modelled by using stochastic units with a specific distribution (for example Gaussian). Learning procedure involves several steps called Gibbs sampling, which gradually adjust the weights to minimize the reconstruction error. Such NNs are useful if it is required to model probabilistic relationships between variables. The main difference with DBN is that the former possesses undirected connections (conditionally independent) between all layers of the network. In this case, calculating the posterior distribution over the hidden units given the visible units cannot be achieved by directly maximizing the likelihood due to interactions between the hidden units.

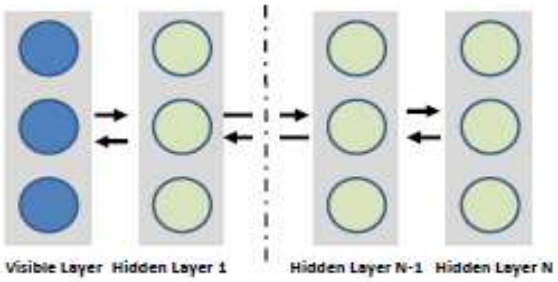

Fig. 4. Deep Boltzmann Machine

F. Recurrent Neural Networks

RNN is an NN that contains hidden units capable of analyzing streams of data. This is important in several applications where the output depends on the previous computations, such as the analysis of text, speech and DNA sequences. The RNN is usually fed with training samples that have strong interdependencies and a meaningful representation to maintain information about what happened in all the previous time steps. The outcome obtained by the network at time t - 1 affects the choice at time t. In this way, RNNs exploit two sources of input, the present, and the recent past, to provide the output of the new data.

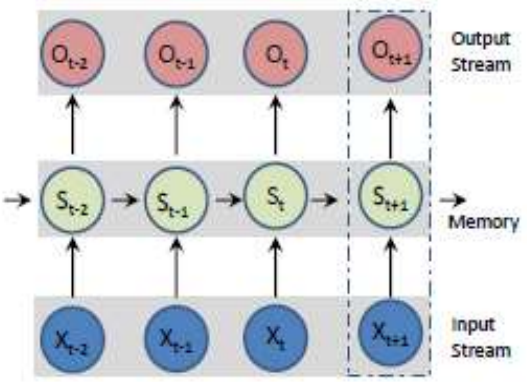

Fig. 5. Recurrent Neural Network

G. Convolutional Neural Networks

A convolutional neural network (CNN) is a type of feed-forward artificial neural net- work in which the connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. Individual cortical neurons respond to stimuli in a restricted region of space known as the receptive field. The receptive fields of different neurons partially overlap such that they tile the visual field. The response of an individual neuron to stimuli within its receptive field can be approximated mathematically by a convolution operation. Convolutional networks were inspired by biological processes and are variations of multilayer perceptrons designed to use minimal amounts of preprocessing. They have wide applications in image and video recognition, recommender systems and natural language processing.

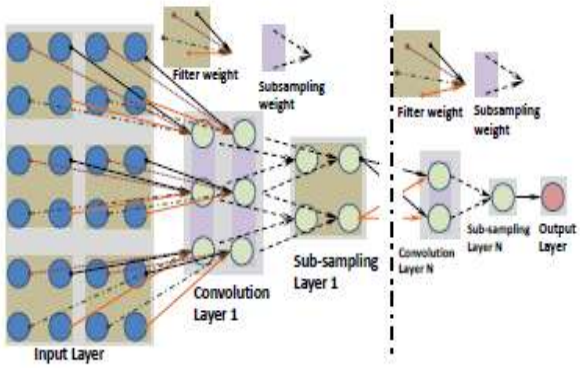

Fig. 6. Convolutional Neural Network

III. PROPOSED WORK

A. Dimension Reduction

Dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction. Feature selection approaches try to find a subset of the original variables (also called features or attributes). There are three strategies: the filter strategy (e.g. information gain), the wrapper strategy (e.g. search guided by accuracy), and the embedded strategy (features are selected to add or be removed while building the model based on the prediction errors). Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Dimension Reduction is done using RBM.

$$P(v,h) = \frac{1}{z}\exp(-E(v,h))$$

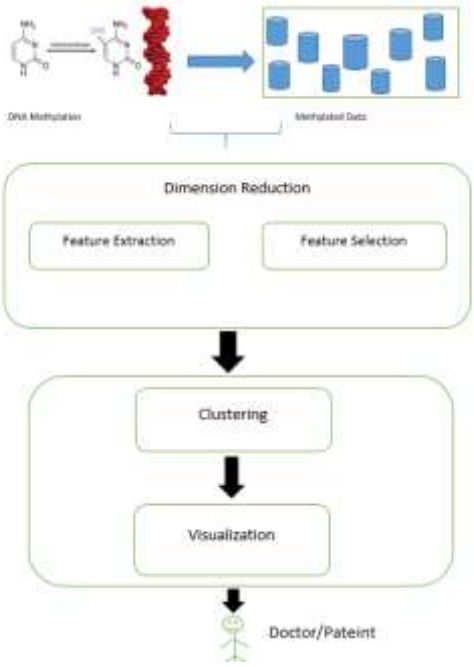

Fig. 7. Architectural Framework

B. Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Clustering is done using Self Organizing Maps (SOM) technique. A self-organizing map (SOM) or self-organizing feature map (SOFM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction. Self-organizing maps differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as back propagation with gradient descent), and in the sense that they use a neighborhood function to preserve the topological properties of the input space.

$$W_v(s+1) = W_v(s) + \theta(u, v, s).\alpha(s).\\ (D(t) - W_v(s))$$

C. Visualization

Visualization is done using the t-SNE technique. It is a nonlinear technique that is particularly well-suited for embedding high-dimensional data into a space of two or three dimensions, which can then be visualized in a scatter plot. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points. First, t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked, whilst dissimilar points have an extremely small probability of being picked. Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the KullbackLeibler divergence between the two distributions with respect to the locations of the points on the map.

IV. CONCLUSION

DNA methylation states are part of a process that changes the DNA expression without changing the DNA sequence itself. This can be brought about by a wide range of reasons, such as chromosome instability, transcription or translation errors, cell differentiation or cancer progression. The datasets are usually high dimensional, heterogeneous, and sometimes unbalanced. The conventional workflow includes data pre-processing/cleaning, feature extraction, model fitting, and evaluation. These methods do not operate on the sequence data directly but they require domain knowledge. For example, the ChEMBL database, used in pharmacogenomics, has millions of com- pounds and compound descriptors associated with a large database of drug targets.

Such databases encode molecular fingerprints and are major sources of information in drug discovery applications. Traditional machine learning approaches have been successful, mostly because the complexity of molecular interactions was reduced by only investigating one or two dimensions of the molecule structure in the feature descriptors. Reducing design complexity inevitably leads to ignoring some relevant but uncaptured aspects of the molecular structures. However, using deep learning to model structural features for DNA methylation can have significant advantages.

V. FUTURE SCOPE

In the prior sections, we discussed some recent applications of Deep Learning algorithms for Big Data Analytics, as well as identified some areas where Deep Learning research needs further exploration to address specific data analysis problems observed in Big Data. Considering the low-maturity of Deep Learning, we note that considerable work remains to be done. In this section, we discuss our insights on some remaining questions in Deep Learning research, especially on work needed for improving machine learning and the formulation of the high-level abstractions and data representations for Big Data. An important problem is whether to utilize the entire Big Data input corpus available when analyzing data with Deep Learning algorithms. The general focus is to apply Deep Learning algorithms to train the high-level data representation patterns based on a portion of the available input corpus, and then utilize the remaining input corpus with the learnt patterns for extracting the data abstractions and representations. In the context of this problem, a question to explore is what volume of input data is generally necessary to train useful (good) data representations by Deep Learning algorithms which can then be generalized for new data in the specific Big Data application domain.

REFERENCES

1. Zhongwei Si, Hong Yu and Zhanyu Ma, "*Learning Deep Features for DNA Methylation Data Analysis*", 10.1109/ACCESS.2016.2576598 IEEE-2016

2. W. Raghupathi and V. Raghupathi, "*Big data analytics in healthcare: promise and potential*", Health Information Science and Systems, Springer-2014.

3. Shaung Ch eng, Maozu Guo, Chunyu Wang, Xiaoyan Liu, Yang Liu and Xuejian Wu, "*miRTDL: a deep learning approach for miRNA target prediction*",2015 IEEE/ACM Transactions on Computational Biology and Bioinformatics.

4. Najmul Ikram, Muhammad Abdul Qadir and Muhammad Tanvir Afzal, "*Investigating Correlation between Protein Sequence Similarity and Semantic Similarity Using Gene Ontology Annotations*", 2017 IEEE/ACM Transactions on Computational Biology and Bioinformatics

5. V. N. Vapnik, "*An overview of statistical learning theory*", IEEE Trans. Neural Netw., vol. 10, no. 5, pp. 988-999.

6. Ngoc Giang Nguyen, Vu Anh Tran, Duc Luu Ngo, Dau Phan, Favorisen Rosyking Lumbanraja, Mohammad Reza Faisal, Bahriddin Abapihi, Mamoru Kubo, Kenji Satou, "*DNA Sequence Classification by Convolutional Neural Networks*", J.Biomedical Science and Engineering, 2016, 9, 280-286

7. Wullianallur Raghupathi, Viju Raghupathi, "*Big data analytics in healthcare: promise and potential*", Health Information Science and Systems, Springer-2014.

8. B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "*Predicting the sequence specicities of DNA-and RNA-binding proteins by deep learning*", Nature Biotechnol., vol. 33, pp. 831-838, 2015.

9. K. Sohn, G. Zhou, C. Lee, and H. Lee, "*Learning and selecting features jointly with point-wise gated Boltzmann machines*", in Proc. Int. Conf. Mach. Learn., 2013, pp. 19
10. Daniele Ravi, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu Perez, Benny Lo, and Guang-Zhong Yang, "*Deep Learning for Health Informatics*", IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 21, NO. 1, JANUARY 2017.
11. Muxuan Liang, Zhizhong Li, Ting Chen and Jianyang Zeng, "*Integrative Data Analysis of Multi-platform Cancer Data with a Multimodal Deep Learning Approach*", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, MAY 2014.
12. A. M. Deaton and A. Bird, "*CpG islands and the regulation of transcription*,"
Genes Develop., vol. 25, pp. 1010-1022, May 2011.
13. K. D. Siegmund, P. W. Laird, and I. A. Laird-Offringa, "A comparison of cluster analysis methods using DNA methylation data," *Bioinformatics*, vol. 20, no. 12, pp. 1896-1904, 2004.
14. D. Peter Augustine, "*Leveraging Big Data Analytics and Hadoop in Developing*
Indias Healthcare Services", International Journal of Computer Applications (0975-
8887), Volume 89 No 16, March 2014.
15. Z. Ma and A. E. Teschendor, "*A vibrational Bayes beta mixture model for feature selection in DNA methylation studies*," J. Bioinform. Comput. Biol., vol. 11, no. 4, p. 135-005, 2013.
16. C. M. Bishop, "*Pattern Recognition and Machine Learning*", Berlin, Germany:
Springer, 2006.
17. B. A. Flusberg et al., "*Direct detection of DNA methylation during single- molecule, real-time sequencing*," Nature Methods, vol. 7, pp. 461-465, Jun. 2010.

