Machine Learning Models and Its Applications in Bioinformatics – A Study

[1]*Rama Devi Chalasani, Research Scholar, KLEF, Vaddeswaram, Guntur*
[2]*Raja Rajeswari.P, Professor, KLEF, Vaddeswaram, Guntur*

Abstract- The Use of computers was drastically increased during the last decade in the Biological and medical research. In this area of research, the focus is now a days is getting shifted from the collection and storage of data to interpretation of the data. The abnormal increase of data demand newer, efficient methods and programs to store, retrieve, process, analyse and formulate the results. Huge biological databases provide challenges and opportunities also. The conventional computer algorithms and methods can solve these problems and most of the times they are unable to handle these problems due to the complexity of the systems. Machine Learning methods are considered to be a best fit for these problems which use huge databases. Machine learning methods like neural networks, Markov Models, support vector machines are proved successful in bioinformatics and found suitable. For building computational models Machine learning also uses statistical theories and produce a result from a sample. In this paper the use of Machine learning in bioinformatics is studied extensively.

Key Words- Bioinformatics, Machine learning, neural networks, support vector machines.

I. INTRODUCTION

Bioinformatics is an interdisciplinary field that uses methods and tools for understanding biological data. Bioinformatics combines biology, computer science, mathematics and statistics to analyse and interpret biological data. The enormous growthin the inflow of data, there arises two problems, one is the storage and the other one retrieval and usage of the data[1]. These problems are critical with regard to Computational biological data which requires tools and softwares for transforming the heterogeneous data in to biological knowledge. The biological data may be like small strings of data and may be to an extent of complex graphs and may be from sequential data or three dimensional structures like protein and RNA structures. With the increasing volumes and complexity of the data the conventional methods and tools of computation may not be sufficient and accurate enough[2]. The high throughput data needs tools and methods with high level capabilities and precision. Hence the scientists and Computer engineers are always in search of high end computational methods and tools. Another problem faced by the analysis of protein and DNA sequences is the redundancy of the data. Many entries in protein or genomic databases represent members of protein and gene families, or versions of homologous genes found in different organisms. Several groups may have submitted the same sequence, and entries can therefore be more or less closely related, if not identical. In the best case, the annotation of these very similar sequences will indeed be close to identical, but significant differences may reflect genuine organism or tissue specific variation. In sequencing projects redundancy is typically generated by the different experimental approaches themselves. Hence it is observed that the repositories should be used in an appropriate and a suitable manner. The bioinformatics approaches till recent times used and still using tools that can process the data in a desired manner. But the increasing volumes and the increasing complexity of data demands applications with higher computing capabilities and accuracy[3]. Hence scientists always look for newer technologies and approaches. One technology or methodology that comes to the minds of the scientists is Machine learning.

Machine Learning in Bioinformatics:

With the increasing computational requirements in Computational biology and bioinformatics, the use of Machine learning techniques is now a days very much increased. Innovative methods of computation to process high potential data in a variety of formats like sequences, pathways and protein and gene expressions have become important for identifying, understanding diseases. Machine learning methods such as Markov models, neural networks, support vector machines and graphical models are proved successful in processing biological data which is more complex in nature. Artificial Intelligence and machine learning techniques are extensively used in the bioinformatics domain for discovery and mining of knowledge. Machine learning uses the example data or previous experience to optimize a performance criterion[4].

A machine learning algorithm can learn from experience with respect to some class of tasks and a performance measure. These are suitable for molecular biology data due to the learning algorithm's ability to construct hypotheses that can explain complex relationships in the data. The classifiers or hypotheses can then be interpreted by a domain expert who suggests some wet-lab experiments to validate or refute the hypotheses.

There are two types of learning schemes in machine learning:

Supervised Learning: The output has been given *a priori* labelled or the learner has some prior knowledge of the data and

Unsupervised Learning: where no prior information is given to the learner regarding the data or theoutput.

Machine Learning Algorithms

Types of Machine Learning Algorithms

There are 3 type of Machine Learning Algorithms. They are

1. Supervised Learning:
This algorithm consist of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data. Examples of Supervised Learning are Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc.

2. Unsupervised Learning
In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning are Apriori algorithm, K-means.

3. Reinforcement Learning
Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning Markov Decision Process.

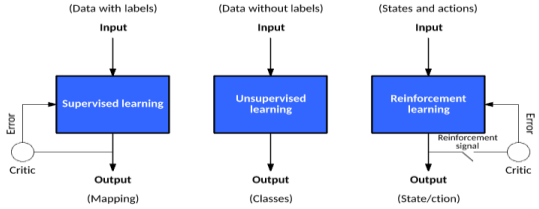

Fig.1: The Three Types of Machine Learning

Machine Learning Models

1. Neural networks
A neural network processes an input vector to a resulting output vector through a model inspired by neurons and their connectivity in the brain. The model consists of layers of neurons interconnected through weights that alter the importance of certain inputs over others. Each neuron includes an activation function that determines the output of the neuron[5]. The output is computed by applying the input vector to the input layer of the network, then computing the outputs of each neuron through the network (in a feed-forward fashion).

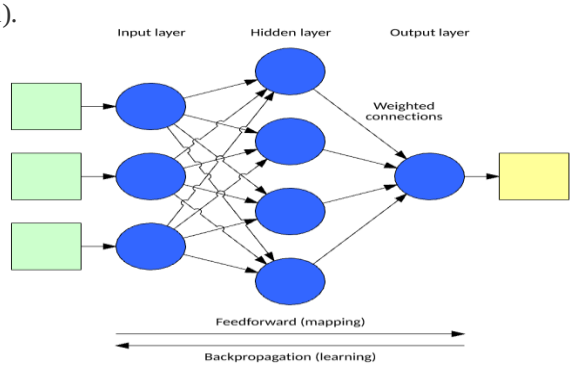

Fig.2: Layers of a typical neural network

2. Decision Trees:-
A decision tree is a supervised learning method for classification. Algorithms of this variety create trees that predict the result of an input vector based on decision rules inferred from the features present in the data. Decision trees are useful because they're easy to visualize so you can understand the factors that lead to a result.
Two types of models exist for decision trees: classification trees, where the target variable is a discrete value and the leaves represent class labels and regression trees, where the target variable can take continuous values.

3. K-means clustering:-
k-means clustering is a simple and popular clustering algorithm that originated in signal processing. The goal of the algorithm is to partition examples from a data set into k clusters. Each example is a numerical vector that allows the distance between vectors to be calculated as a Euclidean distance.
The simple example below visualizes the partitioning of data into $k = 2$ clusters, where the Euclidean distance between examples is smallest to the centroid of the cluster, which indicates its membership.

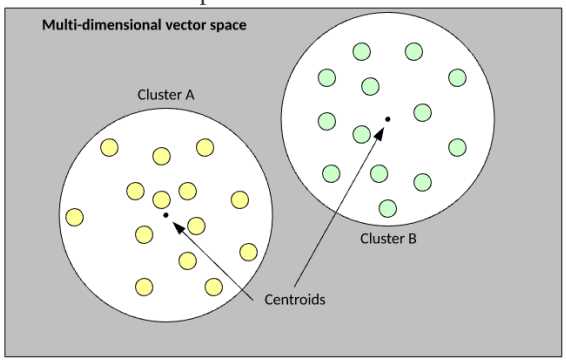

Fig.3: *k*-means clustering

The *k*-means algorithm is extremely simple to understand and implement. If the member is closer to the alternate cluster, the example is moved to the new cluster and its centroid recalculated. This process continues until no example moves to the alternate cluster.

4. Q-learning:-
Q-learning is one approach to reinforcement learning that incorporates *Q* values for each state–action pair that indicate the reward to following a given state path. The general algorithm for Q-learning is to learn rewards in an environment in stages. Each state encompasses taking actions for states until a goal state is reached[6]. During learning, actions selected are done so probabilistically (as a function of the *Q* values), which allows exploration of the state-action space. When the goal state is reached, the process begins again, starting from some initial position.
Q values are updated for each state–action pair as an action is selected for a given state. The *Q* value for the state–action pair is updated with some reward provided by the move (may be nothing) along with the maximum *Q* value available for the new state reached by applying the action to the current state. This is further discounted by a learning rate that determines

how valuable new information is over old. The discount factor indicates how important future rewards are over short-term rewards. Note that the environment may be filled with negative and positive rewards, or only the goal state may indicate a reward.

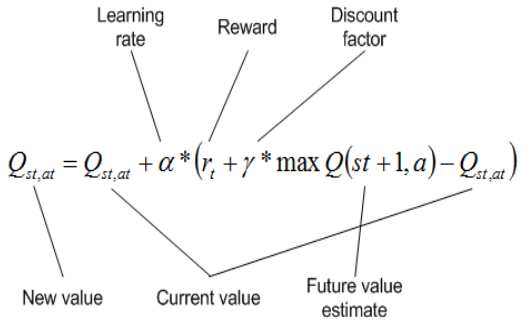

Fig.4: A typical Q-learning algorithm

II. CONCLUSION

Machine learning is a very useful domain for use in Bioinformatics research which involves complex data. Various models of the machine learning can be used in different context of the Bioinformatics Research.

III. REFERENCES

[1]. Baldi P, Brunak S (2001). Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge. ISBN 0-262-02506-X.

[2]. Hastie T, Tibshirani R, Friedman JH (2001). The Elements of Statistical Learning.SpringerVerlag, New York. ISBN 0-387-95284-5.

[3]. Cohen, A.M. &Hersh, W.R. (2005). A survey of current work in biomedical text mining. Briefings in Bioinformatics, 6(1), 57–71.

[4]. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., &Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.AAAI Press/MIT Press, Menlo Park, California, USA.

[5]. Hunter, L. (2004). Life and Its Molecules: A Brief Introduction. AI Magazine, 25(1), 9-22.

[6]. Yeung, Y.K., Medvedovic, M. &Bumgarner, R.E. (2003). Clustering Gene-Expression Data with Repeated Measurements.Genome Biology, 4(5), R34.