International Journal of Engineering &Technology, 7 (1.7) (2018) 201-204

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Study on machine learning methods and applications in genetics and genomics

K. Jayanthi [1*], C. Mahesh [2]

[1]*Research Scholar, Department of Computer Science and Engineering, School of Computing, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai-62, TamilNadu, India.*

[2]*Associate professor, Department of Information Technology Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai-62, TamilNadu, India.*
**E-Mail: jayanthi2contact@gmail.com*

Abstract

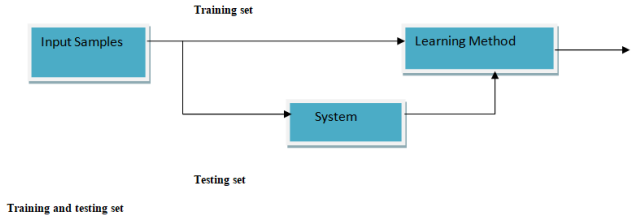
Machine learning enables computers to help humans in analysing knowledge from large, complex data sets. One of the complex data is genetics and genomic data which needs to analyse various set of functions automatically by the computers. Hope this machine learning methods can provide more useful for making these data for further usage like gene prediction, gene expression, gene ontology, gene finding, gene editing and etc. The purpose of this study is to explore some machine learning applications and algorithms to genetic and genomic data. At the end of this study we conclude the following topics classifications of machine learning problems: supervised, unsupervised and semi supervised, which type of method is suitable for various problems in genomics, applications of machine learning and future views of machine learning in genomics.

Keywords: *Machine Learning Methods, Genomics Classification Problems, Future Application Of Genomics.*

1. Introduction

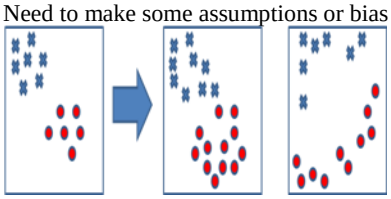
Machine learning is one of an application of artificial intelligence that make systems should have the ability to learn automatically and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves [1]. For example in figure.1 consider a data set D, with task T and Performance measure M, a system is intention to learn from D to perform the task T if after learning the system's performance on T improves as measured by M.

Learning system model



Training and testing set

- Training is the process of making the system able to learn.
- Training set and testing set come from the same distribution
- Need to make some assumptions or bias



In genomics, machine learning can be used to learn how to extracting the location and structure of various genes, to identify regulatory elements, to identifying non- coding RNA genes, to predicting gene function, to predicting RNA secondary structure. Here we use machine learning methods can used to make computers to learn from this scenario for analysing the locations of transcription start sites (TSSs) in a genomic sequence [2]. It has three process stages [11].

1. First develop a machine learning algorithm that will leads to successful learning.

2. In that algorithm, we provide a large collection of TSS sequences as well as optionally a list of sequences that are known not to be TSS, when annotating that identifies whether a sequence is TSS or not is known as the label. By using the algorithm it process these labelled sequences and stores a model.

3. Unlabelled sequences are collectively given as the input to an algorithm again and it uses the model to predict labels for each sequence. If the learning was successful,

then all or most of the predicted labels will be correct, if the labels associated with test set examples are known. i.e if these the learning system then the performance of the machine learning algorithm can be assessed immediately.

Example of a machine learning application with DNA sequence

In Figure.2 shows DNA sequences training set is provided as input to learning procedures, along with labels indicating whether each sequence is centred on a TSS or not. A model can be produced by machine learning algorithm which can then be subsequently used with a prediction algorithm to assign predicted labels to unlabeled test sequences.

2.

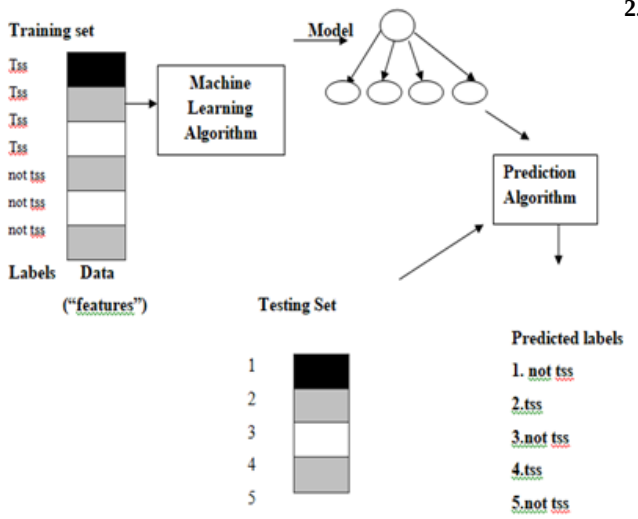


Fig. 2: Machine Learning

1.1 Application of Machine learning in genomics and genetics:

To annotate a huge variety of genome sequencing elements we can use machine learning methods. Generally if we can compile a list of sequence elements of a given type, then we can probably train a machine learning method to recognize those elements, then models can be combined along with logic about their relative locations. Many application areas that are related to machine learning such as gene sequence, gene expression, protein structure, gene regulatory networks, microarrays. Before trying to solve the problems in the above area of genomics we use machine learning algorithm to train the system to classify the gene data and create a model. Some modern techniques to handle the gene data can be used. There are so many machine learning algorithms can be used for identification, prediction, selection, recognization, and also in classification of DNA sequences. For the identification of gene in DNA sequence the neural network based multiclassifier can be used. Gene expression can be identified by promoters. To predicting the location of the promoter promoter neural network has been used. For predicting the promoters in the genes and evaluating the performance of the gene the artificial neural network classifier has been used. In a genome research, the vital part is to learning to DNA sequences pattern recognization. Machine learning can take as input data generated by other genomic assays, such as microarray or RNA – Sequence expression data, transcription factor, binding chip - sequence data, etc. Another example of genome data is gene expression data can be used to learn to distinguish between different disease phenotypes and in the process, to identify potentially valuable disease biomarkers. We can also use machine learning to assign annotations to genes these kind of annotations maximum taken from the gene ontology assignment terms[3].

Gene expression can predicted by various machine learning method (promoters) from huge variety of gene data set. To predict the expression of a gene based solely on the DNA sequence [4]. These are some techniques are available to generate jointly model of the expression of all genes in a cell by training a network model can do this kind of genomics data.[5] maximum number of problems in genomic research can be solved by various techniques drawn from field of statistics.

1.2 Scope of this study

In this paper we mainly discuss the following topics such as major classification of machine learning problems, which type of method should be used for genomics, applications of machine learning methods with feature selection, handling missing data, future of machine learning in genomics.

3. Classification of Machine Learning Problems

a) Supervised Learning

Supervised learning algorithms can apply what has been learned in the past to new data using labelled examples to predict future events starting from analysis of a known training dataset that learning algorithms produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with correct, intended output and find errors in order to modify the model accordingly. In Figure.3 shows the supervised learning architecture.

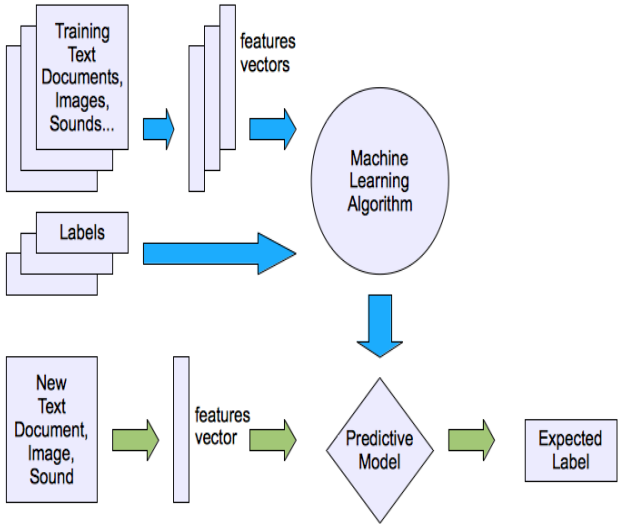


Figure.3: Supervised Machine learning structure

Low E-out or maximize probabilistic terms

$$error = \frac{1}{N}\sum_{n=1}^{N}[y_n \neq g(x_n)]$$

E-in: for training set
E-out: for testing set

$$Eout(g) \leq Ein(g) \pm O\left(\sqrt{\frac{d_{VC}}{N}\ln N}\right)$$

Consider an example for finding genes here algorithm is to predict the locations and detailed intron/exon structure of all the protein – coding genes on the chromosomes. For gene finding requires input a training set of labelled DNA sequences by supervised learning. Where the labels specify the locations of the start and end of the gene (TSS and not TSS). After

that model the training data can be used to learn general properties of genes for example what DNA sequence pattern, the trained model the use these learned properties to identify novel genes that resemble the genes in the training set.

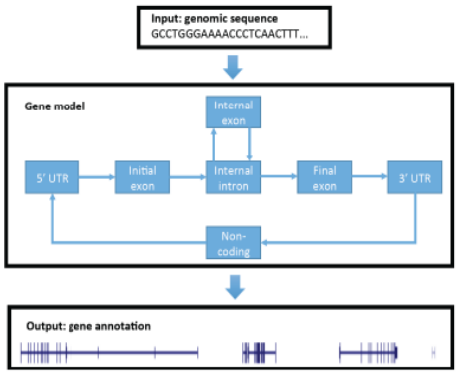


Fig.4: A gene finding model-Simplified

In figure.4 shows a gene finding model it captures the protein-coding gene basic properties.DNA sequence of a chromosome or a portion thereof can taken as model based on that it produces as output of detailed gene annotations.

b) Unsupervised Learning

In unsupervised algorithms from unlabelled data to describe a hidden structure. When the system is trying to figure out the output its does not provide the right output, but it explores the data and can inferences can be drawn from data sets to describe the hidden structures from unlabelled data. In this method the user only have input data(X) and there is no labels. The data given for learning model has to be analysed more. Unsupervised algorithm can use these methods Clustering, Probability distribution estimation, Finding association (in features),Dimension reduction for creating labels from the given inputs.

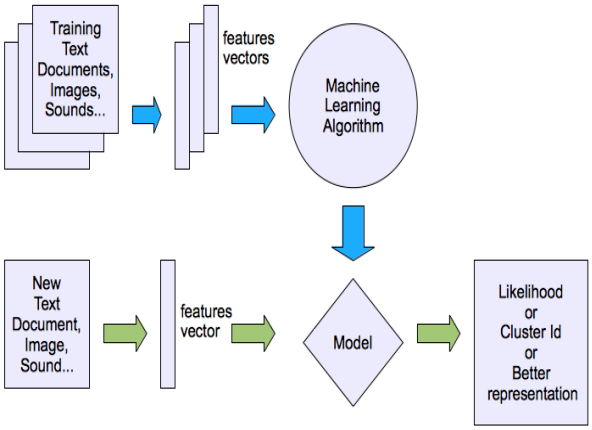
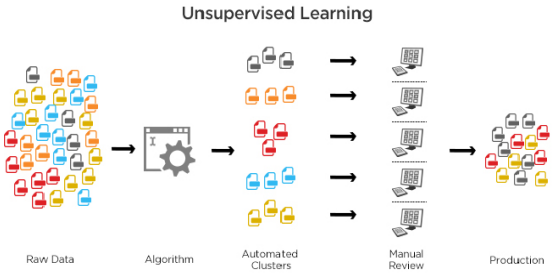


Fig.5: Unsupervised Machine learning structure

Unsupervised learning algorithms example says k- means for clustering problems, Apriori algorithm for association rule learning problems. Instead of imposing a pre- determined set of labels on the data, can allow computer system to predict what types of labels are possible that can of prediction must be determined unsupervised rather than supervised learning. This type of approach, the machine algorithm takes only unlabelled data and the desired number of different labels to assign. [6, 7, 8].Now the machine learning algorithm then automatically separates the genome into segments and assigns a label to each and every segment, by assigning the same label to segments can provide the similar group of data. The main advantage of this unsupervised approach is providing the ability to train when labelled examples are unavailable and the ability to identify novel types of genomic elements.

c) Semi-supervised learning

It is a combination of both supervised and unsupervised learning [22]. In supervised learning, a collection of data points received by machine learning algorithm, everything with an associated label, but in unsupervised learning no labels can be received for an algorithm. In semi – supervised learning does not require any labels but some cases it accepts labels for predicting the features labels.



Using a semi-supervised approach, gene findings are maximum trained in genomics data where collection of input annotated genes plus a complete (unlabeled) genome sequence. The learning procedure starts by creating an initial gene finding model based solely on the labelled subset of the input training data. The constructed model can then been used to scan the genome, now we get some tentative labels that are assigned throughout the genome. By using utilizing the tentative labels we can improve the learned model which has modelled already. These kinds of procedures are iterated until no new genes are found. We can say that compare with supervised learning method semi supervised is better because the model is able to learn more from the large collection of gene data, all genes in the genome, rather than only the subset of genes that have been identified with high confidence and more of iteration can lead to perfect learned model for further analysis.

Which type of method can use to solve genomic data?

So far we have seen about various approaches to solve machine learning in genomics data, if we want to solve a new machine learning task, the first question is often whether to use a supervised, unsupervised or semi-supervised approach. In that moment we conclude to this question is obvious. Based on the label availability we can choose appropriate machine learning approach. For supervised we have proper labels then only we properly learn the model and one more notified thing is that always not to choose blindly that having labels then go with supervised learning its not a good idea. In this approach planning to train an algorithm to work with training set that should generated differently from the testing data to which the trained model will eventually be applied. In a gene finder using a training set of human genes will not probably work at finding all genes. In supervised learning should be applicable only in the case where training as well as test data both are expected to generate similar statistical properties. If your genomic data is likely to be large in size that we could not find our labels for training sets to model that time need of clustering algorithms for grouping similar data to find the gene there we can choose unsupervised learning approach.

4. Future of machine learning in genomics

1. As genomic data relatively large in size so machine learning approaches can make that to be easily analyse and make the things as simplified.

2. Gene sequencing can be very easy to analyse only by using machine learning methods. The sequence of the various genes must have labels so using supervised learning algorithms easily

gets the sequence accurately, Deep Genomics maximum the machine learning algorithms are used to for gene sequencing.

3.Gene Editing one of the main research area in the next generation for analysing the genes and finding the exact matches in genes and changes the gene sequence according to the way its need to targeted. Especially gene editing with machine learning can reduce the time, cost and effort needed to identify the targeted sequence.

4. Pharmacogenomics field provides more advantage for initiating the personalized medicine that is the drug which is given to the patient for a particular disease that should adapt to the genetic makeup of the individual patient. Identifying the dose of drug machine learning can achieve a mass challenge in future.

5. New Born genetic screening tools can make use of machine learning approaches in identifying the metabolism defects.

5. Conclusion

This study paper is about machine learning methods and where we can use these approaches in problems of genomics. Mainly discuss the following topics such as major classification of machine learning problems, which type of method should be used for genomics, applications of machine learning and future of machine learning in genomics.

References

[1] Mitchell, T. Machine Learning. McGraw-Hill; 1997.
[2] Ohler W, Liao C, Niemann H, Rubin GM. Computational analysis of core promoters in the drosophila genome. GenomeBiology. 2002;
[3] Picardi E, Pesole G. Computational methods for ab initio and comparative gene finding. Methods in Molecular Biology. 2010; 609:269–284. [PubMed: 20221925]
[4] Gene Ontology Consortium. Gene ontology: tool for the unification of biology. Nature Genetics. 2000; 25:25–29. [PubMed: 10802651]
[5] Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell. 2004; 117:185–198. [PubMed: 15084257]
[6] Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 2004; 303:799–805. [PubMed: 14764868]
[7] Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, Noble WS. Unsupervised segmentation of continuous genomic data. Bioinformatics. 2007; 23:1424–1426. [PubMed: 17384021]
[8] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012; 9:215–216. [PubMed: 22373907]
[9] Hoffman MM, et al. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012; 9:473–476. [PubMed: 22426492]
[10] Lanckriet GRG, Bie TD, Cristianini N, Jordan MI, Noble WS. A statistical framework for genomic data fusion. Bioinformatics. 2004; 20:2626–2635. [PubMed: 15130933]
[11] W. Libbrecht "Machine learning in genetics and genomics Maxwell" Nat Rev Genet. Author manuscript; available in PMC 2017 January 02.
[12] http://graveleylab.cam.uchc.edu/WebData/mduff/MEDS_6498_SPRING_2016/machine_learning_genomics_Noble_2015.pdf