DNA SEQUENCE CLASSIFICATION USING MACHINE LEARNING

Hima M S
PG Student
Dept. of computer Applications
Amal Jyothi college of Engineering
Kanjirappally, Kottayam
2018himams@gmail.com

Ms. Rini Kurian.
Asst. Professor
Department of Computer Applications
Amal Jyothi college of Engineering
Kanjirappally, Kottayam
rinikurian@amaljyothi.ac.in

***Abstract*—Machine learning is a data processing technology that uses training data to help make judgments, predictions, classifications, and recognitions. It is a subset of Artificial Intelligence that performs tasks without regard to any theory. Agriculture, computer vision, the gaming industry, and linguistics are some of the uses of machine learning. DNA classification is one of the most relevant topic today's. Therefore, understanding those data related to DNA is very important to improve prediction accuracy. The goal of this research is to see if machine learning techniques like natural language processing and the Nave Bayes algorithm can be used to predict DNA vulnerability. The performance of models developed using the Nave Bayes Classifier Algorithm and Natural Language Processing, as well as the k-mer counting approach, was compared in this article. The datasets are categorized into sequences and classes. The performance of models developed using the Nave Bayes Classifier Algorithm and Natural Language Processing, as well as the k-mer counting approach, was compared in this article. The datasets are categorised into sequences and classes.**

Keywords:

Natural language processing, k-mer counting, Classification Techniques, Machine Learning Algorithms.

I. INTRODUCTION

Machine learning is a broad field of artificial intelligence that focuses on improving performance over time and identifying patterns in large amounts of data. Machine learning is a subset of Artificial Intelligence that allows machines to learn from data and perform real-world tasks intelligently ANN, CNN, Deep Learning, Genetic Algorithms, and other subcategories of machine learning exist. When it comes to neural networks, they may be used to identify subsets of data in a sequence and can also be used to identify individual data. Machine learning is being used in a variety of situations., from foundational research in computational chemistry, quantitative anthropology, physics, agro, computer vision, gaming, and semantics to bioinformatics and physics.

The DNA sequence expresses a variety of characteristics about a species, including habits, looks, and information about their parents. These details aid in distinguishing a species from other species. As a result, determining the order of DNA sequences and their classification is critical in today's world. The purpose of DNA sequencing is to determine the nucleotide order of a specific DNA region.

The k-mer counting approach can be used to categories DNA sequences.. There are millions of genes in a single DNA sample. As a result, classifying this is a difficult task. These days, the majority of the time, sequencing technologies are used to accomplish that work.

Machine learning is a relatively new tool for addressing this issue. Recent machine learning techniques used to tackle this issue include ANN, CNN, Deep Learning, and Genetic Algorithms. The input sequence of DNA to these models and obtain the results Similarly, several Machine learning features can be utilized to make decisions. The categorization of DNA sequences These models are used to improve the accuracy of the categorization approach. This model has the ability to predict as well as classify. These models were created using the Python programming language.. The Nave Bayes algorithm and Natural Language Processing (NLP) are two of the most basic types of algorithms that are also the easiest to implement. The Nave Bayes Classifier is based on the Bayes Theorem and the concept of probability. It frequently plays an important part in the decision-making process.

II LITERATURE SURVEY

This section examines and evaluates some papers that are connected to similar investigations. Machine Learning Algorithms for DNA Classification Prediction Maintaining the Specifications' Integrity. Apply a classification model that can predict the function of a gene solely based on the coding sequence's DNA sequence. [4]

Hemaletha Garunanth et.al[1]proposed In a generic computational framework for biomedical data processing, DNA sequence categorization is a key task. In this study, we used CNN, CNN-LSTM, and CNN-Bidirectional LSTM architectures with Label and k-mer encoding for DNA sequence categorization.

Naglaa Fathi Soliman et.al proposed Classification of DNA Sequences Deep learning (DL) methods have shown to be extremely effective in tackling a wide range of issues in a variety of sectors, particularly in the area of large data. With the advancements of the big data age in bioinformatics, DNA sequences can be identified with accurate and sca prediction using DL approaches. [2]

DNA is a deoxyribonucleic acid (DNA)-based biological macromolecule with the primary function of data storage. Due to breakthroughs in sequencing technology, DNA sequence data is currently accumulating at an exponential

rate. This has ushered in the era of big data in the study of DNA sequences. Machine learning is also a powerful technique for digesting large volumes of data and learning on its own. [2014]

Unlike previous work on heart DNA classifications, this research proposes a comparison technique for effective heart DNA classification utilizing Nave Bayes and a Natural Language Processing model..

III. DNA SEQUENCING

The DNNA sequencing is an important method of today's. The study of DNA sequence data is a major focus of bioinformatics. When we talk about DNA sequencing, we're talking about the process of determining which genes are present in a person's DNA. The order in which nucleotides appear in a nucleic acid sequence. The phrase classification refers to the classification of nucleic acids. or their combinations, which are referred to as genes, into various categories sections.

In bioinformatics, analyzing and interpreting DNA data are two of the most difficult tasks. The main tools for addressing these tasks are classification and prediction methods. The approaches for classifying DNA sequences are divided into three categories in [9]. Distance-based methods, feature-based methods, and mode-based approaches are the three types.

Heart disease risk factor include:

- G protein coupled receptors
- Tyrosine kinas
- Tyrosine posthaste
- Synthesize
- Lon channel
- Transcription factor

k-mer length overlapping "words".: "ATGCATGCA

- 'ATGCAT',
- 'TGCATG',
- 'GCATGC',
- 'CATGCA'.

IV.METHODOLOGY

A. Data Source

The dataset for gene prediction was obtained from kaggle. The kaggle database is a collection of datasets for implementing machine learning techniques. The data set is mostly made up of two parameters. The sequence parameter, as well as the types of sequences that it corresponds to.

A). Natural Language Processing(NLP)

NLP algorithms frequently employ machine learning algorithms. Instead of manually coding large sets of rules, NLP can utilize machine learning to automatically learn these rules by reviewing a set of examples (for example, a large corpus, such as a book, and breaking it down into a collection of phrases) and generating a statistical conclusion. .

1)*Algorithm 1*: A classification approach that can predict the function of a gene solely based on the coding sequence's DNA sequence.

a) Step 1: Apply dataset and read data .

Step 2: predict gene's function.

Step 3: convert to vector with uniform count

Step 4: using k-mer counting method and .divide data to hemaers words.

Step 5. Aplly bag of words using CountVectorizer with the help of NLP.

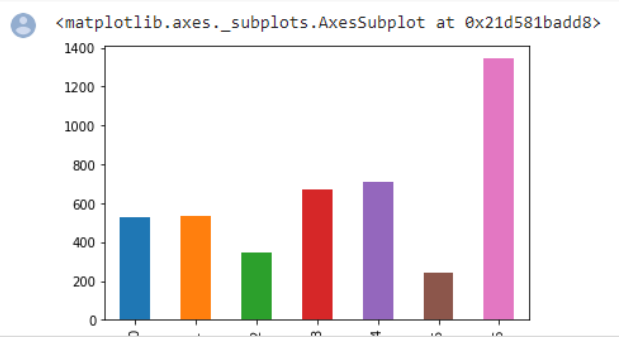
b) Experimental Results: These types of manipulations are referred to in genomics as "k-mer counting," or counting the occurrences of each potential k-mer sequence. Although there are specialist tools for this, Python's natural language processing tools make it extremely simple.Any sequence (string) can be converted to overlapping k-mer words with this function:

	class	words
0	4	[atgccc, tgcccc, gcccca, ccccaa, cccaac, ccaac...
1	4	[atgaac, tgaacg, gaacga, aacgaa, acgaaa, cgaaa...
2	3	[atgtgt, tgtgtg, gtgtgg, tgtggc, gtggca, tggca...
3	3	[atgtgt, tgtgtg, gtgtgg, tgtggc, gtggca, tggca...
4	3	[atgcaa, tgcaac, gcaaca, caacag, aacagc, acagc...

Figure[1]converted image using k-mer counting

```
(4380, 232414)
(1682, 232414)
(820, 232414)
```

Figure[2]vector form.



Figure[4]classified gene's

2) *Naive Bayes Classifier*

Based on the Bayes Theorem A statistically based classifier is the Naive Bayes classifier. It is assumed that the qualities are statistically independent. This classifier is built using probabilities. Given two occurrences A and B, Bayes theorem states that P (A) is the prior probability and P (A|B) is the posterior probability.

P (A|B) = P (B/A) P (A)/P (B) and P (B|A) is computed as P (A ∩ B) =P (A)

The most crucial component in identifying the most likely next occurrence for a given instance based on all of the training data is Bayesian probabilities. Conditional probabilities are calculated using the training dataThe conditional independence model of each predictor determines the target class in the Naive Bayes model [5]. This classifier provides the most accurate prediction (given the assumptions). It can also deal with attribute values that are numeric or discrete. The algorithm for this method is shown in Algorithm 1 below.

c)Algorithm 2: Find the Accuracy ,precision,recall and f1-score

Step 1: It will be developed a multinomial naive Bayes classifier.

Step 2: The parameter tuning and found ngram size.

Step 3: using multinomialNB to create a model alpha 0.1 like grid search.

Step 4: Confusion matrix, accuracy, precision, recall, and f1 score

b) Experimental Results: Accuracy recorded by our model is 0.984 The precision value ,recall and f1-score is same.. If both have the same value, sensitivity equals specificity, and so accuray equals sensitivity. We're getting pretty good results on our unseen data, so it doesn't appear that our model overfitted to the training data. Because we have a tiny data set, I would go back and sample many more train test splits in a real project.

```
Confusion matrix

Predicted   0    1   2    3    4   5    6
Actual
0          99    0   0    0    1   0    2
1           0  104   0    0    0   0    2
2           0    0  78    0    0   0    0
3           0    0   0  124    0   0    1
4           1    0   0    0  143   0    5
5           0    0   0    0    0  51    0
6           1    0   0    1    0   0  263
accuracy = 0.984
precision = 0.984
recall = 0.984
f1 = 0.984
```

Figure[3]confusion metrix

V. CONCLUSION

Since the introduction of machine learning, a variety of systems have grown more familiar to users and easier to use as a result of the various applications that have been developed. People became increasingly interested in machine learning after that. This research uncovers its application in the field of DNA sequence classification. There were several obstacles, issues, and downsides encountered during this procedure, only a few of which are mentioned here. Various people have conducted various studies in order to overcome these obstacles and have attempted to develop new thoughts and ways for implementing this procedure with greater efficiency and precision.

REFERENS

[1] Hemalatha Gunasekaran,1 K. Ramalakshmi,2 A. Rex Macedo Arokiaraj,1 S. Deepa Kanmani,3 Chandran Venkatesan,4 and C. Suresh Gnana Dhasekaran 5 Classification of DNA Sequences Using CNN and Hybrid Models[2020]

[2] Naglaa. F. Soliman°,Salah Eldin S. E. Abdulrahman** , Nabil A. Ismail**, Fathi E. Abd El- DNA Sequences Classification with Deep Learning[2010]

[3] Timothy Cohen, Bryan Ostdiek, Spencer Chang, What is the Machine Learning?," vol. 97, no. 5, p. 6, 2018.

[4] Zhaoli, "Machine Learning in Bioinformatics," in . International Conference on Computer Science and Network Technology, 2011.

[5] Wajdi Bellil, Chokri Ben amar, Abdesselem Dakhli, . ."Wavelet Neural Networks for DNA Sequence Classification Using the Genetic Algorithms and the Least Trimmed Square," vol. 96, p. 10, 2016.