Classification of Cancer Model for Clinically Actionable Genetic Mutations Using Machine Learning

Anandh B.[1], Akash U.[1], Subhashini R.[2], Sethuraman R.[2], Saravanan M.[2]

[1]Student, Department of Information Technology, Sathyabama Institute of Science and Technology, Chennai-600119, India, [2]Faculty, School of Computing , Sathyabama Institute of Science and Technology, Chennai-600119, India

Abstract

Classification of Cancer Model Clinically Actionable Genetic Mutations Using Machine Learning Algorithms. Its task is to classify genes based on text evidence from clinical issues with good results. If a normal person has symptoms of cancer we can find it easily, but we have nine types of viruses in cancer in that which type of viruses has been attacked to the person cannot be easily predict by the doctors. So in hospitals there will be a clinical pathologist. Clinical pathologist has the data's of cancer attacked before and he will collect the gene sample and the person blood sample and predict which type of virus of cancer will attack to the person.

Keywords: *Genetic Mutations, Clinical Evidences, Clinical Pathologist, Natural Language Process.*

Introduction

Cancer is one of the most dangerous diseases. If cancer attacked to a person and not treated properly, it may lead to death. Health wise cancer may lead to blood vomit, loss of hair falls, body weakness etc. Researchers who are researching about the cancer has said that if a cancer is attacked a person would be dangerous for his life, which may lead to death . If the cancer not only it will affect him but also affect his future generation through his genes attacks a person, a person attacked by cancer also led to interrupt the normal routine work.

Our objective is to solve to these types of problem, so in this way person and person concerned to the victim can be alerted before it's late. To accomplish this, we need the data's which clinical pathologist will have victim's gene sample and blood samples to check which type of class(virus) of cancer will attack the person[1-2].

This project is a real time project based on "TF-IDF, TF-IDF/BOW, XGB, Meta Bagging, Text Reduction, and SVD" techniques. It will predict which types of class (virus) of cancer will attack the person so that doctors and the person concerned can help him to recover from the cancer[3-5].

Review of Literature: The machine-learning task is to addressing the crucial clinical method. In the previous research of cancer some other researchers have done for breast cancer and lung cancer by using some the machine learning algorithms like 5-fold cross validation and logistic regression etc. In our project what we have done is cancer classification with genetic mutations by using machine learning algorithms are Logical-Regression, K means ++, Naive Bayesian, K-fold cross validation, X-boost, LDA Cosine Similarity, K Nearest Neighborhood[6-12]. Naïve Bayesian is a supervised learning, in supervised learning input data's should be well labeled it collects the data and predict the before results and make it better accuracy[10]. The theory applied here is to compare the genetic mutations with the collection data samples from the patients and predict the possibility of getting cancer[2].

The importance of using more algorithms than the previous research is to find better accuracy than the existing one. When we use algorithms one by one it will take much more time to complete the process so we us the method called hyper tuning which will take less amount of time to complete the process[13-15].

System Overview: Classification of Cancer Model for Clinically Actionable Genetic Mutations is done for the theranostics. The architecture is described as shown in Fig 1. The program of this projected system are:

1. Included Sample Data (ISD)
2. Experimental Data Analytics (EDA)
3. Data Pre-processing:
    - Text Reduction
    - Dimension Compression
    - Component Extractions
4. Component Extraction against classified data
5. Genetic mutation classes 1-9 for classification of cancer
6. Result

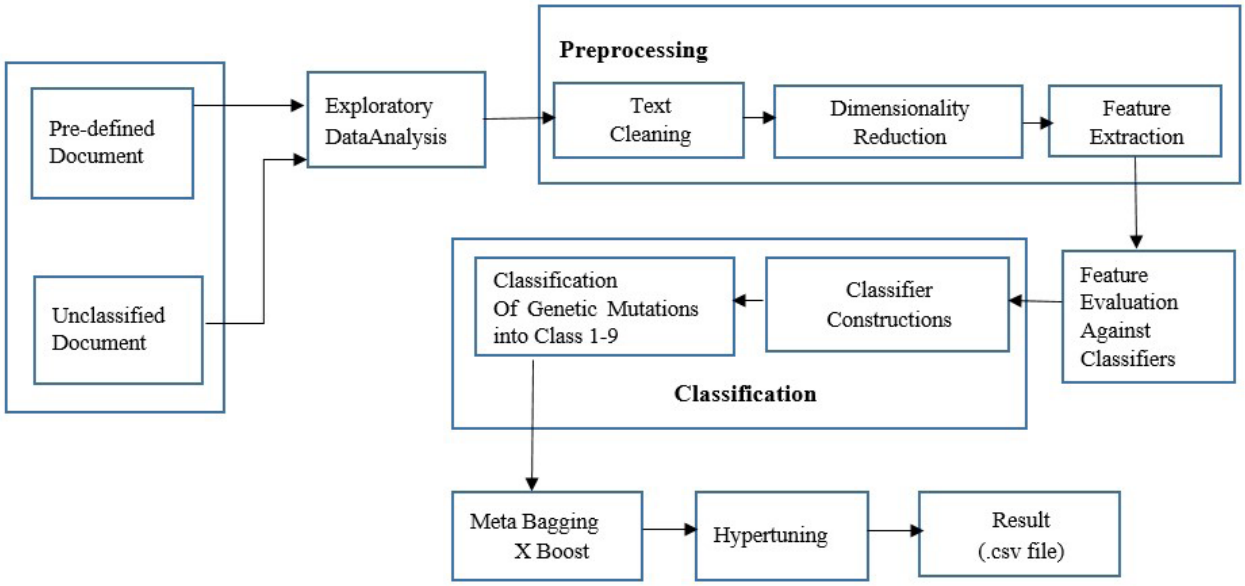

Fig. 1. Architecture of Described Technique

After analyzing the data. The clinical evidences must be extracted for the components[1]. The Neuro Linguistic Programming method are used here. The text data is represented using VSM. The features is being represented via formula given below:

$$\text{Assume t1} = wa, wb, \ldots\ldots, w_n \quad \ldots(1)$$

$$W_j = tf_j \times idf_j \quad \ldots(2)$$

$$Idf_j = log \quad \ldots(3)$$

$$(X/H) = P(X)/P(H) \quad \ldots(4)$$

Let n consist of total amount of unique contents in the text variant data, and w_j is the weight of the $-n^{th}$ term in wj . The recurrance of the term is given by tf_j, and the transposed document recurrence is given by idf_j . The N is the total amount of documents in the training_variant data set, and r_j is the amount of documents that contains the term j(text_variant). These extracted components are used for future operations.

The components are extracted using one hot encoding technique from genes and variation[2]. For test train the sample input data we are using the naïve Bayesian classification. The above naïve Bayesian formula consist the predictor and hypothesis[9].

The logical regression technique is used to validate the accurateness and logarithmic loss.

Included Sample Data: The datasets has been taken from kaggle. There 9 different type of categories of classified for the genetic mutations. The datasets square measure provided via 3 totally different files - coaching and take a look at. One (training_variants) provides training process for the data, (test_variants) provides the sample data texting and (training_text) will train the text formatted clinical evidences.

Experimental Data Analytics (EDA): Experimental Data Analytics (EDA) approach is employed for knowledge analytics. It used to analysis the experimental data of the classification method.

Data Pre-processing: The data pre-processing is used to check the input analysis data and gives result. The options of clinical text are extracted using TF-IDF technique. The options of genes and their variations are extracted using one hot cryptography technique.

Component extraction against classified Data: It is constructed for extracting the numerous classification algorithms already mentioned above. The options are being validated in form of logarithmic loss and accurateness.

Result

The result show that the naïve Bayesian and logistic regression gives the better result analysis than the other algorithms.

System Analysis: For implementing, using EDA the dataset is first analyzed. For the good results and relationship between the different features the experimental data analytics is performed. The solution of the Experimental Data Analytics are helpful in Data pre-processing. The TD-IDF is the technique show the output of the experimental data analytics is given below:

	Gene	Variation	Class	Text
ID				
0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...
1	CBL	W802*	2	Abstract Background Non-small cell lung canc...
2	CBL	Q249E	2	Abstract Background Non-small cell lung canc...
3	CBL	N454D	3	Recent evidence has demonstrated that acquired...
4	CBL	L399V	4	Oncogenic mutations in the monomeric Casitas B...

Fig. 2. Joined Train text data and training variants data

Fig. 2 shows the joined data that is training text, training variants and text_variants data is joined based on data analytics.

	ID	Gene	Variation	Class	Text
count	3321.000000	3321	3321	3321.000000	3321
unique	NaN	264	2996	NaN	1921
top	NaN	BRCA1	Truncating Mutations	NaN	The PTEN (phosphatase and tensin homolog) phos...
freq	NaN	264	93	NaN	53
mean	1660.000000	NaN	NaN	4.365854	NaN
std	958.834449	NaN	NaN	2.309781	NaN
min	0.000000	NaN	NaN	1.000000	NaN
25%	830.000000	NaN	NaN	2.000000	NaN
50%	1660.000000	NaN	NaN	4.000000	NaN
75%	2490.000000	NaN	NaN	7.000000	NaN
max	3320.000000	NaN	NaN	9.000000	NaN

Fig. 3. Exploratory stats on training data

Fig. 3 Shows Exploratory stats on the data such as par, structure of sample dataset, four quarters, Count, Highest, lowest values and standardized deviation. This info is necessary for experimental data analytics (EDA).

Table: Experimentation Results

Sr.No.	Component Extraction	Algorithm Classifier	Accurateness	Logarithmic Loss
1	TF-IDF	Logical Re- gression	86%	0.6812
2	TF-IDF/BOW	K Means++	83%	0.7288
3	XGB	Naïve Bayesian	77%	0.8757
4	XGB	K-Fold Cross Validation	85%	0.6438
5	Meta Bagging	X-Boost	64%	2.0852
6	Text Reduction	LDA Cosine Similarity	69%	1.0731
7	SVD	K Nearest Neighbourhood	62%	2.0395

Conclusion

Here by we have done our project classification of cancer model for clinically actionable genetic mutations with prior seven machine learning algorithms (Logical Re- gression, K Means++, Naïve Bayesian, K-Fold Cross Validation, X-boost, LDA Cosine Similarity, K Nearest Neighbourhood). Here we are using the meta bagging method to separate the data sample and ID-TDF showing the better result analysis. The result shows 86% accuracy. Further experimentation can be applied to increase the accuracy.

Ethical Clearance: No Clearance Required

Source of Funding: Self

Conflict of Interest: Nil

References

1. Waykole RN, Thakare AD. Intelligent Classification of Clinically Actionable Genetic Mutations Based on Clinical Evidences. In2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) 2018 Aug 16 (pp. 1-4). IEEE.
2. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015 Jun; 16(6):321-32.
3. Zhang Z, Lin H. Genomic profiling by machine learning. In2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW) 2011 Nov 12 (pp. 662-668). IEEE.
4. Yeung KY. Signature discovery for personalized medicine. In2013 IEEE International Conference on Intelligence and Security Informatics 2013 Jun 4 (pp. 333-338). IEEE.
5. Yoon J, Davtyan C, van der Schaar M. Discovery and clinical decision support for personalized healthcare. IEEE journal of biomedical and health informatics. 2016 Jun 1;21(4):1133-45.
6. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Jan 1;13:8-17.
7. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer informatics. 2006 Jan; 2:117693510600200030.
8. Vellido A, Biganzoli E, Lisboa PJ. Machine learning in cancer research: implications for personalised medicine. InESANN 2008 Apr (pp. 55-64).
9. Turki T, Wei Z, Wang JT. Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. IEEE Access. 2017 Apr 24: 5:7381-93.
10. Jagga Z, Gupta D. Machine learning for biomarker identification in cancer research–developments toward its clinical application. Personalized medicine. 2015 Aug:12(4):371-87.
11. Martín-Navarro A, Gaudioso-Simón A, Álvarez-Jarreta J, Montoya J, Mayordomo E, Ruiz-Pesini E. Machine learning classifier for identification of damaging missense mutations exclusive to human mitochondrial DNA-encoded polypeptides. BMC bioinformatics. 2017 Dec;18(1):158.

12. Vural S, Wang X, Guda C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. BMC systems biology. 2016 Aug;10(3):62.
13. Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, Feng DD. Deep Gene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC bioinformatics. 2016 Dec 1;17(17):476.
14. Subramanion R, Balasubramanian P, Noordeen S. Enforcement of Rough Fuzzy Clustering Based on Correlation Analysis. International Arab Journal of Information Technology (IAJIT). 2017 Jan 1;14(1).
15. Jeyanthi P, Kumar VJ. Image classification by K-means clustering. Advances in Computational Sciences and Technology. 2010;3(1):1-8.