EXPLORING COLORECTAL CANCER GENES THROUGH DATA MINING TECHNIQUES

R. DEVENTHIRAN1, S. SENTHILKUMAR[2]

[1] PG Scholar, Department of MCA, Arulmigu Meenakshi Amman College of Engineering, AnnaUniversity, Vadamavandal (near Kanchipuram), India.
[2] Assistant Professor, Department of MCA, Arulmigu Meenakshi Amman College of Engineering, Anna University, Vadamavandal (near Kanchipuram), India.

---***---

ABSTRACT: - Data mining is used in various medical applications like tumor classification, protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical model of protein-protein interaction etc. Drug events in prediction of medical test effectiveness can be done based on genomics and proteomics through data mining approaches. Cancer detection is one of the hot research topics in the bioinformatics. Classification of colon cancer dataset using WEKA 3.6, in which Logistics, IBK, KSTAR, NNGE, AD Tree, Random Forest Algorithms show 100 % correctly classified instances, followed by Navie bayes and PART with 97.22 %, Simple Cart and Zero R has shown the least with 50 % of correctly classified instances. Mean absolute error and Root mean squared error are shown low for Logistics, KSTAR and NNGE.

1 INTRODUCTION:

Accumulation of data all over the globe is growing at a fast rate especially in the health field. One of the main interesting fields is cancer in general. In UAE, for example, there are 35% of residences affected by the colon cancer. Generally, the Colon Cancer often occurs with the Rectal Cancer and called Colorectal Cancer (CRC). All people over 68 are more disposed to have the Colorectal Cancer, and the percent of Colon and Rectum Cancer deaths is highest among people aged 75-84.

Colon and Rectum Cancer represent 8% of all new Cancer cases where there will be an estimated 50,000 new cases of Colon Cancer in the United States in 2017. In UAE, the Colorectal Cancer is considered as the second largest cause of death from cancer after breast cancer.

1.1 EXISTING SYSTEM:

Various types of medical cancer research are in top priority around the globe. Data mining can help clear out unseen cancer treatment for different kinds of populations, ethnicities, genders, economic and social influences that may affect assessments worked on Extreme Learning Machine (ELM) for classification.

EXISTING TECHNIQUES:

Zero R algorithm.

TECHNIQUES DEFINITION:

Zero Rule (Zero R) algorithm is the simplest classification method which relies on the target and ignores all predictors. It can be used to determine a baseline performance as a benchmark for other classification methods.

DRAWBACKS:

There is nothing to be said about the predictors contribution to the model because Zero R does not use any of them.

1.2 PROPOSED SYSTEM:

Bayesian networks are a powerful probabilistic representation, and their predictor or attributes. In particular, the naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors.

A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets.

PROPOSED TECHNIQUE:

Naïve Bayes Multinomial algorithm.

TECHNIQUE DEFINITION:

This method goes by the name of Naïve Bayes, because it's based on Bayes' rule and "naïvely" assumes independence; it is only valid to multiply probabilities when the events are independent.

ADVANTAGES:

Make the rule assign that class to this value of the predictors

Calculate the total error of the rules of each predictor.

2. SYSTEMS SPECIFICATION:

2.1 HARDWARE REQUIREMENTS:

PROCESSOR	:	DUAL CORE
RAM	:	4GB
HARD DISK	:	250 GB

2.2 SOFTWARE REQUIREMENTS:-

FRONT END: J2EE (JSP, SERVLET), STRUCTS

BACK END: ARFF DATASETS

OPERATING SYSTEM: WINDOWS7

IDE : ECLIPSE.

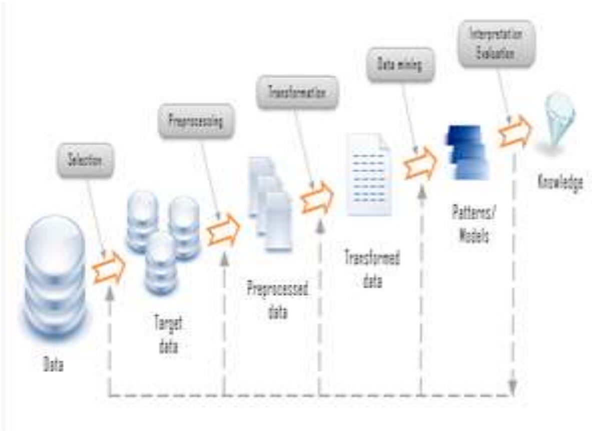
3. PROJECT DESCRIPTION:

MODULES:

- **Data Preprocessing Module**
- **Data Predictor Module**
- **Data Analysis Module**
- **Data Reporting Module**

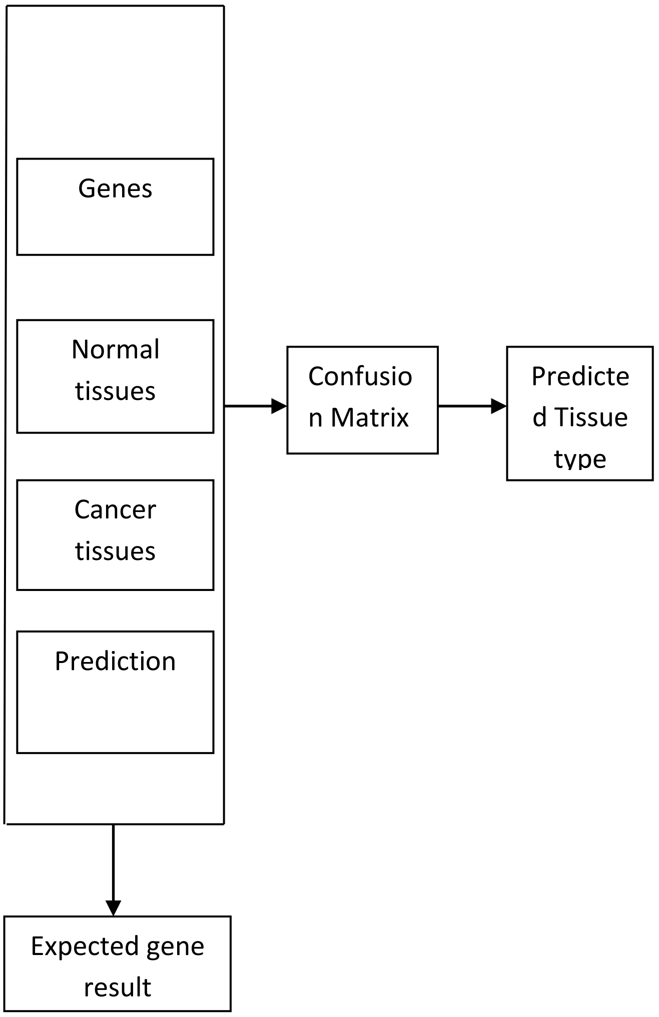
Data Preprocessing Module:

We are first collecting all the data sets from available sources ,and based on our project we are preprocessed the data like separating the cancer tissues ,normal tissues ,cancer values ,normal values and their names based on the signed values .



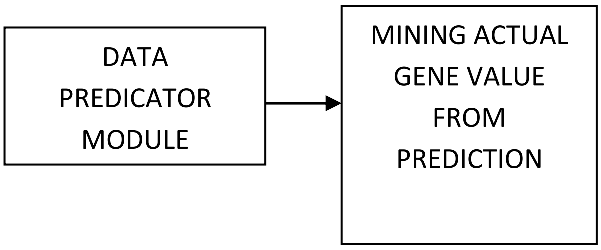
Data Predictor Module:

In data prediction module, all the data sets like genes, tissues classification based on signed values, if it is positive value, the prediction will be normal tissue if it is negative range that will be cancer, based on that the tissue analysis happen and the expected genes result and after that we got the confusion matrix like the overall analysis of the tissue, genes, elements etc..



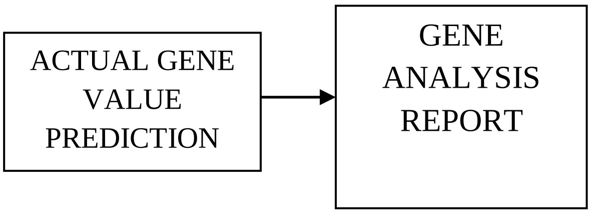
Data Analysis Module:

In the data analysis module all the datasets are analyzed based on 4 Algorithm's and it will give the analyzed output based on analyze techniques.



Data Reporting Module:

In the data reporting module all these prediction and analysis taken into account and our colon detector tool summarize all the output based on confusion matrix result like which algorithm gives the best accuracy and medicines for cancer tissues.



4. SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE:

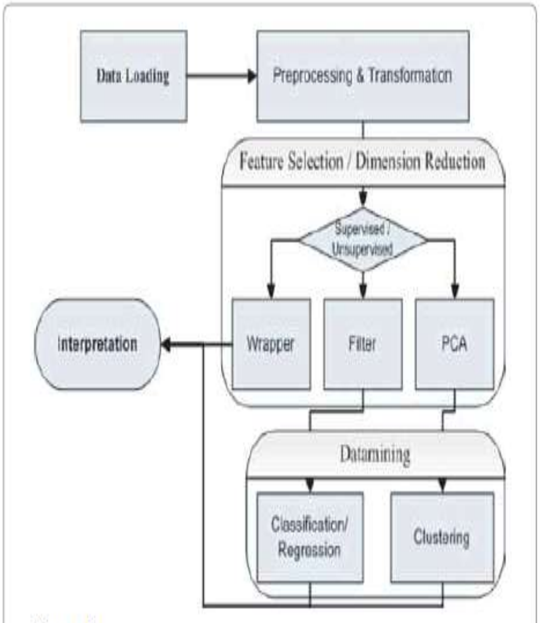

Data Loading
Preprocessing & Transformation
Feature Selection / Dimension Reduction
Supervised / Unsupervised
Interpretation
Wrapper
Filter
PCA
Datamining
Classification/ Regression
Clustering

Figure 1:

5. SCOPE OF THE PROJECT:

The project works for the Cancer detection from gene expression data. We presented a deep learning technique to detect cancer and to identify critical genes for the diagnosis of cancer. Their results showed that the highly interactive genes could be useful as cancer biomarkers for the detection of cancer.

By using Zero R algorithm, Naive Bayes algorithm, Ripper algorithm, and J48 algorithm. We are detecting the cancer genes and normal By using Zero R algorithm, Naive Bayes algorithm, We are detecting the cancer genes and normal genes from the output of each algorithm, the accuracy level gets vary.

Colorectal Cancer qualities in a Nutshell:

By and large, Cancer is a crazy cell development illness. Colon Cancer and Rectal Cancer frequently happen together and are called Colorectal Cancer (CRC). The Colon is the most reduced piece of the stomach related framework. Roughly, individuals more than 68 are increasingly inclined to have the Colorectal Cancer, and the percent of Colon and Rectum Cancer passing is most noteworthy among individuals matured 75-84. Colon and Rectum Cancer speak to 8% of all new Cancer cases.

DATA MINING TECHNIQUES:

Data mining techniques process the information by using specific methods to discover this data and extract vital points from the big data set. Including methods such as generalization, characterization, classification, clustering, association, evolution, pattern matching and data visualization, and not limited to meta-rule guided Mining.

(i) Zero R Algorithim

Zero Rule (Zero R) algorithm is the simplest classification method which relies on the target and ignores all predictors. It can be used to determine a baseline performance as a benchmark for other classification methods.

(ii) Naïve Bayes Multinomial algorithm

This method goes by the name of Naïve Bayes, because it's based on Bayes' rule and "naïvely" assumes independence; it is only valid to multiply probabilities when the events are independent. Bayesian networks are a powerful probabilistic representation, and their predictor or attributes.

(iii)Repeated Incremental Pruning to Produce Error Reduction (RIPPER)

RIPPER is a well-known algorithm that utilizes separate-and-conquer strategy to distinct between positive and negative instances. It is considered as the master of the inductive rule learning that has the ability to deal with over-fitting problem.

(iv)*J48*

J48 is a prediction algorithm that can be used successfully in many demanding problem domains and also to create Decision Trees. It is like a dataset including two lists: a list of predictors (independent variables) and a list of targets (dependent variables) that would allow you to predict the target variable of a new dataset record. Decision tree J48 implementation is based on the C4.5 algorithm ID3 (Iterative Dichotomies 3) developed by the WEKA project team.

PROJECT IMPLEMENTATION:

(i) Data Collection
(ii) Classification
(iii) Clustering
(iv) Prediction

The Colon Cancer is a forceful and understood illness that influences individuals all around the globe. To examine this sort of malignancy, this examination endeavors to investigate the relations between the qualities in charge of colon disease by means of Data Mining arrangement procedures. WEKA has been utilized as a Data mining device to investigate a known dataset for Colorectal Cancer qualities from writing. Distinctive procedures were utilized and thought about. The outcomes were promising and the investigation step can be talked about with a specialist soon.

Accumulation of data all over the globe is growing at a fast rate especially in the health field. One of the main interesting fields is cancer in general. The Colon Cancer, in particular, is an aggressive and well known disease that affects people all around the world. In UAE, for example, there are 35% of residences affected by the colon cancer. Generally, the Colon Cancer often occurs with the Rectal Cancer and called Colorectal Cancer (CRC).

(i)Data Collection:

In data collection phase different cancer genes collected as in the form of data sets. The information includes the normal tissues, cancer tissues etc.

(ii)Classification:

Classification is the most common data mining technique, which employs a set of pre classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network - based classification algorithms. The data classification process involves learning and classification.

(iii)Clustering:

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes.

(iv)Prediction:

Predicting the cancer genes based upon four important algorithms based on previously cancer genes and normal genes behavior and activities. In this module can predict which tissues are affected by cancer genes and normal genes.

6. CONCLUSIONS:

This simple study aims at exploring the Colorectal Cancer disease using the available data in literature. The used dataset contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues with 40 tumor tissues and 22 normal ones. The best algorithm in terms of classification performance was Naïve Bayes Multinomial with very close results.

The next step was to filter the 2000 genes into 26 using a well-known feature selection algorithm and apply a rule-based classification algorithm on the new dataset. The generated rules were very clear and adding a new insight to the database in general. It seems that the gene “M26383” is a very important gene that can determine a normal tissue if it is below 56.9. As a general finding, these rules can be further studied by an expert in the field.

7. FUTURE ENHANCEMENT:

Designing efficient online mechanisms for a bi-directional market is significantly more challenging. We consider server energy cost minimization in social welfare maximization, and reveal an important property, sub modularity, of the objective function in the resulting significantly more challenging offline problem.

8. REFERENCES:

[1] (2017, February) Centers for Disease Control and Prevention.[Online]. HYPERLINK "https://www.cdc.gov/" https://www.cdc.gov/

[2] National Cancer Institute. [Online]. HYPERLINK "https://seer.cancer.gov/statfacts/html/colorect.html" https://seer.cancer.gov/statfacts/html/colorect.html

[3] (2017, February) The Surveillance, Epidemiology, and End Results.[Online]. HYPERLINK "https://seer.cancer.gov/" https://seer.cancer.gov/

[4] Centers for Disease Control and Prevention (CDC).[Online]. HYPERLINK "https://www.cdc.gov/cancer/colorectal/index.htm" https://www.cdc.gov/cancer/colorectal/index.htm

[5] U. Alon et al., "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proceedings of the National Academy of Sciences, vol. 96, no. 12, pp. 6745-6750, 1999.

[6] Jesse Samuel Moore and Tess Hannah Aulet, "Colorectal Cancer Screening," Surgical Clinics of North Ameria, vol. 97, no. 3, pp. 487-502, 2017.

[7]"https://seer.cancer.gov/statfacts/html/colorect.html"
[8]https://seer.cancer.gov/statfacts/html/colorect.html.