A Feature Selection Framework for DNA Methylation Analysis in Predicting Bladder Cancer

Sreeshma P.S[1]

[1]*Department of Computer Sci. & Engg, Thejus Engineering College, Vellarakkad, Thrissur, Kerala, India*

Abstract – *DNA Methylation is a procedure by which methyl groups are included to the DNA molecule. DNA methylation characterize an important role for the origination and the growth of human cancer. So for the early prediction of bladder cancer, DNA methylation can be used. This paper presents two different feature selection methods namely, Correlation based Feature Weighting(CFW), and Differential Mean Feature Selection(DMFS). CFW is a process by which each feature is assigned a weight according to their correlation. And this weight is compared with the threshold value, if the weight is greater than the threshold that feature is selected. Otherwise the feature will be rejected. DMFS uses the Differential mean values of both normal and cancer samples.*

***Key Words*: DNA Methylation, Feature Selection, Correlation based Feature Weighting, Differential Mean Feature Selection.**

1. INTRODUCTION

Throughout our lives, healthy cells in our bodies divide and replace themselves is a controlled fashion. Cancer may be a general term for large cluster of diseases, whose causes, characteristics and prevalence vary greatly. It's refers to an abnormal growth of cell tissue. Genetic and epigenetic changes contribute to the event of cancer. DNA methylation is a procedure by that methyl teams superimposed to DNA molecule. Methylation will modify the activity of a DNA section while not changing the sequence. The inflated methylation is named as hypermethylation and also the massive loss of DNA methylation stated as hypomethylation. complete hypomethylation and promoter region specific hypermethylation are 2 forms of typical DNA methylation found in human cancer. DNA methylation plays a awfully vital role for the origination and growth of human cancer. So, it are often used as a biological marker for the primary prediction of cancers [1].

The main problems with DNA methylation dataset is that huge number of features in each sample and less number of the available samples. One possible solution is to use feature selection methods to obtain relevant feature set and to eliminate redundant and irrelevant features. So , this paper introduces two methods for feature selection namely, Correlation based Feature Extraction (CFW) and Differential Mean Feature Selection (DMFS). Sample size goes through two reduction steps, correlation based feature weighting and redundant feature elimination and unsupervised feature selection via vertical analysis of the feature.

The main contributions of this paper are:

- Building a framework for DNA methylation classification.
- Uses correlation based feature weighting method for the feature selection process.
- Differential Mean feature Selection by utilizing vertical analysis of features across the sample set.

2. LITERATURE REVIEW

[3] A datamining approach for predicting a disease using the most relevant genes associated with it. A few existing feature selection methods have been applied independently and their results combined to form fused feature set and then applying genetic algorithm which generate optimal feature set. They have been applied to a classifier for identifying the disease.

[4] A framework based on both feature selection methods and feature extraction methods to predict cancer. This framework is made foe detecting cancer based on the methylated DNA probes.

[5] A method to select the small subset of informative gene relevant to the classification. This method is based on a binary migration model and binary mutation model.

[6] Supervised learning technique that have been employed to classify cancers. A two step feature selection method is done. First one is an attribute estimation method(eg. ReliefF) and the second one is the genetic algorithm.

[7] Tried to find efficient feature selection methods to select a small number of informative genes using mutual information and rough sets .

[8] Sparse Compact Learning Incremental Machine (SCLIM) for cancer classification on microarray gene expression data that is robust against diverse noises and outliers.

3. METHODOLOGY

With the advancement of technology, real time monitoring system to the early prediction of cancers have become common. This paper proposes a framework for the early prediction of bladder cancer. For the early prediction of cancer DNA methylation analysis is used. One of the important role is for classifying normal samples from cancer ones. For that combining different feature selection methods for classification. The proposed framework consists of three modules ; Data pre-processing means that the data in the available format is loading and then pre-processing the data in the required format, Feature selection module contains 2 methods correlation based feature weighting (CFW) and Differential Mean Feature Selection (DMFS) by utilizing vertical analysis of the methods for each sample, finally classification method is applied to the selected features.

The fig 1 illustrates the workflow of the method. Initially the data is pre-processed. For that, the data in the available format is changes to the required format. The pre-processed data consists of samples in rows and features in columns. After pre-processing the data two feature selection methods are applying CFW and DMFS. Then finally Naïve Bayes classification is done.

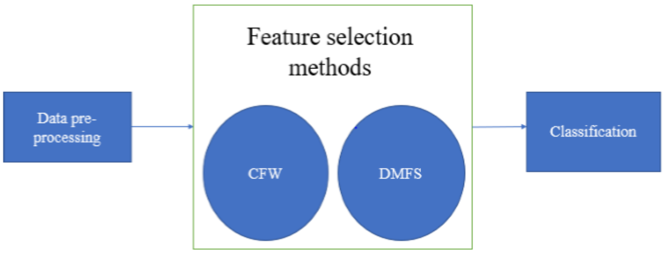

Fig -1: Workflow of the method

3.1 Data Pre-Processing

Data pre-processing is a datamining technique that involves transforming raw data into an understandable format. The main aim of this module is to load the available data. The dataset contains 24 samples, where 6 normal samples and 18 cancer samples. In each of the sample there is 27,578 features.

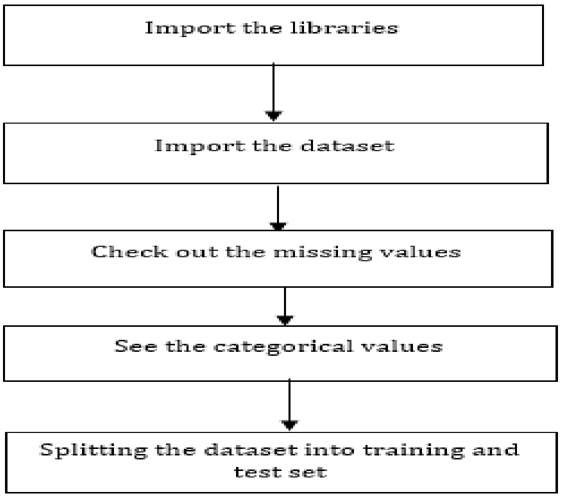
Import the libraries
Import the dataset
Check out the missing values
See the categorical values
Splitting the dataset into training and test set

Fig -2 : Steps in data pre-processing

3.2 Feature Selection

Correlation based Feature Weighting(CFW) : The weight for a feature is proportional to the difference between the feature - class correlation (mutual relevance) and the average feature - feature intercorrelation (average mutual redundancy) [2]. In this method each feature is assigned a different weight according to their correlation.

The equation for feature - class correlation and feature – feature intercorrelation can be respectively defined as [2],

$$I(A_i;C) = \sum a_i \sum c \, P(a_i,c) \log (P(a_i,c) / P(a_i)P(c)) \quad (1)$$

$$I(A_i;A_j) = \sum a_i \sum a_j \, P(a_i,a_j) \log (P(ai,aj) / P(a_i)P(a_j)) \quad (2)$$

where C is the class variable, A_i and A_j are two different feature variables, c, a_i and a_j represent the values that they take, respectively.

$$D_i = NI(A_i;C) - 1/m-1 \sum j=1 \Lambda j!=1 \, NI(A_i ; A_j) \quad (3)$$

where $NI(A_i;C)$ is the normalized $I(A_i;C)$ representing mutual relevance and $NI(A_i;A_j)$ is the normalized $I(A_i;A_j)$ representing mutual redundancy.

$$NI (A_i;C) = I (A_i;C) / (1/m \sum i=1 \text{ to } m \, I (A_i;C)) \quad (4)$$

$$NI(A_i;A_j)=I(A_i;A_j)/(1/m(m-1) \sum i=1 \text{ to } m \sum j=1 \Lambda j!=1 I(A_i;A_j) \quad (5)$$

The final weight of the feature A_i is

$$W_i = 1 / 1+e^{-D_i} \quad (6)$$

This weight is compared with the threshold value. If the weight is greater than the threshold that feature will be selected otherwise rejected.

Differential Mean Feature Selection (DMFS) : Analyzing each features across all samples vertically. Finding the mean values of both normal and cancer samples separately. And then find the absolute difference between mean values of both normal and cancer samples (weight vector), selecting the features corresponding to the weight vector, and comparing with the threshold, If the weight vector is greater than the threshold that feature will be selected.

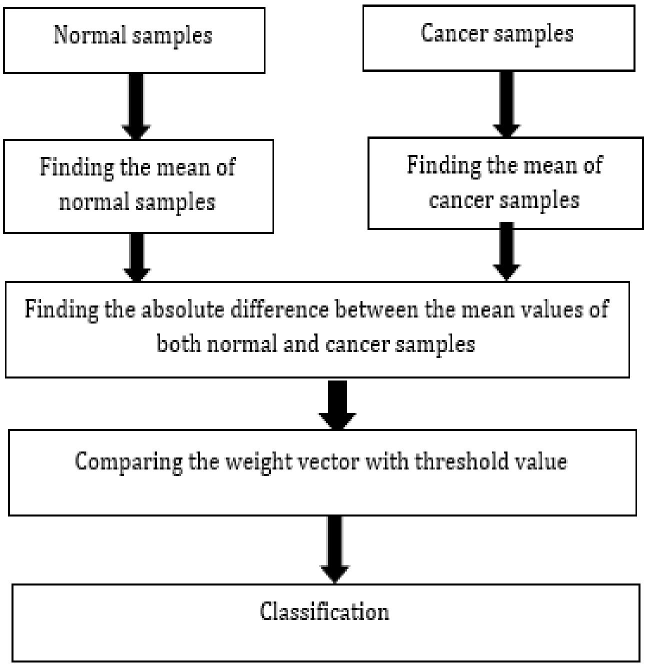
Normal samples
Cancer samples
Finding the mean of normal samples
Finding the mean of cancer samples
Finding the absolute difference between the mean values of both normal and cancer samples
Comparing the weight vector with threshold value
Classification

Fig -3 : DMFS approach

3.3 Classification

There are many classification techniques and each of them have its own advantages and disadvantages. The Naïve Bayes classifier is the direct and simplest classifier that utilizes Bayes theorem for conditional probabilities of random variables given known observations to build the classifiers [1]. In this classifier all features are assumed independent from each other and it calculates independently the probability of each feature for a particular class label [9],[10].This classifier is simple and computationally fast to reach a decision.

3. CONCLUSION

The main aim of the paper is to provide efficient feature selection methods for the early prediction of bladder cancer. So, building a framework for DNA methylation classification. Basically it is a classification problem which classifies bladder cancer dataset into normal samples and cancer samples. Mainly two feature selection methods are used namely, Correlation based Feature Weighting (CFW) and Differential mean Feature Selection (DMFS). CFW assigns weights for each of the features according to their importance. And then comparing the weight vector with the threshold value and if the weight is greater than threshold, then that feature is selected. DMFS uses differential mean between the mean of normal and cancer samples. The number of features can be reduced by applying the feature selection methods.

REFERENCES

[1] Abdul Majid f. Al-Junaid and Talal s. Qaid," Vertical and Horizontal DNA Differential Methylation Analysis for Predicting Breast Cancer," IEEE Access.,vol. 6 , pp. 53533,Aug 2018.

[2] Liangxiao Jiang, Lungan Zhang, Chaoqun Li, Jia Wu," A Correlation-based Feature Weighting Filter for Naive Bayes", IEEE Transactions on Knowledge and Data Engineering., pp. 2836440, 2018.

[3] Mohammed Siyad B, Visakh R," Disease Prediction using Optimal Feature Selection from Epigenetic Data," IEEE conf. on Innovations in Power and Advanced Computing Technologies, 2017.

[4] A. A. Raweh, M. Nassef. And A. Badr," Feature selection and extraction framework for DNA methylation in cancer, Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 7, pp. 3036,2017.

[5] X. Li and M. Yin, "Multi-objective binary biogeography based optimization for feature selection using gene expression data," IEEE Trans. Nanobiosci., vol. 12, no. 4,pp. 343352,Dec. 2013.

[6] H. Hijaziand C. Chan ," A classification framework applied to cancer gene expression profiles", J. Healthcare Eng., vol. 4, no. 2, pp. 255–283, 2013.

[7] W. Zhou, C. Zhou, G. Liu, and H. Zhu, "Feature selection for microarray data analysis using mutual information and rough set theory", in Proc. IFIP Int. Fed. Inf. Process., Artif. Intell. Appl. Innov., vol. 204, I. Maglogiannis, K. Karpouzis, and M. Bramer, Eds. Boston, MA, USA: Springer, 2006, pp. 492–499.

[8] M. Nayyeri and H. S. Noghabi, "Cancer classification by correntropy based sparse compact incremental learning machine," Gene Rep., vol. 3, pp. 31–38, Jun. 2016.