ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 11, Issue, 01, pp.515-517, January, 2019

DOI: https://doi.org/10.24941/ijcr.34100.01.2019

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

CANCER PREDICTION USING HFSE APPROACH BASED ON DNA METHYLATION OVER MACHINE LEARNING

[1]Nasiya, PM., [2]Priya, K.V. [3]Dr. Rajeswari, M. and [2]Krishnadas, J.

[1]M.Tech Scholar, Department of CSE Sahrdaya College of Engineering and Technology, Kodakara, Thrissur
[2]Assistant Professor, Department of CSE Sahrdaya College of Engineering and Technology, Kodakara, Thrissur
[3]Associate Professor, Department of CSE Sahrdaya College of Engineering and Technology, Kodakara, Thrissur

ARTICLE INFO

Article History:
Received 28th October, 2018
Received in revised form
20th November, 2018
Accepted 09th December, 2018
Published online 31st January, 2019

Key Words:

Cancer prediction,
DNA methylation,
Feature selection,
Feature Extraction

ABSTRACT

Breast Cancer serve as one of the diseases that make a high number of deaths every year. It is the common type of all cancers and the main cause of women's deaths worldwide. Due to the vital role of the aberrant DNA methylation during the disease development such as cancer, prediction mechanism had become essential in the recent years for early detection and diagnosis. The high-dimensionality and noisiness of the DNA methylation data may lead to the reduction of the prediction accuracy. Thus, it becomes more important in a wide range to employ robust computational tools such as feature selection and extraction methods to extract the informative features amongst thousands of them, and hence improving cancer prediction. This paper aims at predicting cancer with a hybridized feature selection and feature extraction (HFSE) techniques. The suggested approach shows a filter feature selection method called (F-score) to overcome the high-dimensionality problem of the DNA methylation data, and proposes an extraction model which employs the peaks of the mean methylation density in order to exact cancer classification and reduce training time. To evaluate the reliability of our approach, machine learning algorithms such as The naïve base and support vector machine, knn algorithms are introduced to predict cancer. The results show that, the classification accuracy improves in all cases and it also proves the reliability indirectly.

Citation: Nasiya, PM., Priya, K.V. Dr. Rajeswari, M. and Krishnadas, J.**, 2019.** "Cancer prediction using hfse approach based on dna methylation over machine learning", *International Journal of Current Research*, 11, (01), 515-517.

INTRODUCTION

Cancer is a leading cause of death world wide, it begins when some cells in a part of the body start to grow out of control. Despite the presence of more than one type of cancer that differ in the way of growing cells and spreading, the development of all these types is driven by "genetic changes of the DNA sequence (Terry, 2016). Recent research increases resources that the epigenetic modifications play a critical role in human cancer. These modifications are the changes in a cellular phenotype that are independent of alterations in the DNA sequence (Baur and S.Bozdag, 2016 and Wu, 2017). Many studies of epigenetic aberrations in tumors prove that the biology of DNA methylation is the most potential epigenetic marker for cancer (Pouliot, 2015). Actually, DNA methylation acts as a gene-alteration mechanism to turn off specific genes due to its significant effects on gene expressions and the structure of the nucleus of the cell (Hira, 2015). Chemically, DNA methylation is a relatively stable chemical modification resulting from the addition of a methyl (CH3) group at the carbon 5 position of the guanine or cytosine nucleotides in the sequence of 5'-CG-3' (CpG di nucleotide) by DNA methyl transferase (DNMT) enzymes (Liu, 2017). There are two types of cancer-associated with DNA methylation, based on the methylation level called hypo methylation and Hypermethylation "the methylation exceeds normal methylation level" of tumor suppressor gene in the gene expression. Hypomethylation "the methylation beneath normal methylation level" has been observed in solid tumors (Singh, 2015). Ideally, dimensionality reduction method should eliminate these irrelevant probes while at the same time retain all the highly discriminative probes. Hybridized feature selection and extraction techniques in cancer predication becomes essential. In this paper, we propose a framework based on feature selection and extraction methods, to eliminate irrelevant information and improve cancer classification accuracy based on DNA methylation data. First, a feature selection based on statistical variation and standard deviation is utilized for identifying the small set of discriminative methylated DNA probes, then, the average methylation density of three regions (hypermethylation, midmethylation and hypomethylation) is calculated as new extracted features to predict cancer. Section II elaborates on previous work, Section III presents the proposed framework, Section IV discusses our experimental results and the last Section V contains conclusion.

***Corresponding author: Nasiya,**
M.Tech Scholar, Department of CSE Sahrdaya College of Engineering and Technology, Kodakara, Thrissur

Related Works

To increase the accuracy and increasing tumor feature data and information, a number of researchers have turned to feature selection and extraction techniques for predicting cancer. Feature selection is one of the important steps in classification modeling of cancer based on DNA methylation (Liu, 2017), it is used for eliminating unnecessary information to reduce the high dimensionality of the data. Whereas feature extraction also known as data transformation, is the process of transforming the feature data into a quantified data type instead of recognizing new patterns to represent the data. In the past, many feature selection and extraction methods have been proposed, resulting in improvements of classification. Li *et al.* (Ren, 2013), proposed a gene extraction method, using two standard feature extraction methods, are the T-test method and kernel partial least squares (KPLS) in tandem. Zheng *et al.* (Assenov, 2014), proposed a hybrid of K-means and support vector machine (K-SVM) algorithms to predict breast cancer. Kopriva *et al.* (Jaganathan, 2012), proposed a general feature extraction method for cancer detection based on the linear transformation constructed by tensor decomposition. A novel method using wavelet analysis, Bayes classifier and genetic algorithm proposed by Liu *et al.* (Sheather, 2004), was applied to detect the biomarkers of survival in colorectal cancer patients. Fontes *et al.* (Schneider, 2011), used feature extraction techniques such as *p-value rank, F-score* and *wrapper approaches* in order to identify which probes Presented higher significance in breast cancer prediction. D.L. Tong (Nguyen, 2013), proposed an innovative hybridized model based on genetic algorithms (GAs) and artificial neural networks (ANNs), to extract the highly differentially expressed genes for specific cancer pathology. Anuradha *et al. (*Kaur and S.Kalra, 2016) provided a comparative study to identify the best feature extraction technique to classify Oral cancers. Zhuang *et al.* (Jaganathan, 2012), gave another good comparison study of feature selection and classification methods in DNA using the Illumina Infinium platform. Cai *et al.* (Assenov, 2014) applied Ensemble-based feature extraction methods to capture the unbiased, informative as well as compact molecular signatures followed by SVM trained with Incremental Feature Selection (IFS) strategy to predict subtypes of lung cancer. A novel feature selection and classification system proposed by Sebastian *et al*. (Nguyen, 2013), used for data merged from different molecular biomedical techniques demonstrated that the feature selection step is crucial in high dimension data classification problems. Furthermore, Baur *et al.* (2016), make a feature selection algorithm based on sequential forward selection to compute gene centric DNA methylation using probe level DNA methylation data. Valavanis *et al*. (2015) captured the semantics information included in the Gene Ontology (GO) tree by graph-theoretic methodology in order to select cancer epigenetic biomarkers.

PROPOSED METHODS

Feature Selection Methods: Feature selection methods in cancer classification are aimed to identifying the minimal-sized subset of markers that are necessary to accurate prediction. To achieve this goal, Propose two novel feature selection methods. The first one uses statistical variation in terms of standard deviation in order to select the most informative probes which separate normal tissue from cancer. This measures the differences of probe methylation in all samples compared with the dispersion of this probe methylation in each Normal and Cancer class separately. The proposed feature selection according to discriminative value (DV) for each probe (X) based on DNA methylation as an input is defined as:

$$\mathbf{DV(X)} = \frac{\frac{\sum(x-\bar{x})^2}{n-1}}{\sqrt{\frac{\sum(x^+-\bar{x}^+)2}{n^+-1}}+\sqrt{\frac{\sum(x^--\bar{x}^-)2}{n^--1}}}$$

where n+,n- are the number of positive and negative Instances and x^{-},$x^{-(+)}$,and $x^{-(-)}$ are the average of the i^{th} feature of the whole, positive, and negative datasets respectively.

Feature Extraction Method: The most discriminative probes (10,000 probes) are selected using the feature selection *DV(X).* Then these are extracted using feature extraction methods. Feature extraction is the process which involves for clarifying and detecting the methylation patterns or behavior in the selected probes. As a first step, then use kernel density estimator method [2]; which infers population probability density function of the selected probes; as a feature extraction method, in order to extract 512 features for each sample. The kernel density estimate of at the point is given by $\hat{f}h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-xi}{h}\right)$, where denotes to so-called Gaussian kernel function that integrates to one and has mean zero. It defined as: $K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$, And h denotes to a smoothing parameter >0 called the bandwidth. The optimal bandwidth that gives better results can be obtained by: $h_{opt} = \frac{0.9\times\partial}{\sqrt[5]{N}}$ where, ∂=min (∂,IQR/1.34) and is the interquartile range that measures the difference between the 75th percentile (Q_3) and the 25th percentile (Q_1):IQR=Q_3-Q_1

Classification

To evaluate the ability of the proposed method for cancer classification based on methylated probes, the following classifiers: Naïve Bayes, SVM and KNN were used.250 samples from breast tissue were used as training data and 348 samples were used as testing data.fig 3.1. Different approaches were used to study classifier's accuracy in cancer prediction, where the first experiment used the methylation density of whole probes (485,577 probes), the second one used methylation density of most discriminative probes by DV(X) (10,000 probes) and the Last experiment used three features only average methylation density of three regions :Hypo, Mid, Hyper methylation. The next section shows the testing accuracy, F-Measure, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of each machine learning technique. According to this experiments, the reader can observe the accuracy of classifier in cancer prediction using only the extracted three features.

RESULTS AND DISCUSSION

In order to compare the behaviours of SVM, NB, and k-NN, we conducted an experiment that focuses on assessing both the effectiveness, and the efficiency of the algorithms. More precisely, the research questions posed for the experiment are: Which algorithm exploits better effectiveness? Which algorithm is more efficient? Which algorithm provides a higher accuracy?

Experiment Environment: All experiments on the classifiers in this paper were conducted using libraries from Weka machine learning environment. WEKA contains a collection of machine learning algorithms for data preprocessing, classification, regression, clustering and association rules. Machine learning techniques implemented in WEKA are applied to a variety of real world problems. The program offers a well-defined framework for experimenters and developers to build and evaluate their models.

Breast cancer dataset: The Wisconsin Breast Cancer (original) datasets from the UCI Machine Learning Repository is used in this study. It has 699 instances (Benign: 458 Malignant: 241), 2 classes (65.5% malignant and 34.5% benign), and 11 integer-valued attributes.

Experimental results: During the last decades, accumulating evidence confirmed the significant of the classification process in the biology field; due to the capability of its algorithms that helps researchers and scientists to expand their knowledge and improves tumor diagnosis among different tumor types. Here we present the results of some classification algorithms of weka 3.8 that have been applied on the dataset such as NaïveBase, KNN, and SVMwith 10-fold cross-validation procedures. These classification techniques utilized the DNA methylation degree to distinguish between normal and cancer samples. The following table 1 demonstrates the prediction accuracy and f-measure of Naïve Base,KNN, and SVM in promoters region respectively.

Table 1. Accuracy of classification algorithms

CLASSIFIERS	ACCURACY
Naïve Base	96.32
K-Nearest Neighbour	97.8
Support Vector Machine	98.54

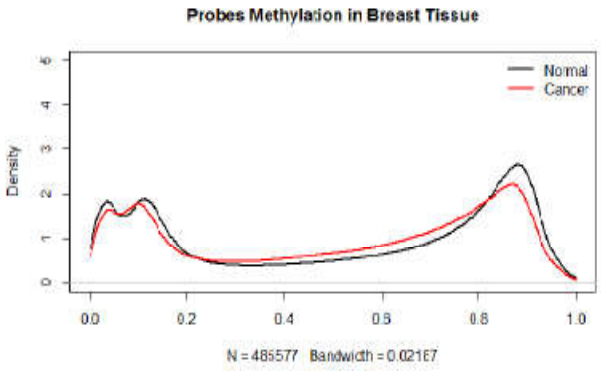

Fig. 1. Whole probe methylation in Breast tissue

Conclusion

To analyze medical data, various data mining and machine learning methods are available. An important challenge in data mining and machine learning areas is to build accurate and computationally efficient classifiers for Medical applications.

In this study, we employed four main algorithms: SVM, NB, and k-NN on the Wisconsin Breast Cancer (original) datasets. We tried to compare efficiency and effectiveness of those algorithms in terms of accuracy, precision, sensitivity and specificity to find the best classification accuracy of SVM reaches and accuracy of 98.54% and outperforms, therefore, all other algorithms. In conclusion, SVM has proven its efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of precision and low error rate.

REFERENCES

Assenov, Y.F.Müller, P.Lutsik, J.Walter, T.Lengauer, and C.Bock,"Comprehensive analysis of DNA methylation data with Rn Beads," *Nature Methods*, vol. 11, no. 11, p. 1138_1140, 2014.

Baur, B. and Bozdag, S. 2016. ``A feature selection algorithm to compute gene centric methylation from probe level methylation data," *PLoS One*, vol. 11, no. 2, p. e0148977, 2016.

Hira, Z.M. and DF.Gillies,"A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinform.*, vol. 2015, p. 198363, Jun. 2015.

Jaganathan, P. Rajkumar, N. and Kuppuchamy, R. 2012. ``A comparative study of improved F-score with support vector machine and RBF network for breast cancer classification," *Int. J. Mach. Learn. Comput.*, vol. 2, p. 741.

Kaur S.and S.Kalra,``Feature extraction techniques using support vector machines in disease prediction," in *Proc. IJARSE*, May 2016, p. 5.

Liu, J. Cheng, Y. Wang, X., Zhang, L. and Liu, H.`"An optimal mean based block robust feature extraction method to identify colorectal cancer genes with integrated data," *Sci. Rep.*, vol. 7, p. 8584, Aug. 2017.

Nguyen, C. Wang, Y. and Nguyen, H.N. 2013. ``Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic," *J. Biomed. Sci. Eng.*, vol. 6, no. 5, pp. 551_560.

Pouliot, M., Labrie, Diorio, C. and Durocher, F. 2015. "The role of methylation in breast cancer susceptibility and treatment," *Anticancer Res.*, vol. 35, pp. 4569_4574.

Ren, X. Wang, Y., Zhang, X.-S.an Q. Jin,``iPcc:Anovel feature extraction method for accurate disease class discovery and prediction," *Nucl. Acids Res.*, vol. 41, no. 4, p. e143, 2013.

Schneider, R. ``Survey of peaks/valleys identification in time series," Dept. Inform., Zurich Univ., Zürich, Switzerland, Tech. Rep., Aug. 2011.

Sheather, S. J. ``Density estimation," *Statist. Sci.*, vol. 19, no. 4, pp. 588_597, 2004.

Singh R.K.and M. Sivabalakrishnan, ``Feature selection of gene expression data for cancer classification: A review," *Proc. Comput. Sci.*, vol. 50, pp. 52_57, Jan. 2015.

Terry, M. B. McDonald, J. A. Wu, H. C. Eng, S. and Santella, R. M. 2016. ``Epigeneticbiomarkers of breast cancer risk: Across the breast cancer prevention continuum," *Adv. Exp. Med. Biol.*, vol. 882, pp. 33_68, Mar. 2016.

Wu, J. *et al.*, 2017. ``Identification of biomarkers for predicting lymph node metastasis of stomach cancer using clinical DNA methylation data," *Disease Markers*, vol. 2017, pp. 1_7.

****