International Journal of Recent Development in Computer Technology & Software Applications
Vol. 3, Issue 1 - 2019

ANALYSIS OF DATA MINING ALGORITHM FOR ORAL CANCER USING CLASSIFICATION TECHNIQUES IN R PROGRAMMING

T NITHYA*

ABSTRACT

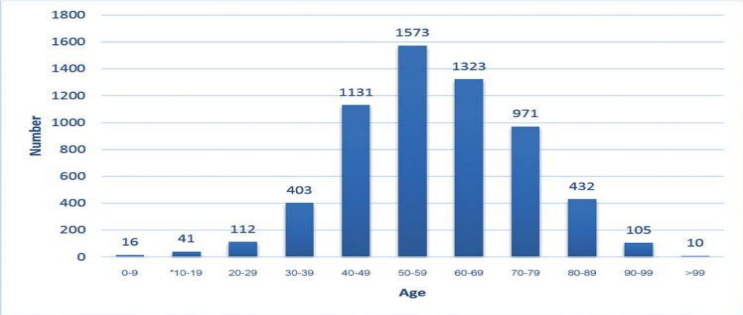
Data mining techniques are used for variety of medical applications.Oral cancer is the most common malignancy worldwide. The aims of this research were to report the prevalence as well as clinic pathologic features of the oral cancer patients from different parts of Asia and Canada and to compare them with patients from other parts of the world. Early detection and prevention of cancer plays a vital role in reducing death of the patients. Three algorithms are used in classification technique for oral cancer such as Naïve Bayes, KNN (k-nearest neighbor's) and SVM for analyzing the best algorithm. Here the cancer risk prediction is proposed which is easy and time saving

KEYWORDS: Classification, Naive Bayes, SVM, KNN.

INTRODUCTION

Oral cancer is the most common malignancy worldwide. As on survey Three hundred thousand patients (2.1% of the total cancer cases) was afflicted with cancer of the oral cavity and lip. It has long been accepted that tobacco consumption including smokeless tobacco and heavy alcohol consumption are the principal factors for the development of oral cancer.

Cigarette smoke contains more than 60 carcinogens according to the International Agency for Research on Cancer. In the Indian subcontinent, some parts of Southeast Asia, and Taiwan, the use of betel quids containing areca nut and lime has long been strongly associated with an increased risk for oral cancer. Alcohol also induces basal cell proliferation and generates free radicals which have the deleterious effects on DNA. In addition, alcohol-associated impairment of the body's ability to breakdown and absorb nutrients and immune suppression may further promote carcinogenesis.



*Assistant Professor, Krishnagiri, Tamilnadu-635108.
Correspondence E-mail Id: editor@eurekajournals.com

 ISSN: 2581-6276

There is a wide variation in the prevalence of oral cancer in different regions of the world or even within the same countries from the minorities or sub-populations. The aims of this research were to report the prevalence as well as clinic pathologic features of the oral cancer patients from different parts of Asia and Canada and to compare them with patients from other parts of the world.

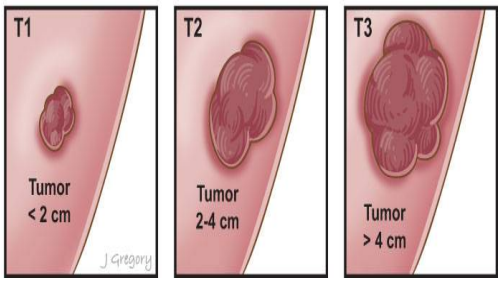


Figure 1.Stages of oral cancer

In this paper, i apply extensive preprocessing techniques to get the accurate nodules in order to enhance the accuracy of detection of oral cancer.

LITERATURE REVIEW

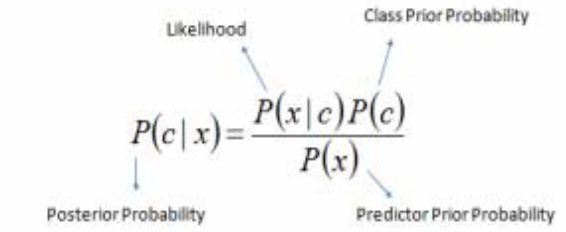
NAÏVE BAYES

Naïve Bayes classifier is statistical classifiers in which it can predict class membership probabilities such that the probabilities of a given tuple fall into a particular class. Naïve Bayes classifier is based on Bayes theorem.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*

 ISSN: 2581-6276

K NEAREST NEIGHBOR

K nearest neighbors is a simple algorithm that stores all available cases and classify new based on similarity measure (e.g. distance function). k nearest neighbor has been used in statistical estimation and pattern recognition Algorithm.

A case is classified by a majority vote of its neighbors. K nearest neighbors measured by a distance function if k=1, then the case is simply assigned to the class of its nearest neighbor.KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output
2. Calculation time
3. Predictive Power

KNN Algorithm is based on feature similarity how closely out-of-sample features resemble. Our training set determines how we classify a given data point:

Example of *k*-NN classification.

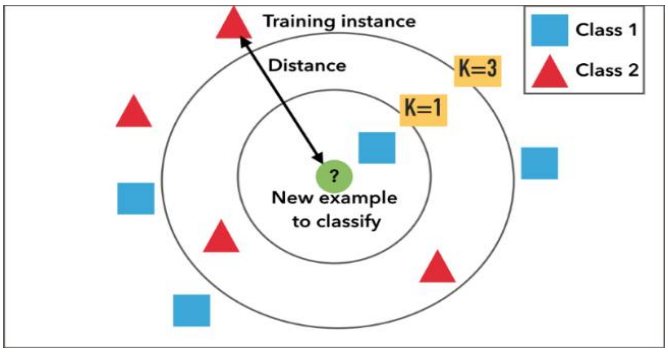


Figure 2.knn Classification

SUPPORT VECTOR MACHINE

The standard SVM implementation SVM takes a input dataset and, for each given input, predicts which of two possible classes the input set belongs to. That's most common use the algorithm to predict if the input belongs to certain dichotomy, or not. Because of this characteristic, SVM is a called a non-probabilistic binary linear classifier.

On Machine Learning-based algorithms such as SVM, the input data has to be separated on two sets: a training set and a test set. The difference between the training and the test set is that, on the training the examples' classes are known beforehand. The test set contains the examples that should have their classes predicted. Given a set of training examples, an SVM algorithm builds a model that predicts what are the categories of the test set's examples.

METHODOLOGY

DATA COLLECTIONS

An oral dataset was taken from the ncbi repository and it is made up of 1077 raw data from which various attributes were published. It have the attributes age, gender alcohol consumption, tobacco intake, dietary supplements and so on.

R PROGRAMMING

R language is an open source program maintained by the R core-development team-team of volunteer developers from across the globe.R language used for performing statistical operations. R is a command line driven program. The user enters commands at the prompt (> by default) and each command is executed one at a time. Many routines have been written for R analytics by people all over

 ISSN: 2581-6276

the world and made freely available from the R project website as packages. R is a consolidated environment for performing statistical operations and generating R data analysis reports in graphical or text formats. R commands entered in the console are evaluated and executed.

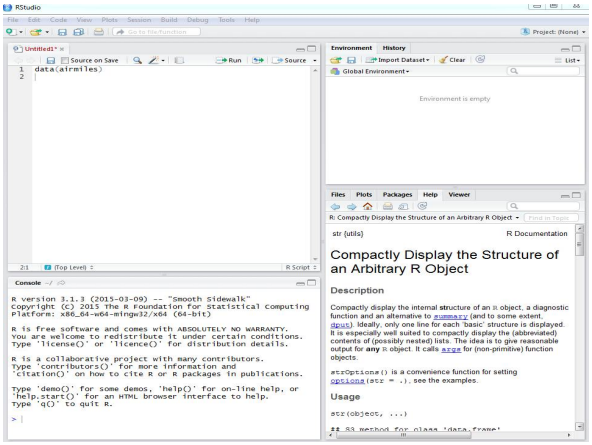


Figure 4.R studio Window

EXPERIMENTAL RESULT

This work is implemented with the help of R studio software contains several classification algorithms that are used to classify and the performance analysis is evaluated. It can work with wide variety of data fields including xlsx and csv file format. Comparative study is made between three algorithms. The performance of three algorithms is calculated based on the accuracy of classification.

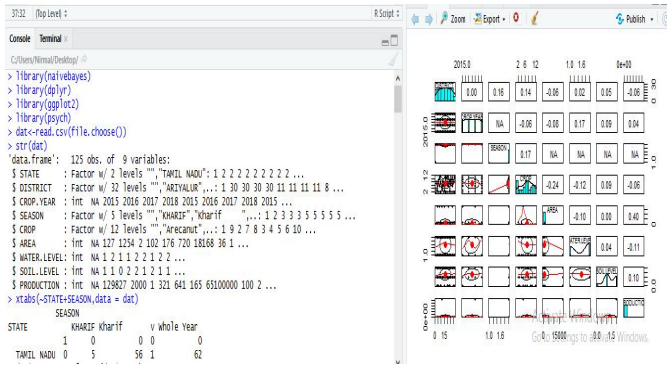


Figure 5.Naïve bayes classification

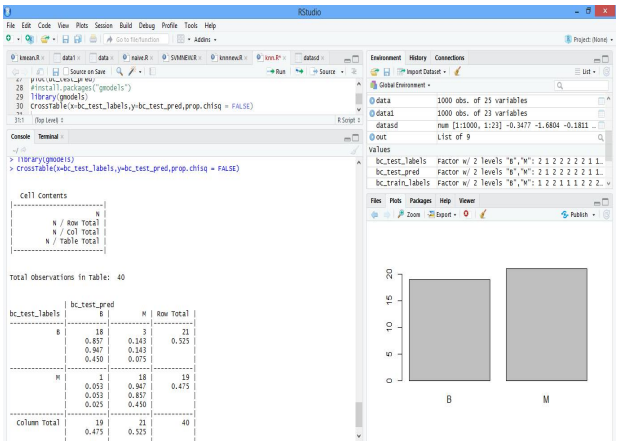


Figure 6.SVM classification output

 ISSN: 2581-6276

CONCLUSION

Although the prevalence of oral cancer is not high compared to other entities, oral cancer pose significant mortality and morbidity in the patients, especially when discovered late in the course of the disease. This study highlights some anatomical locations where oral cancers are frequently encountered. As a result, clinicians should pay attention to not only teeth, but oral mucosa especially in the high prevalence area.

The SVM algorithm proves high accuracy and result are comparing with Naïve bayes algorithm and Knn. Comparing to all other cancers, oral cancer is one of the major causes of death in people. So, the early detection of this cancer is needed in reducing life losses. In this paper we have applied techniques namely classification for predicting oral cancer as accurately as possible.

The dataset will be applied into R tool using classification technique. Finally SVM produces the good accuracy of cancer cell detection.

REFERENCES

[1]. Warnakulasuriya S. Causes of oral cancer – An appraisal of controversies. Br Dent J. 2009; 207:471–5. [PubMed].

[2]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5822535/.

[3]. ICMR Report 2006. Cancer Research in ICMR Achievements in Nineties.

[4]. Osmar R. Zaïane, Principles of Knowledge Discoveryin Databases. [Online]. Available: webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf.

[5]. The Data Mining Process. [Online]. Available: http://publib.boulder.ibm.com/infocenter/db2luw/v9/index.jsp?topic=/com.ibm.im.easy.doc/c_m_process.html.Shelly Gupta et al./Indian Journal of Computer Science and Engineering (IJCSE).

[6]. "R and DATA MINING Examples and Case Studies", Book 2013 Yanchang Zhao. Published by Elsevier Inc.

 ISSN: 2581-6276