CANCER PREDICTION

S. Dhivya, S. Salai Saidhanya & S. Jasima Yasmin
Students/ Department of CSE, ACGCET, KKD, Tamil Nadu, India

Dr. C. Umarani
Faculty / Department of ICE, ACGCET, KKD, Tamil Nadu, India

Abstract

Cancer is one of the leading causes death world wide. Early detection and prevention of cancer is very important for reducing deaths caused by cancer. Identification of genetic and environmental factors is very important in developing novel methods to detect and prevent cancer. Therefore a novel multi layered method combining clustering and decision tree techniques to build a cancer risk prediction system is proposed here which predicts lung, breast, oral, cervix, stomach, blood, brain, eye cancers and is also user friendly, time and cost saving. This research uses data mining technology such a classification, clustering and prediction to identify potential cancer patients. The gathered data is preprocessed, fed into database and classified to yield significant patterns using decision tree algorithm. Then the data is clustered using K-means clustering algorithm to separate cancer and non cancer patient data. Comparing to K- means clustering by using j48 and c4.5 this tool produces an accurate results and easier to use and it's very efficient comparing to other. Further the cancer cluster is subdivided into ten clusters. Finally a prediction system id developed to analyze risk levels which help in prognosis. This research helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming.

Keywords*: Data Mining process, Classification, K means clustering, Decision Tree, Cancer prediction.*

Introduction

Every day we learn something about cancer. Cancer is different from other diseases. Unlike other diseases cancer is not based on "mechanical breakdowns", "biochemical imbalances", etc..,. Instead of these Cancer is a disease in which abnormal cells divide uncontrollably and destroy body tissues. This can result in tumours, damage to the immune system and other impairment that can be fatal. These cells can infiltrate normal body tissues.

Many cancers on the abnormal cells that compose the cancer tissue are further identified by the name of the tissue that the abnormal cells originated from (for example breast cancer, lung cancer, skin cancer) Cancer is not confined to humans. Animals and other living organism can get cancer.

The incidence of cancer and cancer types are influenced by many factors such as age, gender, race, local environment factors, diet, and genetics. Consequently, the incidence of cancer and cancer types vary depending on these variable factors. For example, the World Health Organization (WHO) provides the following general information about cancer worldwide. Cancer is a leading cause of death worldwide. It accounted for 8.2 million deaths (around 22% of all deaths not related to communicable diseases; most recent data from WHO). Lung, stomach, liver, colon, and breast cancer cause the most cancer deaths each year.

Data mining technique involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithms and machine learning methods in early detection of cancer.

In Classification learning, the learning scheme is presented with the set of classified examples from which it is expected to learn a way of classifying unseen examples.

In Association learning, any association among features is sought, not just ones to predict a particular class value. In Clustering, group of examples that belong together are sought. This project uses a new tool called Weka. What makes Weka worthy of try is Easy learning curve. For someone who doesn't coded for a while, Weka with its GUI provides easiest transition into the world of Data Science. Weka is a collection of Machine Learning algorithms for data mining tasks. The algorithms can either be applied directly from a dataset or called from your own Java code. Weka contains tools for

1. Data preprocessing
2. Classification
3. Clustering
4. Association Rules
5. Regression
6. Visualization

The datasets can be collected in ARFF format and then inserted to the Weka tool for processing through Machine Learning techniques. This is tool is used by everyone and it gives a clear understanding to the user to easily know a person's cancer status and severity without screening them for testing cancer. Also it is useful to record and save large volumes of sensitive information which can be used to gain knowledge about the disease and its treatment.

Literature Survey

We referred to a few previously published papers that had relatively similar objective as. Most of the papers had ideas about how to detect and predict the cancer in early stage. Everybody focused on such complex codings like Java, R programming, Python, etc. This codings are difficult and not used by everyone. So we use a new tool called Weka for detecting and predicting the cancer. Most of the papers mainly focused on specified cancer but this paper highly focused on generalized cancer.

V.Krishnaiah et al [2] developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be naïve bayes followed by IF-THEN rule, decision tree and neural network. For Diagnosis of Lung cancer Disease Naïve Bayes observes better results and fared better than Decision Tree.

Labeed K Abdulgafoor et al [10] wavelet transformation and K-mean clustering algorithm have been used for intensity based segmentation

Proposed Model

The following is the model of the Proposed work. The collected data is pre-processed and the entire data is taken as training set to build the classification and clustering model. The classification model is build the classifiers such as Zero Rule, M5 rule, Decision table rule. The clustering model is build using Simple K-Means, Expectation Maximization, canopy, cobweb, Farthest first, Filtered

Clustered, etc. This model is tested for accuracy, sensitivity, specificity using the test data. Finally the model is visualized. The below fig 3.1 shows that the proposed system for our work.

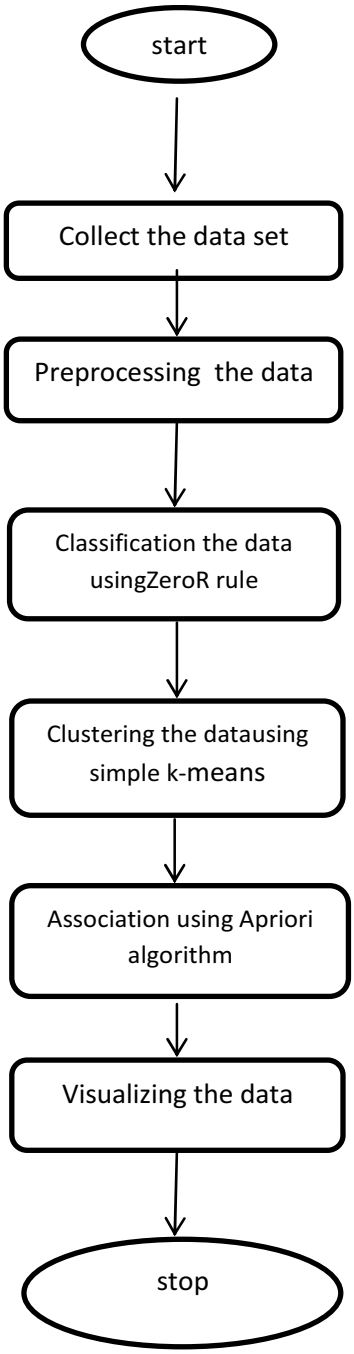


Fig 3.1: Proposed work

Experimental Results

The dataset can be collected in a comma separated variable (csv) format.

Table 4.1: Dataset for Cancer

id	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	0	3	0	1	0	2	2	2	1	
2	0	3	3	3	0	3	1	3	1	
3	0	3	3	4	0	3	3	3	1	
4	0	2	3	4	1	3	3	3	1	
5	0	3	2	3	1	3	3	3	2	
6	0	3	3	4	0	3	3	3	1	
7	0	3	2	3	0	3	3	3	1	
8	0	2	2	3	0	3	1	3	3	
9	0	3	1	3	0	3	1	3	1	
10	0	2	3	4	0	2	2	2	1	
11	0	2	2	2	0	3	2	3	1	

The input is given as in thearff that is attribute related file format.

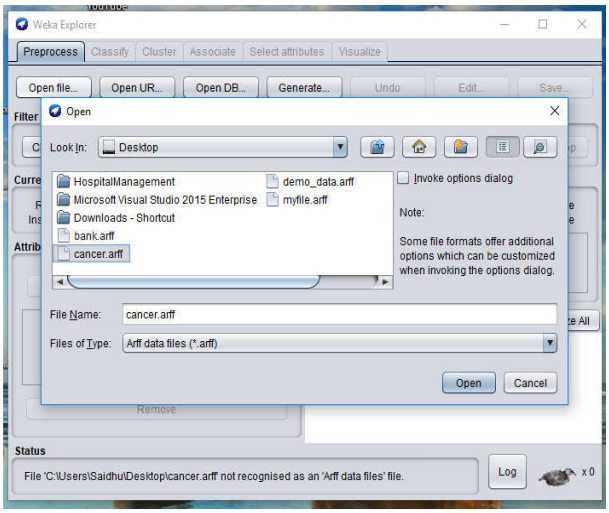


Fig 4.1: Input screen in Weka tool

The table 4.1shows that the data preprocessed and then data to be visualized.
The fig 4.1 shows that the data to be classified and then clustered.

Statistic	Value
Minimum	18
Maximum	67
Mean	42.395
StdDev	14.425

Fig 4.2: Based on age the data is evaluated

The fig 4.2 shows that the data for age variation.
The fig 4.3 shows that the representation of visualized of data.
The fig 4.4 shows that the various cluster data.

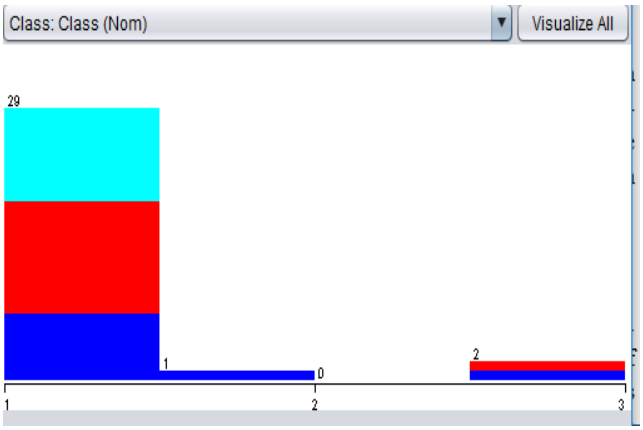


Fig 4.3: Visualized representation of data

The fig 4.5 shows that the classification of data in confusion matrix.

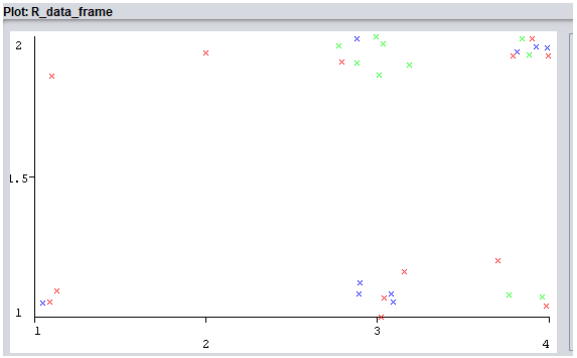
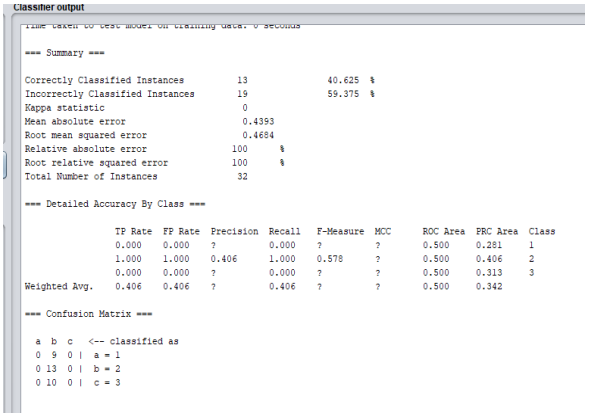


Fig 4.4: Clustering the data



Classifier output

```
=== Summary ===

Correctly Classified Instances          13               40.625  %
Incorrectly Classified Instances        19               59.375  %
Kappa statistic                          0
Mean absolute error                      0.4393
Root mean squared error                  0.4684
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances               32

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.000    0.000    ?          0.000    ?          ?        0.500     0.281     1
               1.000    1.000    0.406      1.000    0.578      ?        0.500     0.406     2
               0.000    0.000    ?          0.000    ?          ?        0.500     0.313     3
Weighted Avg.  0.406    0.406    ?          0.406    ?          ?        0.500     0.342

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0  9  0 |  a = 1
  0 13  0 |  b = 2
  0 10  0 |  c = 3
```

Fig 4.5: Classification of data in Confusion matrix format

Comparison

When Compared to others, we have listed all types of cancer. Also we have given preference to the most deadliest type of cancer which kills an average of 70% of people per year.

Conclusion & Future Enhancement

In this paper a multilayer methods combing preprocessing, classify, cluster and visualization techniques is to build the is prediction system is proposed. The cancer is become the leading cause of death worldwide. The most effective way to reduce cancer deaths is to detect it earlier. Many people avoid cancer screening is due to cost involved in several test for diagnosis this prediction system may provide easy and cost effective way for screening cancer and may play a pivotal role in earlier diagnosis process for different types of cancer and its effective preventive strategy.

In this paper, the overall view of the organs can be taken for achieving the result by using weka tool. In Future, Cancer can be distinctly identified by Men cancer and Women cancer accurately by using this efficient tool.

Acknowledgement

We hereby show our gratitude to our college. “Alagappa Chettiar Govt. College of Engineering and Technology”. We are using this opportunity to prove and improve ourself. We should like to extend our gratitude to the TEQIP-III (a world bank assisted project through NPIU, New Delhi) which provided us with great financial support for us through the endeavor through this work.

References

1. Neelamadhab Padhy “The Survey of Data Mining Applications and Feature Scope” Asian Journal of Computer Science and Information Technology 2:4 (2012) 68-77 ISSN: 2249-5126
2. V.Krishnaiah “Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques” International Journal of Computer Science and Information Technologies, Vol. 4(1) 2013, 39 – 45.
3. Rafaqat Alam Khan “Classification and Regression Analysis of the Prognostic Breast Cancer using Generation Optimizing Algorithms” International Journal of Computer Applications (0975-8887) Volume 68- No.25, April 2013.
4. K.Kalaivani “Childhood Cancer-a Hospital based study using Decision Tree Techniques” Journal of Computer Science 7(12): 1819-1823, 2011 ISSN: 1549-3636.
5. Boris Milovic “Prediction and Decision Making in Health Care using Data Mining” International Journal of Public Health Science Vol. 1, No. 2, December 2012, pp. 69-78 ISSN: 2252-8806.
6. T.Revathi “A Survey on Data Mining Using Clustering Techniques” International Journal of Scientific & Engineering Research Http://Www.Ijser.Org, Volume 4, Issue 1, January-2013, Issn 2229-5518.
7. Shomona Gracia Jacob “Data Mining in Clinical Data Sets: A. Review” International Journals of Applied Information System (IJAIS) - ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA, Volume 4-No.6, December 2012-www.ijais.org.

8. G. Rajkumar " Intelligent Pattern Mining and Data Clustering for Pattern Cluster Analysis using Cancer Data" International journal of Engineering Science and Technology Vol. 2(12), 2010, ISSN: 7459-7469.
9. M. Durairaj "Data Mining Applications in Healthcare Sector: A Study" International journal of Scientific & Technology Research, Volume 2, Issue 10, October 2013, ISSN: 2277-8616.
10. Labeed K Abdulgafoor "Detection of Brain Tumour Using Modified K-Means Algorithm and SVM" International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013.