An Analysis of Cancer Affected People using Classification Data Mining Algorithms

H. Lookman Sithic, R. Uma Rani

Abstract*— Data mining is a collection of exploration techniques based on advanced analytical methods and tools for handling a large amount of information. The techniques can find novel patterns that may assist as enterprise in understanding the business better and in forecasting. Much research is being carried out in applying data mining to a variety of applications in healthcare[1]. This article explores data mining techniques in healthcare management. Particularly, it talk about data mining and its various application in areas where people are mostly affected rigorously by cancer in Erode District, Tamil Nadu, India. The people affected by cancer using tobacco, chemical water. This paper identifies the cancer level using classification algorithms and finds meaningful hidden patterns which gives meaningful decision making to this socio-economic real world health venture

Keywords: Data Mining, Cancer, Classification algorithms.

I. INTRODUCTION

A. Cancer

Cancer is actually a group of many related diseases that all have to do with cells. Cells are the very small units that make up all living things, including the human body. There are billions of cells in each person's body. Cancer happens when cells that are not normal grow and spread very fast. Normal body cells grow and divide and know to stop growing. Over time, they also die. Unlike these normal cells, cancer cells just continue to grow and divide out of control and don't die when they're supposed to. Cancer cells usually group or clump together to form tumors . A growing tumor becomes a lump of cancer cells that can destroy the normal cells around the tumor and damage the body's healthy tissues. This can make someone very sick.

i. Oral cancer

India has the dubious distinction of harboring the world's largest number of oral cancer patient with an annual age standardized incidence of 12.5per 100,000.

The treatment is successful only if the lesion is diagnosed early. Globally , about ,5,75000 new cases and 3,20,000 deaths occur every year from oral cancer[2]. Most oral cancers in india present in advanced stage of malignancy. One of the main barriers to treatment and control of oral cancer is the identification and risk assessment of early disease in the community in a cost effective fashion. Oral cancer is a subtype of head and neck cancer and is any cancerous growth located in any subsites of the oral cavity [3]. Oral cancers may originate in any of tissues of the mouth. Oral cancer most commonly involves the tongue. The symptoms for an oral cancer at an earlier stage [4] are : 1)patches inside the mouth or on lips that are white, red or mixture o white and red.2)Bleeding in the mouth.3)dIfficulty or pain when Swallowing.4)lump in the neck. These symptoms should raise the suspicion of cancer and needs proper treatment. The treatment is successful only if the lesion is diagnoses early, but sadly many times, it is ignored and the patient reports late when the lesion is untreatable. Most people contract cancer owing to environmental problems. Food path cancer is on the increase and oral cancer is decreasing in Erode district. Erode is located on the banks of Cauvery River and there are many villages on the banks of Kalingarayan Canal. Farmers and the public complain that owing to abundant use of chemicals and large-scale discharge of effluents into water sources many farmers and cattle are affected. that heavy discharge of effluents from tanning and textile industries into Kalingarayan canal contaminated the canal water and also the ground water. The farmers who used this fell victims to cancer. The goal of this paper is to find out the people who are affected by the cancer by using the data mining classification algorithm.

II. LITERATURE OF REVIEW

Erode is located on the banks of Cauvery River and there are many villages on the banks of Kalingarayan Canal. Farmers and the public complain that owing to abundant use of chemicals and large-scale discharge of effluents into water sources many farmers and cattle are affected. Union Minister for Social Justice and Empowerment Ms Subbulakshmi Jagadeesan said an unofficial survey conducted in many villages found that more than 100 women and 75 men were victims of the deadly disease. Ms. Jagadeesan told the Tamil Nadu Government that heavy discharge of effluents from tanning and textile industries into Kalingarayan canal contaminated the canal water and also the ground water. The farmers who used this fell victims to cancer. Erode district is witnessing an alarming number of cancer cases due to drinking water contamination from the deadly chemical discharge by various factory units into kalingarayan canal. The secretary of Tamilaga Vivasayeegal Sangam T Subbu says that more than 10 textile unit SIPCOT industrial growth centre at Perundurai alone have been letting out harmful effluents into drains in six or seven villages, badly affecting the ground water. "Erode district is one of the worst hit cancer districts in Tamil Nadu and as on date, within just 18 months of starting the IICG cancer hospital, 1,320 cancer cases were examined in Erode alone," says Dr P Suthahar, consultant radiologist at the hospital.

Manuscript Received on May 12, 2015.

H. Lookman Sithic, Asst. Prof, Department of Computer Science, Muthayammal College of Arts & Science, Rasipuram, Periyar university, Salem.

Dr. R. Uma Rani, Associate Professor, Department of Computer Science, Sri Saradha College for Women, Salem, India.

He said 35 to 40 per cent of those examined had liver and bladder cancer, a clear indication that it was due to consumption of water contaminated with dyes and chemicals. Pollution Control Board personnel on their part maintained that they were taking stringent action against polluting units and that 500 units had been sealed for violation of pollution control rules. They also said action was being taken against those who have not set up reverse osmosis plants. Erode district is one of the worst hit cancer district in Tamilnadu. The IICG cancer hospital 1320 cancer cases were examined in Erode alone.

A. Data Preparation

Based on the information from various physician, we have prepared questionnaires to get raw data from too many villagers who affected with high level cancer. People of different age groups with different ailments were interviewed based on the questionnaires prepared in our mother tongue i.e tamil to avoid communication problem Total data collected from villages 250 (men) From the medical practitioners advice, while classifying the data, the degree of disease symptoms are placed in several compartments as follows :

None

Mild cancer

Moderate cancer

Severe cancer

The above types are classified by the following rules:

(i) No symptoms found grouped or any one symptoms as none.

(ii) Those who are found with two symptoms are grouped as Mild disease.

(iii) Those who are found with three symptoms moderate disease.

(iv) Those who are found with than three symptoms severe diseases.

B. Classification as the Data mining application

Classification is a form of data analysis that can be used to extract models describing important data classes. Such analysis will provide us a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels. Classification have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing and medical diagnosis.[5]

C. Weka as a data miner tool

In this paper we have used WEKA (to find interesting patterns in the selected dataset), a Data Mining tool for classification techniques.. The selected software is able to provide the required data mining functions and methodologies. The suitable data format for WEKA data mining software are MS Excel and ARFF formats respectively. Scalability-Maximum number of columns and rows the software can efficiently handle. However, in the selected data set, the number of columns and the number of records were reduced. WEKA is developed at the University of Waikato in New Zealand. "WEKA" stands for the Waikato Environment of Knowledge Analysis. The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and WEKA has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. WEKA expects the data to be fed into be in ARFF format (Attribution Relation File Format).[6]

WEKA has two primary modes: experiment mode and exploration mode .The exploration mode allows easy access to all of WEKA's data preprocessing, learning, data processing, attribute selection and data visualization modules in an environment that encourages initial exploration of data. The experiment mode
allows larger-scale experiments to be run with results stored in a database for retrieval and analysis. [6]

D. Classification in WEKA

The basic classification is based on supervised algorithms. Algorithms are applicable for the input data. Classification is done to know exactly how the data is being classified. The Classify Tab is also supported which shows the list of machine learning tools. These tools in general operate on a classification algorithm and run it multiple times to manipulating algorithm parameters or input data weight to increase the accuracy of the classifier. Two learning performance evaluators are included with WEKA. The first simply splits a dataset into training and test data, while the second performs cross-validation using folds. Evaluation is usually described by the accuracy. The run information is also displayed, for quick inspection of how well a classifier works. [6]

E. Manifold Machine learning algorithm

The main motivation for different supervised machine learning algorithms is accuracy improvement. Different algorithms use different rule for generalizing different representations of the knowledge. Therefore, they tend to error on different parts of the instance space. The combined use of different algorithms could lead to the correction of the individual uncorrelated errors. As a result

Table 1: Selected Attributes

Classifier tool	Experimenter accuracy
Simple Cart	90.8
J48	95.6
Naïve Bayes	88

the error rate and time taken to develop the algorithm is compared with different algorithm.[6]

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

F. Experimental Setup

The data mining method used to build the model is classification. The data analysis is processed using WEKA data mining tool for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 250 instances with 10 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of cancer affected persons. The performance of the classifiers is eval uated and their results are analyzed. The results of comparison are based on ten-fold cross-validations.

According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing.[6]

G. Learning Algorithms

This paper consists of three different supervised machine learning algorithms derived from the WEKA data mining tool. Which include?

- J48 (C4.5)
- Naive Bayes,
- CART

The above algorithms were used to predict the accuracy of Fluoride Dental diseases affected persons.

III. DISCUSSIONS

A. Attributes selection

First of all, we have to find the correlated attributes for finding the hidden pattern for the problem stated. The WEKA data miner tool has supported many in built learning algorithms for correlated attributes. There are many filtered tools for this analysis but we have selected one among them by trial.[7] Totally there are 250 records of data base which have been created in Excel 2007 and saved in the format of CSV (Comma Separated Value format) that converted to the WEKA accepted of ARFF by using command line premier of WEKA. The records of data base consists of 10 attributes, from which 8 attributes were selected based on attribute selection. We have chosen Symmetrical random filter tester for attribute selection in WEKA attribute selector. It listed all 10 attributed with rank, but from which we have taken only 8 attributes . The other attributes S.No., Name, omitted for the convenience of analysis of finding impaction among peoples in the district. [8].

B. Classifier chosen using Ranker testing in WEKA

The Classify option in WEKA has many learning tools for finding hidden patterns based on classification. We can choose the best learning tool for the created learning data base from the ranking test in WEKA Experimenter option. Randomly we have chosen six learning algorithms and applied in Experimenter.

Table 2: Algorithms used in Classification

S. No	Name of the Attribute
1	Age
2	Designation
3	Smoking or Drinking
4	White or red color patch on the tongue
5	Difficulties in sudden swelling
6	wound in mouth
7	Bleeding
8	Class

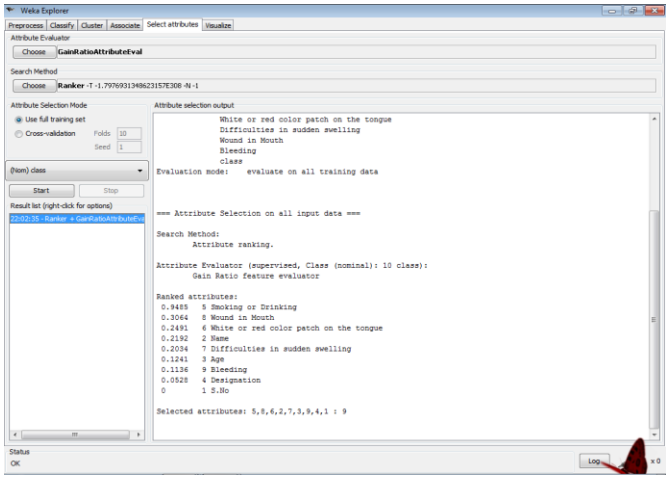

Fig 1 : Attributes selection in Weka 3.6.4

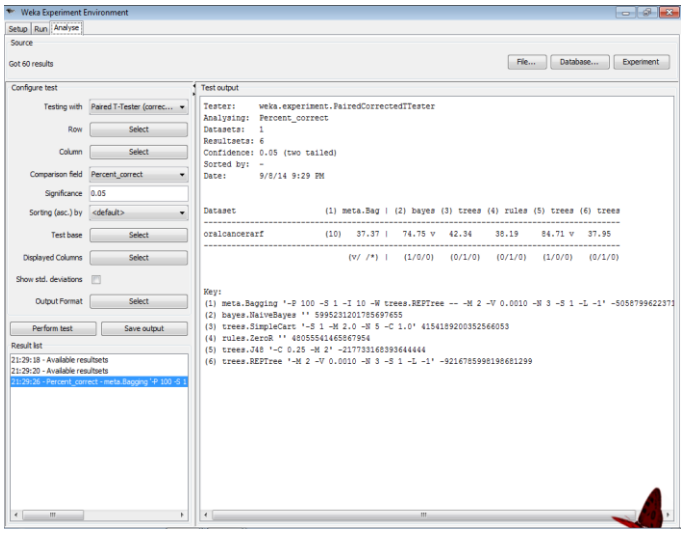

Fig 2 : Algorithms selection in Weka 3.6.4 Experimenter

The Experimenter has given above the accuracy over the created learning data base. We have chosen two high accuracy and randomly chosen one medium accuracy learning algorithms which have highlighted in the above table to find the hidden pattern of the classification. [8]

C. J48 Algorithm in WEKA

The J4.8 decision tree in WEKA is based on the C4.5 decision tree algorithm. The C4.5 algorithm is a part of the multi-way split decision tree. C 4.5 yields a binary split if the selected variable is numerical, but if there are other variables representing the attributes it will result in a categorical split. That is, the node will be split into C nodes where C is the number of categories for that attribute. The learning algorithm J48 in WEKA 3.6.4 accepts the training data base in the format of ARFF. It accepts the nominal data and binary sets. So our attributes selected in nominal and binary formats naturally. So no need of preprocessing for further process.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

We have trained the training data by using the 10 Fold Cross Validated testing which used our trained data set as one third of the data for training and remaining for testing. After training and testing which gives the following results[9]. From the WEKA 3.6.4 classifier Confusion matrix confirms that the Erode District people are impacted by Mild cancer disease.

D. Classification And Regression Tree (CART) Algorithm in WEKA

It builds a binary decision tree by splitting the records at each node, according to a function of a single attribute. CART uses the Gini index for determining the best split. The initial split produces two nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine all the input fields to find candidate splitters. If no split can be found that significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of over fitting.. [1]

At the end of the tree growing process, every record of the training set has been assigned to some leaf of the full decision tree. Each leaf can now be assigned a class and an error rate. The error rate of a leaf node is the percentage of incorrect classification at that node. The error rate of an entire decision tree is a weighted sum of the error rates of all the leaves. Each leaf's contribution to the total is the error rate at that leaf multiplied by the probability that a record will end up in there.

We have trained the training data by using the 10 Fold Cross Validated testing. The CART decision tree classifier Confusion matrix too confirms the same result obtained in the J48 decision tree. That is the Eroade District are impacted by Mild cancer diseases.

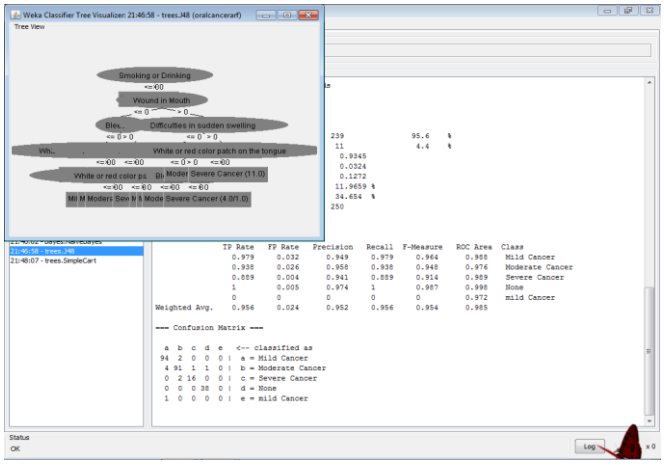

Fig 3 : J48 Decision Tree in Weka 3.6.4

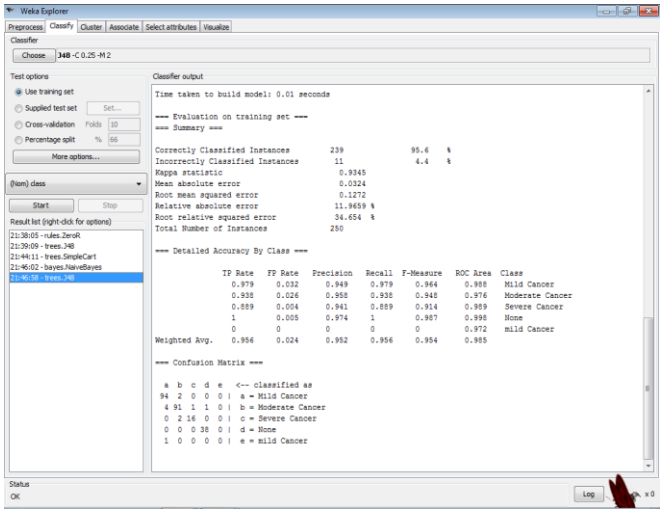

Fig4 : J48 implementation in WEKA 3.6.4

Bayesian classification is quite different from the decision tree approach. In Bayesian classification we have a hypothesis that the given data belongs to a particular class. We then calculate the probability for the hypothesis to be true. This is among the most practical approaches for certain types of problems. The approach requires only one scan of the whole data. The expression P(A) refers to the probability that event A will occur. P(A/B) stands for the probability that event A will happen given that event B has already happened. In other words p(A/B) is the conditional probability of A based on the condition that B has already happened. For example, A and B may be probability of passing a course A and passing another course B respectively. P(A/B) then is the probability of passing A when we know that B has been passed.[1]

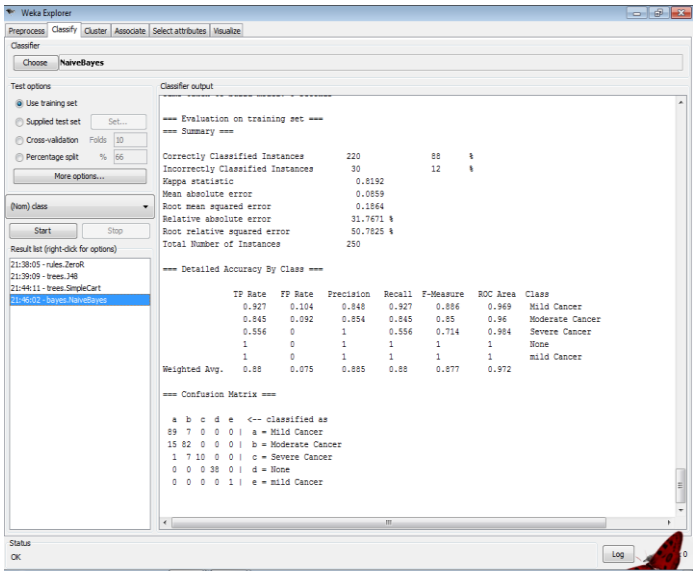

Fig 4 : Naïve Bayes implementation in Weka 3.6.4

Now the Bayes theorem

P(A/B) = P(B/A)P(A)/P(B)

If we consider X to be an object to be classified then Bayes theorem may be read as giving the probability of it belonging to one of the classes C_1,C_2,C_3, etc by calculating $P(C_i/X)$. Once these probabilities have been computed for all the classes, we simply assign X to the class that the highest conditional probability. [1]

$P(C_i/X)$ may be calculated as

$$P(C_i/X) = [P(X/C_i)P(C_i)]/P(X)$$

- $P(C_i/X)$ is the probability of lthe object X belonging to class C_r .
- $P(X/C_i)$ is the probability of obtaining attribute values X if we know that it belongs to class C_r.
- $P(C_r)$ is the probability of any object belonging to class C_i without any other information.
- P(X) is the probability of obtaining attribute values X whatever class the object belongs to.

The Naïve Bayes classifier Confusion matrix also declares the same result obtained in the J48 and CART decision trees. That is the Eroade District is impacted by Mild cancer disease.

IV. RESULT COMPARISION

The above implementation algorithms yields the same results that the Erode District residing people affected by the Mild Cancer disease.. However some key parameters which played important role in which algorithm works better.

Table 3 : Comparison of Accuracy

Clasification algorithm tree type	% of correctly classified instances	Root mean square error	Time take to build the model(In seconds
J48	95.6 %	0.1272	0.01
Simple Cart	90 %	0.1378	0.05
Naïve Bayes	88 %	0.1864	0.47

All the three classified learning algorithms train the data up to 90% so the error rate completely reduced. The time taken to build the algorithm relatively too small. The root mean square error reduced when the % of correctly classified increased. It shows that the J48 have less error and high % of classified rate and time taken to build the model faster than other two algorithms. It clearly shows that the J48 algorithm works better on the cancer affected data.

V. CONCLUSION

Data mining applied in health care domain, by which the people get beneficial for their lives. As the analog of this research found the meaningful hidden pattern that from the real data set collected the people impacted in Erode District by consuming the contaminated Kalingarayan canal water and also the ground water by discharge of effluents from tanning and textile industries.. The farmers who used this fell victims to cancer. By this analysis we can easily know that the people do not get awareness among themselves about the cancer impaction. If it continues in this way, it may lead to some primary health hazards sever oral and breast cancer diseases.

The impaction of Mild cancer in this area disturbed their daily meager life. It is primary duty of the Government to providing good hygienic drinking water to the people and reduce the discharge of effluents from tanning and textile industries with the latest technologies and creating awareness among the people in some way like medical camps and taking documentary films. If continues in this way after 10 to 20 years there may be the possibilities of Severe cancer impaction among people in Erode District. Through this research the problem of cancer in this district came to light. It is a big social relevant problem.

REFERENCE

1. Introduction to Data Mining with Case Studies – G.K.Gupta
2. Langdon JD, Russel RC , Williams NS, Bulstrode CJK Arnold, Oral and Oropharyngeal cancer practice of surgery, London: Hodder Headline Group;2000.
3. Werning, John W (may 16,2007). Oral cancer : Diagnosis, Management, and rehabilitation. P.1.ISBN 978 – 1588903099.
4. crispian scully, Jose.V.Bagan, Colin Hopper, Joel.B.Epstien, "oral Cancer: Current and future diagnostics Techniques – A review article", American journal of Dentistry, vol. 21,No.4,pp 199-209, Augest 2008.
5. Arun K.Pujari, "Data mining Techniques", University Press, First edition, fourteenth reprint, 2009.
6. Peter Reutemann, Ian H. Witten,"The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, issue 1 pages 10 to 18, 2005.
7. Weka 3.6.4 data miner manual. 2010
8. P.Rajeswari, G.Sophia Reena, "Analysis of Liver Disorder Using Data mining Algorithms", Global Journal of computer science and Technology, Volume 1, issue 1, November 2010 page 48 to 52. ISSN:0975-4172
9. A.V.L.N.S.H.Hariharan and K.S.R. Murthy, "Determination of Fluoride in and around Visakhapatnam City, Andhra Pradesh", International Journal of Applied Biology and Pharmaceutical Technology, Volume: I: Issue-3: Nov-Dec -2010, pages 1261-64 ISSN:0976-4550.

AUTHORS PROFILE

Dr. R. Uma Rani received her Ph.D., Degree from Periyar University, Salem in the year 2006. She is a rank holder in M.C.A., from NIT, Trichy. She has published around 40 papers in reputed journals and national and international conferences. She has received the best paper award from VIT, Vellore , Tamil Nadu in an international conference. She was the PI for MRP funded by UGC. She has acted as resource person in various national and international conferences. She is currently guiding 5 Ph.D., scholars. She has guided 20 M.Phil., scholars and currently guiding 4 M.Phil., Scholars. Her areas of interest include information security, data mining, fuzzy logic and mobile computing.

H. Lookman Sithic received his M.S (IT) nce in Jamal Mohamed College, Trichy under Bharathidasan University and M.Phil Degree from Periyar University. Now persuing his Ph.D research under Bharathiar University, Coimbatore. Doing research under health care domain in Datamining applications. He published research papers in various National, International conferences and International journal.