A Grouping of Cancer in Human Health using Clustering Data Mining Technique

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 clustering algorithms and finds meaningful hidden patterns which gives meaningful decision making to this socio-economic real world health venture.

Keywords: Data Mining, Cancer, Clustering 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.

B. 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 sub sites 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,

Manuscript Received on May 12, 2015.

H.Lookman Sithic, Asst. Professor Department of Computer Applications, Periyar University, Muthayammal College of Arts & Science, Namakkal, India.

R. UmaRani, Associate Professor, Department of Computer Science, Sri Saradha college for Women, Salem, India.

consultant radiologist at the hospital.

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 424 peoples.
From the medical practitioner's 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.

Those who are found with than three symptoms severe diseases

B. Clustering as the Data Mining application

Clustering is one of the central concepts in the field of unsupervised data analysis, it is also a very controversial issue, and the very meaning of the concept "clustering" may vary a great deal between different scientific disciplines. However, a common goal in all cases is that the objective is to find a structural representation of data by grouping (in some sense) similar data items together. A cluster has high similarity in comparison to one another but is very dissimilar to objects in other clusters.

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 clustering 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). 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.[8]

D. Clustering in WEKA

The basic classification is based on supervised algorithms. Algorithms are applicable for the input data. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.. The Cluster tab is also supported which shows the list of machine learning tools. These tools in general operate on a clustering algorithm and run it

Table 1 Classification of Symptoms of Diseases

Smoking or Drinking	Patch on the Tongue	Sudden Swelling	Wound in Mouth	Bleeding	Remark
No	No	No	No	None	--
Low	Low	Low	--	Mild	Any three Low Symptom
Low	Low	Low	Low	Mode-rate	--
Low	Low	Medium	Medi-um	Mode-rate	Any two Medium Symptoms
Low	Medium	High	High	Severe	Any two High Symptoms

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 cluster works.

E. Experimental Setup

The data mining method used to build the model is cluster. 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 424 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. 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]

F. Learning Algorithm

This paper consists of an unsupervised machine learning algorithm for clustering derived from the WEKA data mining tool. Which include:

- K-Means

Above clustering model used to cluster the group of people who are affected by cancer in Erode District

III. DISCUSSION AND RESULT

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.[5]

Totally there are 424 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 15 attributes, from which 10 attributes were selected based on attribute selection in explorer mode of WEKA 3.6.4. (fig 1)

We have chosen Symmetrical random filter tester for attribute selection in WEKA attribute selector. It listed 14 selected attributes, but from which we have taken only 8 attributes . The other attributes are omitted for the convenience of analysis of finding impaction among peoples in the district.

TABLE 2: CLASSIFICATION OF ATTRIBUTES

No.	Attributes
1.	S.No.
2.	Name
3.	Age
4.	Designation
5.	Smoking orDrinking
6.	White or Red color patch on the tongue
7.	Difficulties in Sudden swelling
8.	Wound in Mouth
9.	Bleeding
10.	Class

Table 3: Selected Attributes for Analysis

No.	Attributes
1.	Age
2.	Designation
3.	Smoking orDrinking
4..	White or Red color patch on the tongue
5.	Difficulties in Sudden swelling
6.	Wound in Mouth
7..	Bleeding
8.	Class

B. K-Means Method

The k-Means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed a the cluster's *centroid or center of gravity.*

The K–Means algorithm proceeds as follows:

First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterated until the criterion function converges. Typically, the square-error criterion is used, defined as

$$E=\sum_{I=1}^{K}\sum_{P\in C_i}|p-m_i|^2 \qquad (1)$$

Where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and m_i is the mean of cluster C_i. In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible.

i. K-Means algorithm:

Input;
= k:the number of clusters,
= D:a data set containing n objects
Output: A set of k clusters.
Method:

(1)arbitrarily choose k objects from from D as the initial cluster centers;
(2)repeat
(3) (re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;

Suppose that there is a set of objects located in space as depicted in the rectangle shown in fig 1a. Let k = 3; i.e., the user would like the objects to be partitioned into three clusters.

According to the algorithm above we arbitrarily choose three objects as the three initial cluster centers, where cluster centers are marked by a "+". Each objects is distributed to a cluster based on the cluster center to which it is the nearest. Such a distribution forms encircled by dotted curves as show in fig 1a.

Next, the cluster centers are updated. That is the mean value of each cluster is recalculated based on the current objects in the cluster. Using the new cluster centers, the objects are redistributed to the clusters based on which cluster center is the nearest. Such a redistribution forms new encircled by dashed curves, as shown in fig1b.

This process iterates, leading to fig 1c. The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as *iterative relocation.*

Eventually, no redistribution of the objects in any cluster occurs, and so the process terminates. The resulting cluster is returned by the clustering process.

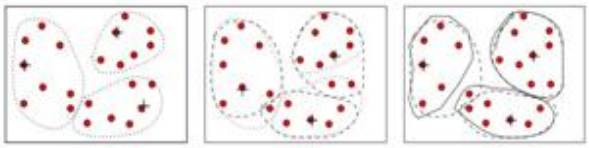

FIG 1: CLUSTERING OF A SET OF OBJECTS BASED ON K-MEANS METHOD

C.K-Means in WEKA

The learning algorithm k-Means 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.

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.(fig 2)

1) Euclidean distance

K-means cluster analysis supports various data such as quantitative, binary, nominal or ordinal, but do not support categorical data. Cluster analysis is based on measuring similarity between objects by computing the distance between each pair.

There are a number of methods are for computing distance in a multidimensional environment.

Distance is a well understood concept that has a number of simple properties.

1. Distance is always positive
2. Distance from point x to itself is always zero
3. Distance from point x to point y cannot be greater than the sum of the distance from x to some other point z and distance from z to y.
4. Distance from x to y is always the same as from y to x.

It is possible to assign weights to all attributes indicating their importance. There are number of distance measures such as Euclidean distance, Manhattan distance and Chebychev distance. But in this analysis Weka tool used Euclidean distance.

```
=== Run information ===

Scheme:     weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S
10
Relation:   oraldb
Instances:  424
Attributes: 10
            S.No
            Name
            Age
            Designation
            Smoking or Drinking
            White or red color patch on the tongue
            Difficulties in sudden swelling
            Wound in Mouth
            Bleeding
            class
Test mode:  evaluate on training data

=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1130.3460950653453
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                 Full Data          0          1
                            (424)      (229)      (195)
=============================================================================
S.No                        212.5     210.3537   215.0205
Name                    Paramasivam  Paramasivam  Rangasamy
Age                       42.7689    40.3493    45.6103
Designation               Labour     Labour     Labour
Smoking or Drinking       0.8373     0.7031     0.9949
White or red color patch on the tongue  0.316   0.131   0.5333
Difficulties in sudden swelling  0.5236   0.3886   0.6821
Wound in Mouth            0.3797     0.1266     0.6769
Bleeding                  0.1745     0.0961     0.2667
class                Moderate Cancer  Mild Cancer Moderate Cancer

Clustered Instances

0     229 ( 54%)
1     195 ( 46%)
```

FIG 2: KMEANS IN WEKA BASED ON CANCER SYMPTOMS

Euclidean distance of the difference vector is most commonly used to compute distances and has an intuitive appeal but the largest valued attribute may dominate the distance. It is therefore essential that the attributes are properly scaled.

Let the distance between two points x and y be D(x,y).

$$D(x,y) = (\sum(x_i-y_i)^2)^{1/2} \quad (2)$$

2) *Clustering of Disease symptoms*

The collected disease symptoms such as Smoking or Drining, White or red color patch on the tongue, Difficulties in sudden swelling, wound in mouth, Bleeding as raw data, supplied to Kmeans method is being carried out in weka using Euclidean distance method to measure cluster centroids. The result is obtained in iteration 4 after clustered. The centroid cluster points are measured based on the diseases symptoms. Found on the diseases symptoms in raw data, the Kmeans clustered two main clustering units. From the confusion matrix above we came to know that the district mainly impacted by Mild Cancer.

The above implementation algorithm yields results that the Kaligarayan palayam and surrounding villages in Erode Dist people affected by the Mild Cancer disease. We came to the conclusion of this result by confusion matrix of K-means algorithm

The K-means method clustered train the data up to 100% so the error rate completely reduced. The time taken to build the algorithm relatively too small.

IV. CONCLUSION

The K-means algorithm was implemented using Weka 3.6.4 data miner. It clustered into two major clusters units with the class variables. The clusters were varied in zigzag manner, slightly with each iterations and finally in the 4th iteration and finally maximum of its attributes belongs to Mild Cancer class. Found on that Kaligarayan palayam and surrounding villages in Erode Dist people affected by the Mild Cancer disease...

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 which 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 death level.

Through this research the problem of Cancer in Erode District come to light. It is a big social relevant problem. Pharmaceutical industries also can identify the location to develop their business by providing good medicine among people with service motto.

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

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 journals.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication