ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Near The Beginning of Non Small Cell Lung Cancer Avoidance in Human Way of Life Risk Factors Classification Using Clustering Algorithm in the R Environment

R.Kaviarasi
Department of Computer Applications,
University College of Engineering (BIT Campus)
Tiruchirappalli, India

Dr.A.Valarmathi
Department of Computer Applications,
University College of Engineering (BIT Campus)
Tiruchirappalli, India

Abstract: Cancer dominance is on the rise. The Lung Cancer is the Uncontrolled growth of abnormal cells that start off in one or both Lung Usually in the cell that Line the air Passages. The Two main types are Small Cell Lung Cancer and Non-Small Cell Lung Cancer. NSCLC is a complex disease which is characterized by the gathering of human behavior deviation for the duration of the human's lifetime. The prevention of NSCLC is an important and tedious task in medicine. The aim of this paper predicts the NSCLC using clustering algorithm. The attribute selection method is used to reduce the risk factors. The human routine life risk factors are classified into three levels. There are you can change, you cannot change, uncertain or unproven effects of lung cancer. The risk factors are classified using k-means clustering and hierarchical clustering; we identified the best performing of clustering methodology in the context of clinical outcome. Ultimate goal of this research work is to find out which type of clustering algorithm will be most suitable for analysis of Non Small Cell Lung Cancer data.

Key Words: NSCLC, Attribute selection, K-Means clustering, Hierarchical clustering, Validation measure.

I. INTRODUCTION

Lung cancer starts when cells of the lung become abnormal and begin to grow out of control. As more cancer cells develop, they can form into a tumor and spread to other areas of the body. There are two main types of lung cancer. About 80% to 85% of lung cancers are non-small cell lung cancer (NSCLC), about 10% to 15% are small cell lung cancer (SCLC) .There are subtypes of NSCLC, which start from different types of lung cells. But they are grouped together as NSCLC because the approach to treatment and prognosis are often similar. Lung cancer (both small cell and non-small cell) is the second most common cancer in both men and women.

The American Cancer Society's estimates for lung cancer in the United States for 2017 are about 222,500 new cases of lung cancer (116,990 in men and 105,510 in women), about 155,870 deaths from lung cancer (84,590 in men and 71,280 in women) Lung cancer is by far the leading cause of cancer death among both men and women about 1 out of 4 cancer deaths are from lung cancer. Lung cancer mainly occurs in older people. About 2 out of 3 people diagnosed with lung cancer are 65 or older, while less than 2% are younger than 45. The average age at the time of diagnosis is about 70. Overall, the chance that a man will develop lung cancer in his lifetime is about 1 in 14; for a woman, the risk is about 1 in 17. For smokers, and family history of lung cancer the risk is much higher [1], while for non-smokers of environment factors the risk is lower and uncertain effects of lung cancer the risk in average level.

A. *Non-small cell lung cancer risk factors*

A risk factor is anything that affects a person's chance of getting a disease such as cancer. Different cancers have different risk factors [2]. The risk factors are categorized in three levels. There are you can change, you cannot change, and Factors with uncertain or unproven effects on lung cancer. Some risk factors are Tobacco smoke, Secondhand smoke, Exposure to radon, Exposure to asbestos, Exposure to other cancer-causing agent in the workplace, Arsenic in drinking water, certain dietary supplements, can be changed. Others, previous radiation therapy to the lungs, Air pollution, like a person's age or family history, can't be changed. But having a risk factor, or even several, does not mean that you will get the disease. And some people who get the disease may have few or no known risk factor.

II. REVIEW OF THE LITERATURE

Dharmarajan et, al.,[1] proposed the algorithms of k-Means and Farthest First are have been implemented his work. The performance of the partitioning based algorithms were analyzed using the only selected three attributes from the total number of attributes of input dataset. It is very evident from the results that the computational complexity of the k-Means algorithm with LC.arff dataset is better than that of Farthest First algorithm for both of the dataset. The k-Means algorithm is efficient for lung cancer dataset with arff format. It is well suited for requirement clustering of cancer related medical applications.

Arpit Bansal et, al.,[2] we have proposed technique for a modification in K-Means Clustering Algorithm. Here in this proposed modification, the K-Means clustering will vanish off the two major drawbacks of K-Means clustering that are accuracy level and calculation time consumed in clustering the dataset. Although when we use small datasets these two factors accuracy level and calculation time may not matter much but when we use large datasets that have trillions of records, then little dispersion in accuracy level will matter a lot and can lead to a disastrous situation, if not handled properly. So in last considering the whole of the situation, it can be said that this proposed modification can be more extended to achieve the full accuracy level up to 100%, with very little time and with more quality clusters.

Nathan Lawlor et, al.,[3] proposed to most popular and well-understood two clustering algorithms were chosen for comparative analysis because of their low computational intensiveness in comparison to that of model-based algorithms. This allowed for rapid testing of multiple gene ranking and selection options when using our 14 gene expression datasets. However, model-based clustering via parsimonious Gaussian mixture models and linear mixed-effects models are also beneficial options for clustering of tissue samples and gene profiles even when dataset size is small. Such methods have been demonstrated to perform well with high-dimensional gene expression data and produce better clustering when compared to conventional clustering methods.

Prabhakar Chalise et,al.,[4] We have proposed Cluster analysis aims to highlight meaningful patterns or groups inherent in the data that will be helpful in identifying the subtypes of the diseases. Several types of clustering algorithms have been proposed that use several assays of molecular variation of cells most of which are designed for one type of data at a time. Such methods have been successfully implemented in many disease classification studies. A few comprehensive clustering methods have also been proposed and successfully implemented in some studies. In this paper, brief review of those methods has been presented.

III. METHODOLOGY

A. Data collections

The non small cell lung cancer risk factors dataset was taken from the UCI machine learning repository and it is made up of 100 raw attribute from which various attributes were published by various researchers. These attributes are very essential in the identification. The dataset has 1000 instances. The 12 attributes considered in this research work are stated below. The description of Non-Small cell lung cancer dataset is tabulated in below Table 1.

Table 1. Lung cancer data set attributes and characteristics

S.NO	Attributes	Description
1.	Age	>30, 3 30<x>50, 4 50<x, 5
2.	Smoking	Yes=1, No=0
3.	Second hand smoking	Yes=1, No=0
4.	Exposure to radon	Yes=1, No=0
5	Exposure to asbestos,	Yes=1, No=0
6.	Exposure to other cancer-causing agent in the workplace	Yes=1, No=0
7.	Arsenic in drinking water,	Yes=1, No=0
8.	certain dietary supplements,	Yes=1, No=0
9.	previous radiation therapy to the lungs	Yes=1, No=0
10.	Air pollution	Yes=1, No=0
11.	Family history of lung cancer	Yes=1, No=0
12.	Talc and Talcum powder	Yes=1, No=0
13.	Smoking majirina	Yes=1, No=0

B. R Language

R is an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing. It is used to do exploitation and investigation of variety of data in the datasets. Various plots can be developed using R language and it does develop the software development activities in data mining in various areas. It is an efficient, extensible and more than enough environment for various statistical computations and graphics. the key in features of R language is that it supports user-created R packages [3] and we can import data containing variety of file formats such as CSV (Comma Separated Values), XML(), binary files. R language has been a variety of data structures. Its contained arrays, data frames, vectors, metrics and lists. There are different packages are offered for R and we can use the package on every occasion we are in need by using library (package name) command. There are different interfaces are presented for R language. Along with them R Studio is normally used an interface.

IV. METHODS AND WORK FLOW

A. Prevention of NSCLC using Clustering algorithms

In this section, we describe the components and workflow of our R-package cluster. Our integrative R-package contains. Two ways to determine the clustering number. There are fixed and gap statistic [3][4]. This paper to apply clustering algorithm to Euclidian distance values are measured in this paper. Two types are clustering algorithms are applied in this paper. There are Hierarchical clustering and k-means clustering. The clustering algorithm task to calculate the average in each sample cluster. And then a gathering to correlate sample clusters with clinical outcome.

B. Hierarchical clustering

The NSCLC data sets are imported in the R studio. The data set NSCLC, we can run the distance matrix with hclust, and plot a dendrogram that displays a hierarchical relationship among the NSCLC causes of risk factors [5][6]. The dendrogram shows that family history of lung cancer and smoking are classified as close relatives as expected. The hclust function in R uses the complete linkage method for hierarchical clustering by default. This particular clustering method defines the cluster distance between two clusters to be the maximum distance between their individual components. At every stage of the clustering process, the two nearest clusters are merged into a new cluster. The process is repeated until the whole data set is agglomerated into one single cluster. For a data set with 1,000 elements, it takes hclust about 1 minutes to finish the job. The pvclust() function in the pvclust package provides p-values for hierarchical clustering based on multiscale bootstrap resampling. Clusters that are highly supported by the NSCLC data will have large p values.

The algorithm works as follows:

Step 1: Put each data point in its own cluster.
Step 2: Identify the closest two clusters and combine them into one cluster.

Step 3: Repeat the above step till all the data points are in a single cluster.

C. K-Means Clustering

K Means Clustering is an unsupervised learning algorithm that tries to cluster data based on their similarity. Unsupervised learning means that there is no outcome to be predicted, and the algorithm just tries to find patterns in the data. In k means clustering, we have the specify the number of clusters we want the data to be grouped into. The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster.

Then, the algorithm iterates through two steps:

Step 1: Reassign data points to the cluster whose centroid is closest.
Step 2: Calculate new centroid of each cluster.

These two steps are repeated till the within cluster variation cannot be reduced any further. The within cluster variation [7][8] is calculated as the sum of the Euclidean distance between the data points and their respective cluster centroids. In this data set we observe the composition of different NSCLC causes. Given a set of observations (x1,x2,.,xn)(x1,x2,.,xn), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into (k≤n) S={S1,S2,.,Sk} so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find,

$$\text{argmin} \sum_{i=1}^{k} \sum_{X \in S}^{n} \|xj - \mu i\|2 \qquad (1)$$

Where μi is the mean of points in Si. The clustering optimization problem is solved with the function kmeans in R. The “Eq. (1)” is calculate the average value of the parameter k. If we looks at the percentage of variance explained as a function of the number of clusters. One should choose a number of clusters so that adding another cluster doesn’t give much better modeling of the data. If one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information, but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the elbow criterion. The elbow is difference between the WSS [9][10] and number of cluster values.

D. Validating cluster solutions

The fpc package provides a mechanism for comparing the similarity of two cluster solutions using a variety of validation criteria Hubert's gamma coefficient [11][12]. Where d is a distance matrix among the objects of two types of classification results from two different clustering of the same NSCLC data.

V. RESULTS AND DISCUSSION

The work is implemented with R studio Environment. R contains the different clustering algorithms that are used to form clusters. The NSCLC data sets are collected from human way of life risk factors. The risk factors are classified in number of clustering. The WSS is measured in the data sets. The results are performed in sum of square values.

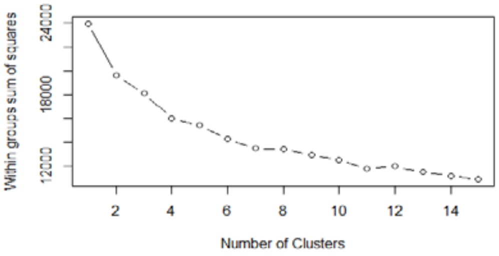

Figure1. Plot showing the within groups of square (WSS) against number of clusters (k)

The hierarchical clustering is applied in the non small cell lung cancer dataset. The data set are created in CSV file. The cluster dendrogram of causes of risk factors samples from the human way of life.

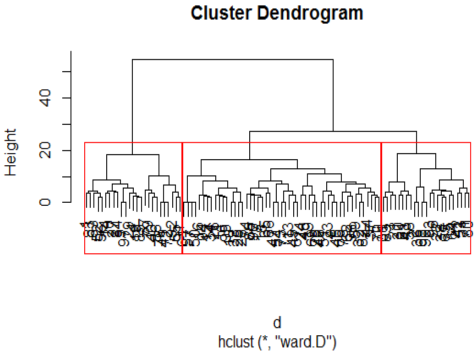

Figure 2. The dendrogram of causes of risk factors samples from the human way of life NSCLC dataset generated using hierarchical clustering. The top 1000 data is selected. NSCLC causes are clustered using Euclidean distance and Ward.D linkage. The risk factors were split into three clusters. The samples in the red boxes represent the clusters.

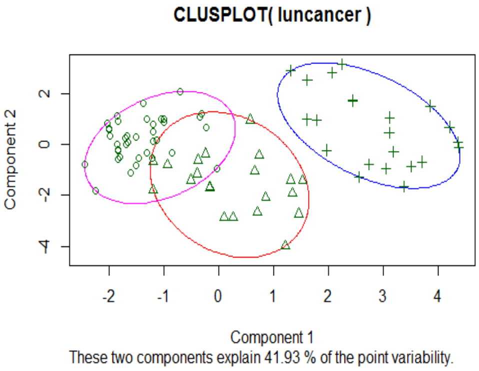

Figure 3. The CLUSPLOT value of luncancer file compared the component 1 and component 2. These two components 41.93% of the point variability is measured. The results produced three clustering values.

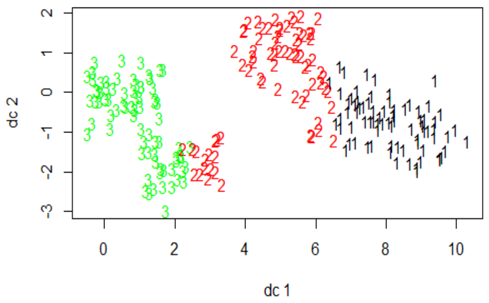

Figure 4. The K-Means clustering centroid plot against first two discriminate function of NSCLC. The dc1 is measured the risk factors of NSCLC. The dc2 is measured the fit values of NSCLC. Three types of cluster values are produced. The block group is can change the risk factors. The red group is cannot change the risk factors. The green in measured unproven effects of lung data.

VI. CONCLUSION

Cluster analysis aims to highlight meaningful patterns or group distance inherent in the data that will be helpful in identifying the high risk factors groups of the NSCLC disease. Two types of clustering are applied in this paper. The hierarchical clustering is produced dendrogram results are produced using Euclidean distance and Ward.D linkage. The K-Means clustering are produced WSS values against number of cluster K values. Then finally K-Means clustering centroid plot against two discriminant function of NSCLC. This paper finally validate to the two type of clustering fit values. The validation measurement result is helped to the distance are measured in two clustering values. The results are helped to at the beginning of NSCLC prevention through human way of life handled risk factors characteristics. This paper work is very helped to the cancer research center and hospitals to prevent the NSCLC.

VII. REFERENCES

[1] A.Dharamarajan, T. Velmurugan to “Lung cancer data analysis by k-means and farthest first clustering algorithms” (IJST), Vol. 8, July 2015.

[2] Arpit Bansal, Mayur Sharma, ”Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining”, IJCA (0975 – 8887) Vol. 157 – No 6, January 2017.

[3] Nathan Lawlor, Alec Fabbri to “Multiclust: an r-package for identifying biologically relevant clusters in cancer transcriptome profiles” Cancer Informatics, 2016.

[4] Prabhakar Chalise, Devin C.Koestler, Milan Bimali, Qing Yu, Brooke L.Fridley to “Integrative clustering methods for high –dimensional molecular data” , Transl Cancer Res 2014.

[5] V.Krishnaiah , Dr.G.Narsimha, Dr.N.Subhash Chandra to “Diagnosis of lung cancer prediction system using data mining classification techniques” (IJCSIT), Vol. 4 (1) , 2013, 39 – 45

[6] Dr. D. P. Shukla, Shamsher Bahadur Patel, Ashish Kumar Sen to “A literature review in health informatics using data mining techniques” (IJSHRE), Vol. 2, Issue 2 , pp.123-129,2014

[7] Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti, Md Zamilur Rahman, Farzana ahmed to “Early detection of lung cancer risk using data mining”(APJCP), Vol.14, pp.595-598, 2013.

[8] R.J. Noel Blessy1, K.Mohamed Amanullah to “Oral cancer detection using apriori algorithm”(IJARCCE), Vol.3,Issue 7, pp. 7376-7379,2014

[9] Divya Tomar and Sonali Agarwal to “A survey on data mining approaches for healthcare” (IJBSBT), Vol.5, No.5, pp.241-246, 2013.

[10] Zehra Karapinar Senturk, Resul Kara - “Breast cancer diagnosis via data mining: performance analysis of seven different algorithms” - An International Journal (CSEIJ), Vol. 4, No. 1, February 2014

[11] V.Krishnaiah, “Diagnosis of lung cancer prediction system using data mining classification techniques” (IJCSIT), Vol. 4, pp. 39 – 45, 2013,

[12] Charles Edeki, “Comparative study of data mining and statistical learning techniques for prediction of cancer survivability”, (MJSS) Vol. 3, , ISSN: 2039-9340, November 2012.