Journal for Research | Volume 03| Issue 05 | July 2017
ISSN: 2395-7549

A Study on Hierarchical Clustering for Identifying Asthma Endotypes

S. Poorani
Assistant Professor
Department of Computer Technology
Kongu Engineering College, Perundurai, India

Dr. P. Balasubramanie
Professor
Department of Computer Technology
Kongu Engineering College, Perundurai, India

N. Sasipriyaa
Assistant Professor
Department of Computer Science and Engineering
Kongu Engineering College, Perundurai, India

Abstract

Large multitudes of medical data are attained due to advancement in recent technology. These large volumes comprise valuable information which helps for diagnosing diseases. Data mining techniques can be used to extract useful patterns from these mass data. It provides a user- oriented approach to the novel and hidden patterns in the data. One of the major challenges in medical domain is the extraction of comprehensible knowledge from medical diagnosis data. Healthcare system needs an automated tool that identifies and disseminates relevant healthcare information. Asthma is one of the dominant disease all over the world. Identifying subtypes of the disease will help in the treatment and prevention approaches. In this paper a clustering model is proposed based on observable symptoms of asthma. The hierarchical clustering technique is used to group the samples based on important characteristics of asthma.

Keywords: Data Mining, Hierarchical Clustering, Asthma, Health Care

I. INTRODUCTION

There are 300 million asthmatics worldwide with 1/10th of those living in India. A recent review about epidemiological studies showed that the asthma prevalence among children was 7.24%[1].

Recently the Global Initiative for Asthma (GINA) defines asthma as a heterogeneous disease usually characterized by chronic airway inflammation. It is defined by the presence of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable airflow obstruction [2].

It is obvious that the asthma is not a solitary disease, but a set of symptoms that includes a number of disease subtypes with similar observable clinical characteristics and it affects all ages of people [3,1].

Data mining is very helpful in mining information and patterns in various fields in the form of association rule mining, clustering and classification techniques. Clustering techniques intend to group the objects based on their similarities. Different clustering techniques can be used for finding similarity between objects based on the nature and size of the data. Hierarchical clustering method is the one which build a hierarchy of clusters.

Thus we made a study to investigate new factors that may associate with asthma disease, where hierarchical clustering technique can be used to group(endotypes) patients with similar symptoms and habits, which helps to create awareness among the people. In addition, it provides more information to the doctors to provide appropriate treatment for the patients.

II. LITERATURE SURVEY

Existing studies depicts that several statistical methods and data mining approaches have been applied into medical datasets for finding important risk factors that affect relevant diseases. This section reviews knowledge discovery using data mining techniques on real datasets of asthma patients.

Juliet Rani Rajan and Chilambu Chelvan[4] proposed a data mining model that can identify the different groups of asthma patients. It can help the doctors to make treatment related decisions at the earlier stage. The predictive factors were captured from the patients in the form of questionnaires.

J. Cathrin Princy and K. Sivaranjani[5] proposed a system, which uses SVM and MLP for predicting asthma attacks. The experimental evaluation results showed a 98.5% of accuracy in asthma attack prediction.

Somayeh Akhavan Darabi[6] developed a case based reasoning system(by using random forest method) which uses the questionnaire to collect asthma related variables without laboratory details to do asthma diagnosis. This system identifies that most important variables of asthma disease in Iran country are symptoms heperresponsivity, frequency of cough, cough.

Behrouz Alizadeh et al[7] developed a model based on neural network for detecting the presence or absence of asthma in which the effective features are used as the input.

Matea Deliu et al[8] provides an overview of different clustering methods to understand the spectrum of asthma syndrome.

Wendy C. Moore et al[9] applied an unsupervised modeling method to the SARP dataset which identifies distinct groups or clusters of people with asthma. Five distinct clusters of asthma phenotypes were identified that differ in lung function, age of asthma onset and duration, atopy, sex, symptoms, medication use, and health care utilization.

Pranab Haldar et al[10] suggest multivariate techniques in the classification of asthma populations.

III. METHODOLOGY

The dataset of asthma population sample during enrolment at the hospital was considered for this study. A baseline, data regarding demographics, home environment characteristics, asthma symptoms, allergy history, and relevant family history of the patient, food habits, and other needed data were captured from the patients by means of questionnaires. After gathering the relevant data hierarchical clustering algorithm was applied on the data to identify the subgroups. A total of 35 records were gathered. The dataset include only the asthmatic patients records.

The following are some asthma features that were included in the questionnaire

- Age
- Sex
- Family history of asthma
- Wheezing
- Shortness of breath
- Chest tightness
- Running nose

The collected reports from hospitals were written and typed in different formats which cause the data inconsistent. Some reports even contained duplicates data or missing values. We pre-processed these reports by removing the duplicates data and providing the missing values according to the past recorded data. Finally 31 records were used for clustering process.

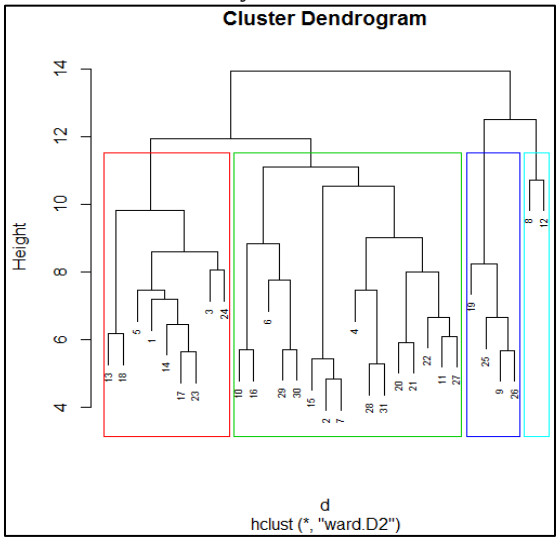

Fig. 1: Cluster dendrogram

The results of hierarchical clustering are usually represented in a dendrogram. The R package was used to generate the dendrogram by using hierarchical clustering shown in fig 1. The rectangles in the dendrogram are considered as four clusters.

IV. RESULTS AND DISCUSSION

Using the Ward's method, a dendrogram was generated (Figure 1). In that, four clusters were identified. The cluster-1 includes 9 samples, cluster-2 includes 16 samples. The clusters three and four were small clusters (n = 4 and n=2 respectively). All the four clusters differ significantly by age. In cluster 1, 2 & 3 no one is having diabetics, but in cluster 4 all are having diabetics. Wheezing and cough are high in all the clusters, but running nose, chest tightness, Heredity and other variables differ in different clusters.

A. Cluster-1

The cluster-1includes both male and female, all are non-vegetarian, no stress, no exposure to AC, all of them having wheezing and cough, few have chest tightness, half of the samples have running nose, no one has the disease in heredity, no shivering, no blood pressure, no diabetics, but most of them take tablets and not using inhaler. All of them are married and age falls between50 to 80.

B. Cluster-2

The cluster-2 includes both male and female people, only few are vegetarians, not work under AC, all of them having wheezing but few people do not have cough, only few people have shivering, chest tightness and stress, no diabetics, only one having Blood pressure, half of samples have running nose and heredity, most of the people use inhaler and take tablets. This includes both married and unmarried in age range from 6 to 75.

Cluster 1 & 2 significantly varies in age, heredity, and usage of inhalers.

C. Cluster-3

The cluster-3 includes both male and female people, vegetarians and non-vegetarians, all of them having wheezing, cough, chest tightness, most of them having stress, no diabetics, only one having Blood pressure, only one have shivering, half of samples have running nose and heredity, most of the people take tablets and only one use inhaler. This includes both married and unmarried in age range from 27 to 60.

D. Cluster-4

The cluster-4 all of the patients have diabetics and blood pressure. All of them use inhaler and take tablets, married and age greater than 50. The remaining variables are equally distributed.

Clusters 3 & 4 significantly varies in age, diabetics and blood pressure.

In all the clusters the wheezing exists.

In all the clusters the people having stress takes tablets as well as inhalers. But the people who do not take inhaler and tablet has no stress.

V. CONCLUSION

Thus, the hierarchical clustering can be used in clustering small dataset. The four clusters created here, show the significance of different asthma characteristics and related variables in identifying endotypes and risk factors of asthma. This work can be extended for large populations or big data. Advanced and more suitable clustering algorithms can be applied to get better accuracy and performance.

REFERENCES

[1] India has 10% of world's asthma patients: Survey. News18.com. May 3, 2016
[2] From the Global Strategy for Asthma Management and Prevention. http://www.ginasthma.org/ . Accessed 1st April 2017.
[3] Wenzel SE. Asthma: defining of the persistent adult phenotypes. Lancet. 2006;368(9537):804–813. doi: 10.1016/S0140-6736(06)69290
[4] Juliet Rani Rajan and Chilambu Chelvan. A Prognostic system for early diagnosis of pediatric lung disease using artificial intelligence. November 2016.
[5] J. Cathrin Princy and K. Sivaranjani. Asthma Prediction Using Classification Technique. I J C T A, 9(27), 2016, pp. 415-421.
[6] Somayeh Akhavan Darabi. Case-Based-Reasoning System for Feature selection and Diagnosing Disease: Case Study: Asthma. Innovative Systems Design and Engineering. 2014.
[7] Behrouz Alizadeh, Reza Safdari, Maryam Zolnoori, Azadeh Bashiri. Developing an Intelligent System for Diagnosis of Asthma Based on Artificial Neural Network. Acta Inform Med. July 2015.
[8] Matea Deliu, Matthew Sperrin, Danielle Belgrave, and Adnan Custovic. Identification of Asthma Subtypes Using Clustering Methodologies. Pulmonary Therapy. June 2016.
[9] Wendy C. Moore, Deborah A. Meyers, Sally E. Wenzel, W. Gerald Teague, Huashi Li, Xingnan Li, Ralph D'Agostino, Jr., Mario Castro Douglas Curran-Everett, Anne M. Fitzpatrick, Benjamin Gaston, Nizar N. Jarjour, Ronald Sorkness, William J. Calhoun, Kian Fan Chung, Suzy A. A. Comhair, Raed A. Dweik, Elliot Israel, Stephen P. Peters, William W. Busse, Serpil C. Erzurum, Eugene R. Bleecker. Identification of Asthma Phenotypes Using Cluster Analysis in the Severe Asthma Research Program. American Journal of Respiratory and Critical Care Medicine 2009;179:A2522. November 2009.
[10] Pranab Haldar, Ian D. Pavord, Dominic E. Shaw, Michael A. Berry, Michael Thomas, Christopher E. Brightling, Andrew J. Wardlaw, and Ruth H. Green. Cluster Analysis and Clinical Asthma Phenotypes. PMC. May 2008.