Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/103107, November 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Aggregated K Means Clustering and Decision Tree Algorithm for Spirometry Data

K. Rohini and G. Suseendran

Department of Information and Technology, School of Computing Sciences, Vels University, P.V. Vaithiyalingam Road, Pallavaram, Chennai - 600117, Tamil Nadu, India; rrohiniphd@gmail.com, suseendar_1234@yahoo.co.in

Abstract

Objectives: The present research work generally focuses on predicting diseases from the lung disease test by using data mining techniques for spirometry data. **Methods/Statistical Analysis**: Spirometry is used to create baseline lung function, check out dyspnea, disclose pulmonary disease, watching effects of therapies used to treat respiratory disease, calculate respiratory impairment, evaluate operative risk, and performs surveillance for occupational-relevant lung diseases. Pulmonary function tests are used to find out lung capacity, based on which the many of the lung diseases can be identified. In this research work, a combination of k-means clustering algorithm and Decision tree algorithm was developed. From the results investigation, it is known that the proposed aggregated k-means algorithm and decision tree algorithm for spirometry data is better which compared to other algorithms such as Genetic algorithm, classifier training algorithm, and neural network based classification algorithms. **Findings**: Existing algorithms are unable to handle noisy data and also with Failure occurrence for a nonlinear data set. It should not classify the data set based on their input attributes. Prediction is not possible for existing system. **Applications/Improvement:** Spirometry data which is used to predict the lung capacity using Aggregated K-means and Decision tree algorithm. Our proposed approach is evaluated for each dataset accordingly.

Keywords: Decision Tree, Pulmonary Function Test Means, Spirometry Data

1. Introduction

Nowadays Lung diseases are a major serious disease which is affected by human health vigorously. Spirometry is an effective tool for finding patients diseases, using pulmonary function test. Pulmonary function test is an only test for analyzing the patient's disorders in an effective manner. Because it has better equipment materials and well experts are there in these tests. Here our research work is based on their spirometry data how their patient's diseases are predicted using K means clustering and decision tree algorithms. Most of the researchers doing their research work in data mining, which provide better results compared to other areas in medical field[1]. Clustering is one of the familiar common unproven data mining approaches that are used to explore the hidden structures enclosed in a dataset. Cluster analysis aim is to standardize a collection of patterns into clusters based on their affinity. The goal of clustering is to provide users to identify different groups in a dataset, and to reduce the amount of data by classifying a similar data items together. Data clustering is most widely used tools in data mining. Clustering technologies allow multiple servers to work in union to present the presence of a single computing environment[2]. It follows variety of steps in Figure 1.

*Author for correspondence

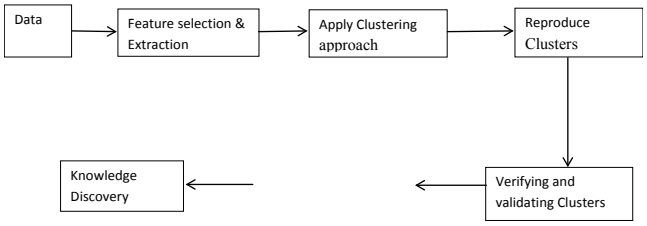

Figure 1. Clustering Operation.

First step is the process of finding the most effective subset of the original features to use in clustering, feature extraction is input features are transformed and to produce salient output feature. Cluster complexity is increased due to the improper selection of the features. Second step is mostly important approach to select the algorithm correctly by applying domain knowledge. Basically many of the algorithms are based upon the various input parameters, like number of clusters, optimization/construction criterion, abandon condition, proximity measure etc. but it is impractical to develop a generalized framework of clustering methods for the application in the different areas like social, scientific, and medical fields. The final step of clustering procedure manages with the representation of the clusters. The k-Means clustering algorithm is mostly used, simplest unsupervised learning algorithms that report the well-known clustering problem[3]. In this technique follows a simple and effective method to organize a given data set through a certain number of clusters. The k-Means algorithm can be run many times to reduce the complexity of grouping data[4]. It is a simple algorithm that has been modified to most of the problem areas and it is a well-born candidate to work for a randomly generated data points[5].

2. Proposed Methodology

The primary objective of this analysis is to predict the diseases from the medical data sets, using the test of Pulmonary Function Test. Many researchers are interested to do their research in this domain. Proposed methodology is aggregated K means and decision tree algorithm for spirometry data, which is used to identify, follow, and managing their patients with lung disorders. This type of test also defines lung movements but it needs more practical equipment and experts were available for pulmonary function laboratory. It is used to calculating lung operation, especially the amount (volume) and/or speed (flow) of air that can be inhaled and exhaled. Spirometer is a most important tool mainly used for creating pneumotachographs, which are more helpful for assessing conditions like cystic, asthma, fibrosis, cystic. Many of the spirometers view the following graphs, called spirograms. The medical dataset contained in all the information aggregated during a survey it needs to be analyzed[6]. Studying how to interpret the results is a major part of the survey process. It is collection of interconnected data with user defined parameters.

Pulmonary Function Testing (PFT)) are most helpful in calculated the general type of lung disorder and determines the severity. Heart disorders may also effect of breath and other symptoms that may suggest a lung disorder and because of lung disorders[7].

2.1 Aggregated K Means and Decision Tree Algorithm

K means algorithm is like a dividing based clustering algorithm, is to classify the given data objects into n different clusters over the iterative, converging to a local minimum. The results generated clusters are minimized and independent[8].

2.2 Algorithm for K Means Clustering

Input: C= {c1, c2, c3…..cn}, cluster sets,
D= {d1, d2, Dn} data sets
Output: find mean value μ_i
Begin
Choose any cluster from Data set D
Repeat
While (C_j € D)
Assign Z as a Cluster centric
Select similar data
Compute Mean Value μ_i
End

Flow Chart of K Means Clustering Algorithm

This algorithm is a mathematical, unsupervised, non-deterministic, iterative technique. It is very fast and understood by each and every one for most of the practical applications. This method is verified to be a very effective way that can generate good clustering results. It is very much suited for generating globular clusters in Figure 2.

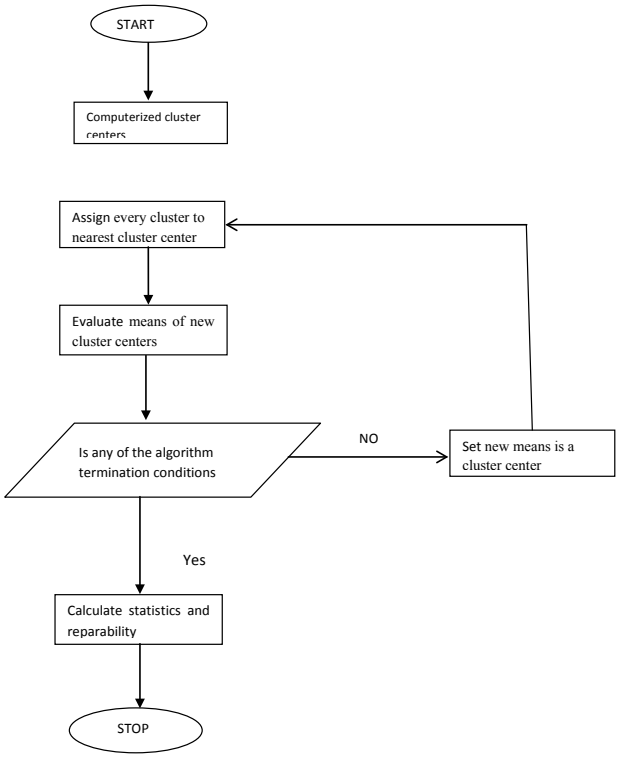

Figure 2. Flow chart of K means clustering algorithm.

2.3 Decision Tree Algorithm

Decision trees are combined of computational and mathematical techniques to aid the representation, generalization and categorization of a given set of data. A Decision tree is a format which contains a root node, branches, and leaf nodes. Each internal node denoted as check on associate degree attribute, every branch denoted as the end result of a check and every leaf node denoted as a category label[9]. The topmost node within the tree is called as root node. The main goal is to produce a model that predicts the value of a required variable based upon many input variables the decision tree model also uses the prediction based rules classification. The known label of test data is compared along with the classified result. Accuracy rate is calculated based on the percentage of test set samples[10].

Algorithm for Decision Tree

Step 1: The leaflet is labeled with the same class if the instances belong to the same class.

Step 2: For each parameters, the potential information will be evaluated and the gain in information will be taken from the test on the parameter.

Step 3: Finally the best parameter will be selected based on the present selection parameter.

Input: Attributes (a1, a2, a3…an)

Output: Predicted value Pv

Begin

Where

R – Root, B-Branches, L_f - Leaf nodes

Select each attribute (A_j)

Calculate potential information P_i

Find best attribute based on the prediction

End

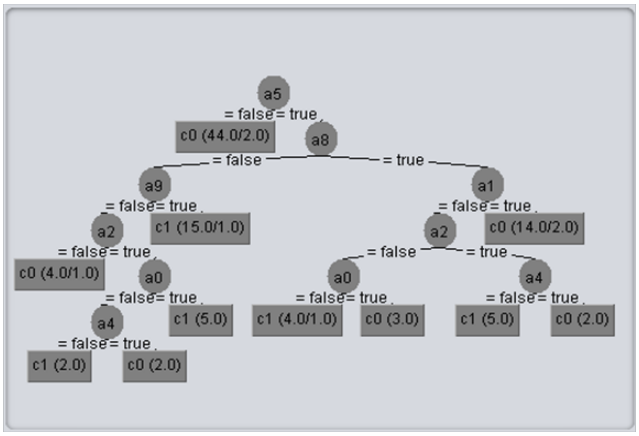

Figure 3. Decision trees of spirometry data.

It states that proposed decision tree for spirometry data. It predicts the lung Diseases based on their pulmonary function test[11]. Clusters are grouped based on their similar characteristics of their clusters. Here root nodes are denoted as a5, a8, a9, a1, a2, a0, a4 etc. Decision trees are used to predict the values for using true/false condition[12]. It is implemented using WEKA tool; it predicted the diseases based on their decision tree algorithm in Figure 3.

3. Experimental Results

The project work executed by WEKA software tool. It contains the variety of clustering algorithms that are used to figure out clusters. WEKA software is a combination of open source Natural language algorithms mainly used for pre-possessing, classifieds, clustering, and association rule. WEKA Tool is based upon Java for data mining. Data's are normally described by flat text files. It also includes different data files such as, "arff","csv" file formats. Performance can be evaluated by aggregated of K means and decision tree algorithm.

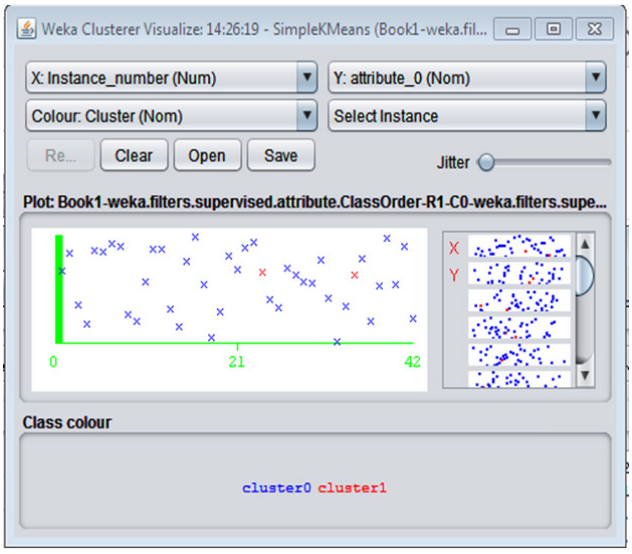

Figure 4. K means algorithm using training set for spirometry data.

The categorization of spirometry data for different attributes such as FVC, FEC, etc., the lung cancer dataset are processed for aggregated clustering algorithms such as k-Means and Decision tree clustering[13]. The above attributes are selected based on the datasets. The k-Means algorithm is mostly used for grouping a data. Existing algorithm drawback was attributes are not categorized, so prediction also not possible it's shown in Figure 4.

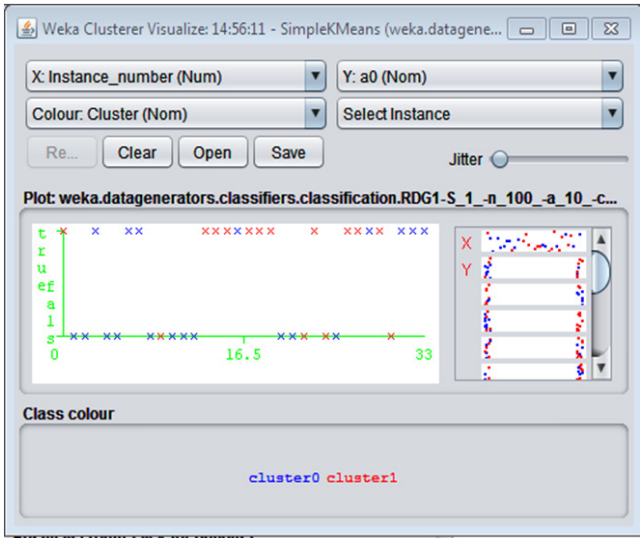

Figure 5. K means algorithm using percentage set for Spirometry data.

In this K means algorithm using percentage filter for spirometry data. Removal of noise is easy for applying those filters[14]. Normally the spirometer can measure by Forced vital capacity (FVC), Forced Expiratory volume one (FEV1). Based on Prediction values only spirometer results are measured[15]. Measuring Performance depends on patients Air quality, smoking habit, allergy details, weather factors etc. The fitted curve coefficients and predicted values for FVC, FEV1, and FEV1% are some of the inputs to the MLPNN (Multilayer perception neural network). Distinct MLP structures were tested. It is shown in Figure 5.

It states that the preprocessed data set of spirometry data, which contains different functional test data of south Indian ethnic group for both male and female. Raw spirometry data was preprocessed and clustered into instance groups which are stated above. Where Instance-I, Instance-II, Instance-III denoted as a Cluster groups of k means algorithm. Different cluster values are shown in Table 1 represented as Instance-I, Instance-II, Instance-III shown in Table 1.

Table 1. Representing data for Spirometry data

Attributes	Instance-I	Instance-II	Instance-III
FVC	1	10	1
FEV1	35	55	26
FEV1/FVC	157	157	154
PEF	50	50	44
FEF2575	1.47	1.47	0.96
FEF25	1.27	1.27	2.55
FEF25	52	52	38
FEF75	0.7	0.7	0.86
FEV3	2.31	2.31	2.13
FET	49	49	42
FIVC	100	100	89.6
FIV1	74.3	74.3	82.5
FIV1/FIVC	107	90	107

3.1 Performance Analysis Graph

The performance analysis graph states that the comparison between the training data and the test data. The training data (X axis plotted graph) was not applied with any preprocessing filters whereas the test data (Y axis plotted graph) was applied with unsupervised instance

filters namely removes percentage. Then this filtered data was applied with the cluster mode called the supplied test set using k-means clustering algorithm. Performance and Accuracy can be improved by applying filters and to detect the errors using this proposed approach shown in Figure 6.

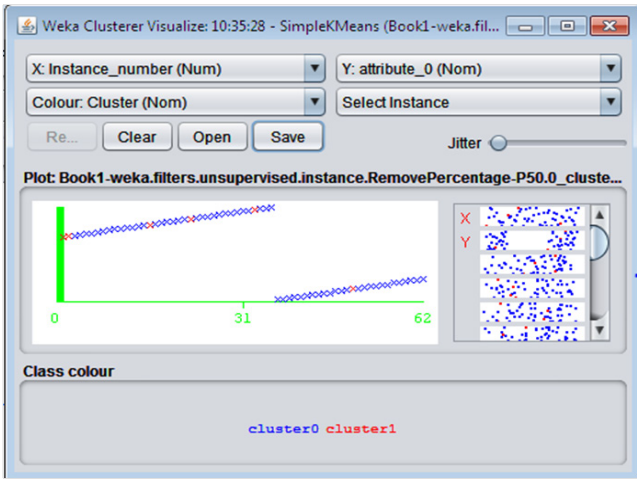

Figure 6. Comparison between training data and test data.

4. Conclusion

This research work is done for effective analyzed their spirometry data using aggregated k means and decision tree algorithm. Our proposed method uses different Input data sets for specifying the spirometry data. Numerous attributes such as FVC, FEV1, and FEV1/FVC, etc., were used for obtaining different Instance values. It uses Instance filters and preprocessing filters to remove their noisy data and it predicts the patient's lung diseases are effectively. Test data was surveyed about south Indian ethni group for male and female candidate. In future work GM-DBSCAN (Gaussian Means-Density Based Spatial Clustering of Application with Noise) algorithm was proposed for predicting different types of diseases based on different variety of parameters. Commonly DBSCAN is the most familiar algorithm for analyzing the clusters. But it failed with choosing parameters. So Gaussian Means is used to calculating the value of DBSCAN's parameters. This is used to produce a better quality of clustering results. Spirometry data is mostly used for medical related applications.

5. References

1. Purusothaman G, Krishnakumari P. A Survey of Data Mining Techniques on Risk Prediction: Heart Disease. Indian Journal of Science and Technology. 2015 June; 8(12):1–5.
2. Soni N, Ganatra A. Categorization of Several Clustering Algorithms from Different Perspective: A Review. International Journal of Advanced Research in Computer Science and Software Engineering. 2012 Aug; 2(8):1–6.
3. Na S, Xumin L, Yong G. Research on k-means Clustering Algorithm. Third International Symposium on Intelligent Information Technology and Security Informatics. 2010; 978-0-7695-4020-7/10.
4. Yadav J, Sharma M. A Review of K-mean Algorithm. International Journal of Engineering Trends and Technology (IJETT). 2013 July; 4(7):2972–75.
5. Venkatesan E, Velmurugan T. Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Indian Journal of Science and Technology. 2015 Nov; 8(29):1–8.
6. Srimani PK, Koti MS. Knowledge Discovery in Medical Data by using Rough Set Rule Induction Algorithms. Indian Journal of Science and Technology. 2014 July; 7(7):905–15.
7. Vijayarani S, Sudha S. An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples. Indian Journal of Science and Technology. 2015 Aug; 8(17):1–8.
8. Kanungo T, David M, Netanyahu NS, Christine D, Piatko, Silverman R, Wu AJ. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. IEEE transactions on pattern analysis and machine intelligence. 2002 July; 24(7):881–92.
9. Chirumamilla V, Sruthi BT, Velpula S, Sunkara I. A Novel approach to predict Student Placement Chance with Decision Tree Induction. International Journal of Science & Technology. 2014; 7(1):78–88.
10. Dharmarajan A, Velmurugan T. Lung Cancer Data Analysis by k-means and Farthest First Clustering Algorithms. Indian Journal of Science and Technology. 2015 July; 8(15):1–8.
11. Zhao ZQ, Vogt KMB, Frerichs I. Customized evaluation software for clinical trials: an example on pulmonary function test with electrical impedance tomography. 978-1-4673-2971-2/13/2013.
12. Yun T, Tengyu H, Bing L, Jing T, Zhongjie Z. Regional Voltage Stability Prediction based on decision tree Algorithm. IEEE International Conference on Intelligent Transportation. 2015 Dec 19-20; 978-1-5090-0464-5/16. DOI: 10.1109/ICITBS.2015.

13. Ghorpade-Aher J, Metre VA. PSO based Multidimensional Data Clustering: A Survey. International Journal of Computer Applications. 2014 Feb; 87(16):1. (0975–8887).
14. Lakshmi KR, Krishna VM, Kumar VS. Performance comparison of data mining techniques for prediction and diagnosis of breast cancer disease survivability. Asian Journal of Computer Science and Information Technology. 2013; 3(5):81–7.
15. Suma VR, Renjith S, Ashok S, Judy MV. Analytical Study of Selected Classification Algorithms for Clinical Dataset. Indian Journal of Science and Technology. 2016 Mar; 9(11):1–9.
16. Bharati M Ramageri. Data mining techniques and applications. Indian Journal of Computer Science and Engineering. 2010 Dec; 1(4):301–05.
17. Mlambo N. Data Mining: Techniques, Key Challenges and Approaches for Improvement. International Journal of Advanced Research in Computer Science and Software Engineering. 2016 Mar; 6(3):59–65.
18. Kumari AD, Gunasekhar T. A Reconstruction Algorithm using Binary Transform for Privacy-Preserving Data Mining. Indian Journal of Science and Technology. 2016 May; 9(17):1–5.