International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 6 567 – 569

Diagnosis of Heart Disease Using K-means Clustering and Bell Curve Fitting

Natasha S Gajbhiye,
Mtech Scholar,
Rungta College of Engineering and Technology,
Bhilai

Mr. Kapil Nagwanshi,
Associate Professor,
CSE Department,
Rungta College of Engineering and Technology,
Bhilai

***Abstract*—** Due to changes in way lots of urban population experiences pathology and different heart related diseases. Many heart issues are because of irregular way and different factors like high cholesterol diets and lack of exercise. If on basis of medical records we are able to confirm patterns of heart issues we tend to scale back viscous connected cases within the health care system. Multiple factors impacting viscous health will be incorporated into the information set for locating different geographical, temporal and spatial correlations. The analysis proposes a strategy exploitation information mining to analyze patterns in tending significantly cardio-vascular diseases. The projected formula uses clustering for feature extraction within the vital organ (ex. Heart rate, sterol levels). It'll cluster the data and tell what per cent of individuals are healthy and how many are sick. The clusters are mapped with the given price information which can facilitate in finding out the insurance cover of the patients. Cleveland information set is employed for mapping of illness teams to price teams, other than Cleveland information sets 2 different information sets are used for comparative calculation of performance of K clusters on the information set.

***Keywords*—** *K- Means Clustering, Bell Curve Fitting.*

*****

I. INTRODUCTION

Heart disease may be a major health issue and it affects an outsized number of individuals. Upset Cardio Vascular Diseases(CVD) is one such threat. Unless detected and treated at an early stage it'll lead to sickness and causes death. There's no adequate analysis focus on effective analysis tools to find relationships and trends in information particularly within the medical sector. Health care industry these days generates great amount of advanced clinical information about patients and different hospital resources. Data processing techniques are used to analyze this collection of information from different views and deriving helpful information. The analysis proposes a strategy exploitation data mining to research patterns in tending significantly cardio-vascular diseases. The projected formula uses clustering for feature extraction within the organ (ex. Heart rate, cholesterol levels).

1.1 DATA MINING

Data mining is an interdisciplinary sub-field of computer science. It is the computational process of discovering patterns in large data sets ("big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

1.2 CLUSTERING

Clustering is a method of partitioning a group of information (or objects) into a group of substantive sub- classes, referred to as clusters. It helps users to grasp the natural grouping or structure during a knowledge set.. K–Means formula is that the simplest clustering formula, that classifies knowledge into k disjoint sets, by finding the Euclidian distance between knowledge points. Cluster analysis itself isn't one specific formula, but the general task to be resolved. It will be achieved by varied algorithms that disagree considerably in their notion of what constitutes a cluster and the way to with efficiency realize them.

II. PROBLEM DEFINITION

Prediction is one amongst the foremost difficult world issues. In large datasets there exists a spread of information things, which may or cannot be classified accurately to fastened categories. Info content of given information will solely be discovered through use of proper techniques and manual intervention to find which means of information points are required for prediction. Determining the quantity of clusters is that the initial drawback in clustering. For example, A batch of product from the producing unit can be classified into elect or rejected labels, on the opposite hand if we have a tendency to use three clusters, we are able to have intermediate quality labels for the merchandise which may be priced consequently. Clustering algorithmic program like k-means works on centralized data repository, area complexness may be a haul for large datasets during this case. Far away information points in dataset square measure usually considered dead i.e of no use once cluster. Thus when solving an information mining drawback with cluster desires manual intervention. The values of vital organ just in case of patient info, usually don't have a strict demarcation, we'd like to contemplate applicable boundaries for such very important characteristics.

SOLUTION OF THE PROBLEM

1. K-Means algorithmic program ought to be used with an effort and error approach to seek out the simplest worth of k.

2. Faraway information points either ought to be clean within the pre-processing stages of KDD, else they ought to be understood to create a special case, which is able to increase the effort of understanding frequent patterns in dataset.

III. METHODOLOGY

In this analysis we are going to use patient dataset, from multiple medical analysis institutes just like the Cleveland information and Hungarian information. Patient dataset include age, gender, chest pain, resting blood pressure, body fluid sterol, abstinence blood glucose, resting electrocardiographs, most pulse achieved, exercise induced angina, ST depression induced by exercise, slope of the peak exercise ST section, no. of major vessels colored by radiology, designation of heart condition (angiographic disease status).

4.1.1 DATA MINING STEPS

The block diagram of methodology is as shown in Figure 4.1

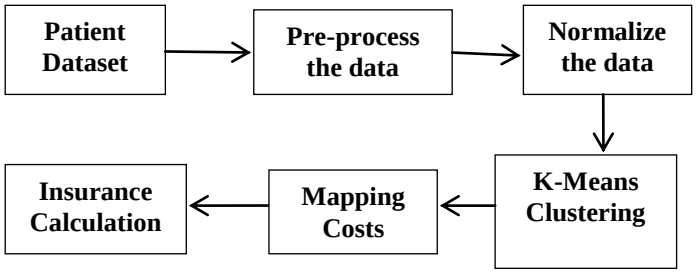

Figure 4.1: Block diagram of k-Means clustering

Total six steps are there to urge the desired output. They are as follows:

- **Patient Dataset**
  The patient data set consist of age, gender, chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographs, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise, slope of the peak exercise ST segment, no. of major vessels colored by fluoroscopy, diagnosis of Heart disease (angiographic disease status).

- **Pre-process the Data**
  Data pre-processing is an important step in the data mining process. If there's abundant tangential and redundant info gift or clamant and unreliable data, then data discovery throughout the coaching phase is harder. Information preparation and filtering steps will take appreciable quantity of process time. Information pre-processing includes cleansing, normalization, transformation, feature extraction and selection. The merchandise of information pre-processing is the final coaching set.

- **Normalize the Data**
  Normalization is a systematic way of ensuring that a database structure is suitable for general-purpose querying and free of certain undesirable characteristics that could lead to loss of data integrity. Normalization could be a systematic method of making certain that a database structure is appropriate for all-purpose querying and freed from bound undesirable characteristics that would cause loss of information integrity. Normalization is often a refinement method once the initial exercise of distinguishing the info objects that should be within the information, distinguishing their relationships, and process the tables needed and also the columns inside every table.

- **K- means Clustering**
  K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. The k-means algorithmic program takes as input the quantity of clusters to come up with, k, and a group of observation vectors to cluster. It returns a group of centroids, one for each of the k clusters. Associate degree observation vector is classified with the cluster range or center of mass index of the center of mass. Vector v belongs to cluster i if it's nearer to centroid i than the other center of mass. If v belongs to i, we say center of mass i is that the dominating center of mass of v. The k-means algorithmic program tries to reduce distortion, which is outlined because the add of the square distances between every observation vector and its dominating centroid. every step of the k-means algorithmic program refines The choices of centroids to scale back distortion. Given a set of observations ($\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS). In other words, its objective is to find:

$$\min_{s} arg \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

  where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

- **Mapping Costs**
  There is another data set which is the cost data. The cost data consists of the group, cost, expense and delay. The costs in this file are for individual tests, considered in isolation. When tests are performed in groups, there may be discounts, due to shared common costs. Tests with immediate results are marked "immediate". Tests with delayed results are marked "delayed". Delayed tests are typically blood tests, which are usually shipped to a laboratory. The full cost is

charged when the given test is the first test of its group that has been ordered for a given patient. The discount cost is charged when the given test is the second or later test of its group that has been ordered. The information in this file is meant to be used together with the information in expense. The tests in a group share a common cost.

For mapping cost, Bell Curve Fitting is used. To begin fitting a regression, put your data into a form that fitting functions expect. All regression techniques begin with input data in an array X and response data in a separate vector y, or input data in a table or dataset array tbl and response data as a column in tbl. Each row of the input data represents one observation. Each column represents one predictor (variable).

It is a linear model in which dataset is passed. After fitting a model, examine the result. Diagnostic plots help you identify outliers, and see other problems in your model or fit.

- **Insurance Calculation**
  The last step is the insurance calculation. On the basis of the patient and the cost data the insurance will be find out that how much insurance should be claimed to the patients.

IV. RESULT

There are several residual plots to help you discover errors, outliers, or correlations in the model or data. The simplest residual plots are the default histogram plot, which shows the range of the residuals and their frequencies, and the probability plot, which shows how the distribution of the residuals compares to a normal distribution with matched variance. The following is the histogram plot of residuals of our project.

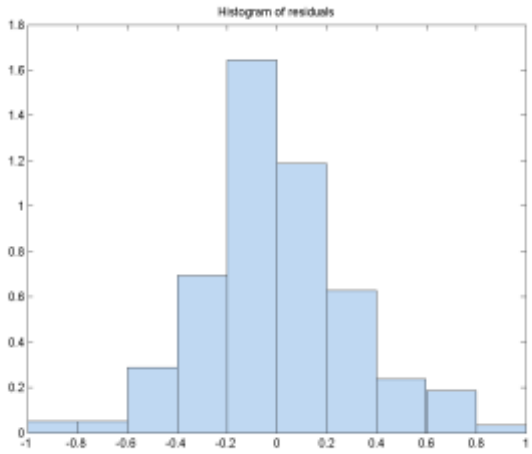

Fig. Histogram Of Residuals

V. CONCLUSION

Clustering is a data mining technique that makes meaningful or useful cluster of objects which have similar characteristics using automatic technique. K-means is a distance based clustering algorithm that partitions the data into a predetermined number of clusters. Each cluster has a centroid. Using this technique the patient data are clustered into clusters and find out that how many people belong to healthy group and how many belong to sick group. Using this the efficiency is also calculated. And for the insurance calculation Bell Curve Fitting is used. For mapping linear regression is used. Linear regression model contains an intercept and linear terms for each predictor.

REFERENCES

[1] G. Adda, J. Mariani, J. Lecomte, P. Paroubek, and M. Rajman. 1998. The GRACE French part-of-speech tagging evaluation task. In A. Rubio, N. Gallardo, R. Castro, and A.Tejada, editors, *LREC, volume I*, Granada, May.

[2] Abhishek Bhola, Twitter and Polls: Analyzing and estimating political orientation of Twitter users in India General #Elections2014, arXiv:1406.5059v1[cs.SI]19Jun2014.

[3] Ethem Alpaydin. 2004. Introduction to Machine Learning (Adaptive Computation and Machine Learning).

[4] Hayter Anthony J. 2007. Probability and Statistics for Engineers and Scientists. Duxbury, Belmont, CA, USA.

[5] Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews.

[6] Alec Go, Lei Huang, and Richa Bhayani. 2009. Twitter sentiment analysis. Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.

[7] Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury. 2009. Micro-blogging as online word of mouth branding. In *CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems.*

[8] Namrata Godbole, Manjunath Srinivasaiah,Steven Skiena, LargeScale Sentiment Analysis for News and Blogs. Google Inc., New York NY, USA, Dept. of Computer Science, Stony Brook University, Stony Brook, ICWSM'2007 Boulder, NY 11794-4400, USA.

[9] Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Alexander Pak, Patrick Paroubek, Universit´e de Paris-Sud, Laboratoire LIMSI-CNRS, B`atiment 508,F-91405 Orsay Cedex, France.

[10] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning.*

[11] Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts Department of Computer Science, Cornell University, Ithaca, NY 14853-7501.

[12] John Blitzer, Mark Dredze, Fernando Pereira, Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification, Department of Computer and Information Science University of Pennsylvania