Available online at www.ijpcr.com
International Journal of Pharmaceutical and Clinical Research 2016; 8(1): 75-79

ISSN- 0975 1556

Research Article

A Centroid-based Clustering Approach to Analyze Examinations for Diabetic Patients

Venkatesan M*, Doshi Aditya Ashvin, Sanket Bhambure, Roney Thomas

School of Computing Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, 632014, India

Available Online: 31st December, 2015

ABSTRACT

Health care sector is exploding with data from various streams such as patient history, insurance details, examination histories, drug prescription and many more. This data has immense potential to serve the health care sector in various ways. But due to the huge outbreak, the sector demands powerful and innovative mining tools for extracting useful information from these data. In this paper we propose a novel and explorative approach for data mining the similarities of examinations and medical conditions prevailing among diabetic patients which are in different age groups. This paper, we use a centroid-based clustering algorithm for grouping the data obtained and a decision tree classifier for classification. The decision tree classifier is chosen because by looking at the tree structure all persons who uses the model can easily understand what the model means. For studying different data attributes, the clustering algorithm has been taken in multi-level fashion.

Keywords: Centroid-based clustering, diabetes, decision tree, k-means, partial parenteral nutrition

INTRODUCTION

Health care industry is growing in different dimensions as the data getting generated. Using machine learning and advanced technology for healthcare, we are in a state of position to deliver the best quality treatment and care to any categorical patients. Due to massive dimensional increase of data in various aspects, it is not easy to manage the same. Data mining, which helps in studying effective and efficient algorithms to transform large amount of data into useful knowledge[5]. If this massive data is well managed, we can use the data for various helpful activities. Information mining has been utilized seriously and widely by numerous associations. In human services, information mining is turning out to be progressively well known, whether not progressively crucial. Information mining applications can incredibly advantage all gatherings included in the medicinal services industry. For instance, information mining can help medicinal services guarantors recognize misrepresentation and misuse, human services associations settle on client relationship administration choices, doctors distinguish compelling medications and best practices, and patients get better and more reasonable social insurance administrations. In this paper, we are proposing an efficient mining approach Centroid-based Clustering Approach called K-means clustering approach data mining, to discover cohesive and well-separated groups of patients with a similar profile (i.e., patient age and gender) and examination history[4]. A number of clustering algorithms is available. We propose the k means approach for clustering because simple and easily separate the cluster. The algorithms prefer mostly clusters of approximately similar size, the nearest centroid have always similar type of objects. Therefore, it can be used to classify the clusters. Also the computational complexity is at lowest level[4]. The paper combines the clustering and classification techniques to discover homogenous group of patients, and then fully characterize these groups[1]. The clustering approach can be used to address different issues such as medicine dose, diet, analysis of patient history so that we can find whether the patient has some allergy to medicines or not.

LITERATURE SURVEY

Diabetics can be classified into 3 types. Type-1 diabetes, named as juvenile-onset diabetes. The body does not produce proper insulin and It's usually caused by an auto-immune reaction that where the body's defense system attacks the cells that produce insulin. Type-1 is also called insulin dependent. Around 10% of all diabetes cases are sort 1. The Type-2 is called non-insulin ward and almost 90% of the diabetes experiences Type-2 and Adult-onset diabetes. This is considered by imperviousness to insulin and insufficiency of relative insulin which might either or both together be available at the season of which diabetes is analyzed. So on account of Gestational Diabetes (GDM) which is regularly seen amid pregnancy coming about because of high glucose levels. This is frequently refined by renal inconveniences, heart ailments and fringe vascular sicknesses.

Danielle M. Hesseler, Lawrence Fisher, Russell E. Glasgow, Joseph T.Mullan, and Umesh Masharani[18,19] has come up with the neglector factor on considering disease diagnosis in adults having Type-2 diabetes and studied life and age of a patient is determining diabetes and its management. Mira KaniaSabariah, SitiSa'adah and

*Author for Correspondence

Table 1: The clustering is based on the patient's age and examination history. All clusters formed are well separated.

C0,C9	Patients with cholesterol
C3,C5,C13,C15,C16	Patients with stable glucose level
C6,C10,C18,C11,C17	Patients partial parenteral nutrition
Other clusters	Patients with normal values

AiniHanifa[19,20] coined diagnosing Diabetes Mellitus which is of Type-2 with Random Forest, regression tree.

Robert Bailey, Kathy Annunziata, Janice, Lopez, Marcia[19,21] proposed the fact that from a diabetic patient viewpoint there are many risk factors that includes family history, obesity, older age and physical dormancy. MemetIsik, ZekeriyaAkturk, Abdul Sattar Khan, UmitAvsarand Turan Set[19,22] focused on hypertension, myocardial infraction, and stroke due to diabetes in different age groups of both genders.

M.Mounika, S.D.Suganya, B.Vijayashanthi and S.KrishnaAnand[19] proposed a predictive analysis for Diabetic treatment using classification. The study focused on classification algorithms like Naïve Bayes, ZeroR and OneR. They studied on the factors like drug intake, diet, obesity, insulin deficiency.

Identify the patients on the basis of his blood report and all other report and will be categories disease. If the patients have high glucose, then its diabetes describe metabolic diseases because insulin production inadequate then body's cell don't respond properly to insulin. The diabetes increase has many serious health problems such as kidney diseases, cardiovascular disease, retinal damage and Hypoglycemia. Patients having type-2 diabetes which is undiagnosed or which is at a high level of risk developing type-2 is one of the most important challenges in medical field.

The availability of enormous amount of medical data is demanding powerful data analysis tools to mine useful knowledge. It has been very big challenge for the researchers to discover statistical and efficient data mining tools that can improve data analysis on large data scale. Diagnosing diseases, such as diabetics at early stages is one of the applications where data mining tools have proved impeccable results in the recent years.

In this paper, we discuss a clustering approach named Centroid-based Clustering Approach for analyzing the diabetic patients. The dimensional area is separated into different clusters and we are going to apply the decision tree on that cluster. We can make a hierarchical structure, since we are having number of clusters. The Clusters can then easily be defined as objects belonging most likely to the same distribution. The aims of K-means clustering is to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster and the results in a partitioning of the data space into clusters.

Data mining would be a valuable benefit for diabetes researchers because it can extract hidden knowledge from a huge amount of diabetes related data[5]. Clustering is the process of analyzing and grouping the data into different clusters or classes such that objects within the same cluster will be having more similarity to each other, but will be having different properties in other clusters. The principle objective of clustering in this area is to find different groups of diabetes patients with similar symptoms within a group but different symptoms of other groups[6]. The supervised learning of algorithm in contrast with Clustering is called Classification. It classifies or maps a data item into any one of many predefined classes.

METHODOLOGY

Nowadays the amount of data gathered and stored in rapidly increasing. The large amount of data being gathered, retrieved and analyzed across the world. Many algorithms are created to find the data over the large databases. Among the several techniques classification is the one of the technique to analyze and gate information that is to predict the data. Classification method used to find or predict the output for the given input. These methods train datasets as well as their attributes and try to find out the relation between those attributes. This method classifies the data in to different classes using its similarity with the previously soared data which can be done with reference of data models which are previously present.

Clustering is nothing but making the group of the set of the objects in such a way that those are almost similar to each other. This is the most important in machine learning techniques. Cluster is nothing but the group of objects which are treated as a single group. While making analysis using clustering firstly creates the set of groups based on similarity between data and that group is assigned with label. Clustering is useful to find broad variety of groups from the large amount of data. Clustering can characterize particular group. Clustering is useful in pattern detection among the huge range of values of the data, it is also useful in the constructing the data concepts as well as in the learning process of the non-supervised data. Clustering based on the distance between the attribute valued is the one of the clustering technique. Some other techniques which are used for the clustering are "distance based, hierarchical, partitioning, center based, property or conceptual and probabilistic". The best clustering algorithm is one which has huge amount of similarity between intra classes and very little similarity between the inter class. Clustering algorithms have the ability to find or discover the almost all the patterns from the data. Simply segmenting data in to groups on the attributes value is not the clustering. To understand and summarize the huge amount of data clustering algorithms are very useful.

Decision tree is the one of the classification technique. The aim to create decision tree is to produce models that can guess or predicts variables target range of a value. This tree represents the subsets of the given dataset which are grouped according to the values of the given attributes. Structure like a flowchart is created when decision tree is produced. Each leaf of the tree has a label according to the classification of the data. Algorithm of decision tree works from head to tails that is from start to bottom by analyzing and choosing attributes or variables at each and every step

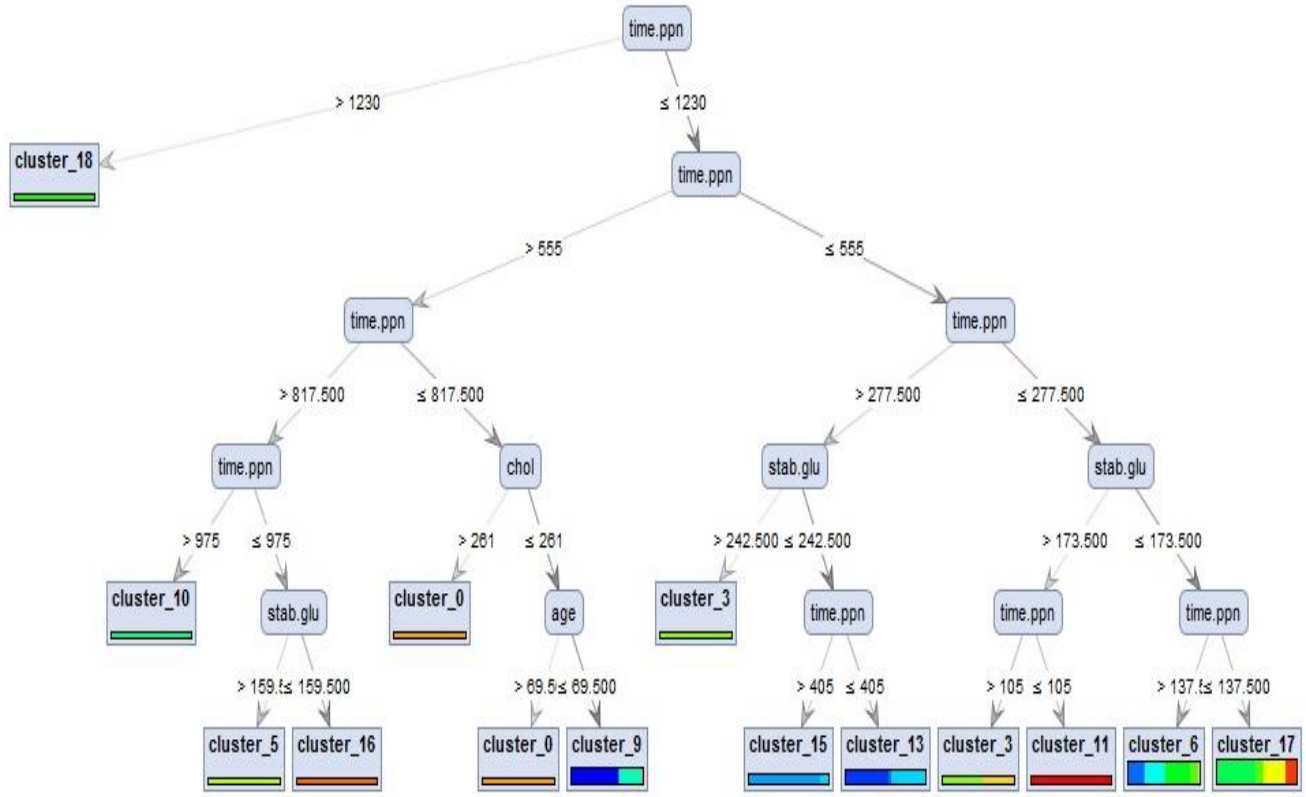

Figure 1: Classification tree

and splits dataset in the best possible way. Different algorithms have different way for measuring good decision. In general, the homogeneity between the attributes of the set of the data is measured.

Patient Representation

Diabetes increases the serious risk for many health problems, such as retinal damage, cardiovascular disease, foot complications and kidney disease if patients has high glucose it belongs to the metabolic diseases and body cells don't respond properly to insulin. Males and females both can be affected by diabetes. To discover groups of patients with similar profile and diabetic status, for this we used the k-means approach in this study. The patient done set of different medical test (ex. Blood test). More specially weight describing the relevance of examination for the patient.

Algorithm: k-means clustering

Let X = {x1, x2, x3, ….,xn} be the set of data points and V = {v1, v2,….,vc} be the set of centers.

1) Select 'c' cluster as center and its random selection.

2) Find the distance between each node or data point and cluster centers.

3) The cluster center should have assign by data point whose distance from the cluster center is minimum of all the cluster centers.4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'ci' is represents the number of data points in i^{th} cluster.5) Recalculate the distance between each data point and new obtained cluster centers.6) If no data point was reassigned then stop, otherwise repeat from step 3).

Algorithm: Decision tree Classifier

1. Create root node
2. If all attributes are positive, return single-node tree as root, label =+
3. If negative, return single-node tree as root, label = -
4. If attributes which used for prediction set is empty, return the single node tree as root, label = common value of target attribute
5. Else

i. A = attribute that has best entropy

ii. Root node for decision tree = A

iii. for each value,v_i, of A

- add a new branch
- If no more values, then below the branch add a leaf node with common target value.

iv. end

return root

Patient Clustering

Cluster analysis partitions objects into groups so that objects within the same group are more similar to each other than those objects assigned to different groups[1]. Clustering dealing with similar group of patient has in same cluster and for clustering its requires medical datasets. They can show various possible examinations for different disease severity, and clusters could be of arbitrary shapes. And then it's get the number of clusters for this clustering k-means clustering algorithm is best suitable for the analysis. The k-means clustering discover the group of patients according to the attributes value.

The proposed k-means clustering approach prefer mostly clusters of approximately similar size, the nearest centroid

have always similar type of objects. Clustering patients requires the definition of a metric to evaluate the distance (or similarity) between patients based on the features describing them. Then, the quality of computed clusters is evaluated according to some indices[1].

RESULTS AND ANALYSIS

The open source RapidMiner toolkit has been used for the classification analysis of cluster. Experiments were performed on Intel(R) Core(TM) 2 Due CPU @ 2.10GHz with 3 Bytes of main memory. This section discusses the results obtained when analysing a real collection of diabetic patients with the proposed approach.

A. Datasets

For this study the dataset was collected on Disease Heath Control Center. The data contain gender, weight, age, blood pressure (bp) etc. of 403 patients. Examinations contain both routine and more specific tests to analyze diabetes complications on various degrees of severity. The dataset includes both male and female patients[1].

B. Cluster Result

The k-means clustering algorithm generated 20 clusters which has been iterated for 2 levels. The cluster details are shown in the Table 1. The clustering is based on the patient's age and examination history. All clusters formed are well separated. Table 1 gives the patients conditions alsoThe clusters at first-level discuss about the patients mainly undergoing regular checkups to monitor diabetic's conditions in blood (C3, C5). Also the clusters contain patients who undergo tests to monitor nutrition level in blood (C18). The second-level of clusters show a wide range of examinations taken by patients for diagnosing several diabetic conditions.

C. Classification Result

The decision tree classifier algorithm is used on the clustering result. Each node of the decision tree represents each characteristic of a patient (like age, nutrition value, glucose level). The branches depict the range of possible value a node can have. Cluster name *Ci* represents the class label for each cluster. The classification result tree is shown in the figure 1.

CONCLUSION

This paper presents a novel approach based on a multiple level clustering strategy to identify groups of patients with similar characteristics and examination history in a dataset with a variable data distribution. From this classification result, it is easy to understand the various scenarios like patient drug prescription, earlier test result values etc. Through this the doctor and understand the pre-historic conditions of the patients even if the regular doctor of the patient gets changed. Also this can be used for Health survey by Government health departments to find the age groups which have the highest and lowest ranges of diabetics.

REFERENCES

1. Giulia Bruno, Tania Cerquitelli, Silvia Chiusano, Xin Xiao.,A Clustering-Based Approach to Analyse Examinations for Diabetic Patients .,in 2014 IEEE International Conference on Healthcare Informatics.
2. B. M. Patil, R. C. Joshi, DurgaToshniwal, "Association rule for classification of type -2 diabetic patients",Proc. of the Second International Conference on Machine Learning and Computing, pp 330-334, 2010
3. Wei Wang, Jiong Yang, Richard Muntz, STING: A Statistical Information Grid Approach to Spatial Data Mining, Proceedings of the 23rd VLDB Conference Athens, Greece, 1997.
4. Pang-Ning T. and Steinbach M. and Kumar V., Introduction to Data Mining. Addison-Wesley, 2006.
5. Kaufman, L. and Rousseeuw, P. J., Finding groups in data: An introduction to cluster analysis. Wiley, 1990.
6. Zheng Jye Ling, Quoc Trung Tran, Ju Fan, Gerald C.H. Koh, Thi Nguyen, Chuen Seng Tan, James W. L. Yip, Meihui Zhang., GEMINI: An Integrative Healthcare Analytics System., Proceedings of the VLDB Endowment, Vol. 7, No. 13
7. K. Chaturvedi, "Geographic concentrations of diabetes prevalence clusters in Texas and their relationship to age and obesity," vol. 9, no. 7, p. 2010, 2003.
8. http://www.ahrq.gov/professionals/prevention-chronic care /improve/ system/pfhandbook/ mod8appbmonicalatte.html
9. https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/
10. http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_033589.hcsp?dDoc Name= bok1_03358
11. https://www.medicare.gov/download/downloaddb.asp
12. http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets
13. DivyaTomar and Sonali Agarwal, 'A survey on Data Mining approaches for Healthcare', International Journal of Bio-Science and Bio-Technology, vol. 5, no. 2552013, 2013.
14. M. Durairaj, V. Ranjani,'Data Mining Applications In Healthcare Sector: A Study', INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, vol. 2, no. 10, 2013.
15. Boris Milovic, Milan Milovic,'Prediction and Decision Making in Health Care using Data Mining', International Journal of Public Health Science, vol. 1, no. 2, 2012.
16. Fawaz AL Hazemi, Chan Hyun Youn, 'Grid-based Interactive Diabetes System', International Conference on Healthcare Informatics, Imaging and Systems Biology, 2011.
17. S. Mougiakakou, A. Prountzou, and K. Nikita, "A Real Time Simulation Model of Glucose-Insulin Metabolism for Type 1 DiabetesPatients", in Proc. 27th Annu. Conf. IEEE Engineering in Medicineand Biology Society, Shanghai, China, 2005, pp. 298-301.
18. Danielle M.Hessler, lawrenceFisher, Joseph T.Mullan, Russel E.Glasgow, UmeshMasharani., Patient age: A neglected factor whenconsidering disease management in adults with type 2 diabetes,Elsevier Journal,85 (2011), 154–159.
19. M.Mounika, S.D.Suganya, B.Vijayashanthi, S.KrishnaAnand, "Predictive Analysis of Diabetic Treatment Using Classification Algorithm, (IJCSIT)

International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2502-2505

20. Mira KaniaSabariah, AiniHanifa and SitiSa'adah, Early Detection of Type-2 Diabetes Mellitus with Random Forest, Classification and Regression Tree, IEEE Transaction, 2014, 238-242,.