Available online at http://acfa.apeejay.edu

APEEJAY JOURNAL OF COMPUTER SCIENCE AND APPLICATIONS

ISSN:0974-5742(P)
Vol. 1, January 2013

Diabetes Classification using K-Means

Gagandeep Singh[1], Gurpreet Singh[2]

[1]Assistant Professor, Sai Institute of Engg.& Tech.Manawala, Amritsar (Punjab)India.
[2]M.tech (CSE) ,Sai Institute of Engg. & Tech. Manawala, Amritsar (Punjab)India

Abstract

This paper presents a survey on proposed quantized techniques in medical field to understand which age group of people are being affecting by diabetes. The overall aspiration of the diabetes data mining procedure is to extract information from diabetes data set and convert it into a comprehensible arrangement for further use. This paper deals with diabetes and proposed a data mining system using simple K-Means and nearest neighbor hierarchical clustering. The main objective is to find the effects of diabetes on the peoples grouped by age and evaluating the survival ratio in efficient manner. Accuracy, Sensitivity and Specificity are different metrics which will be evaluated in this survey.

Keywords: Data Mining,diabetes, K-means, Herarical clustering.

1. Introduction

In Singapore, about 10 percent of the population is diabetic [1]. This disease has many side effects such as higher risk of eye disease, higher risk of kidney failure, and other complications. However, early detection of the disease and proper care management can make a difference. To combat this disease, Singapore introduced regular screening program for the diabetic patients in 1992.Patient information, clinical symptoms, eye-disease diagnosis and treatments are captured into a database. After eight years of data collection, a whole wealth of information has been gathered.

This leads naturally to the application of knowledge discovery [1] & [2] and data mining techniques to discover interesting patterns that exist in the data. The objective is to find rules that can be used by the medical doctors to improve their daily tasks, that is, to understand more about the diabetic disease or to find out something special about a particular patient population. Although knowledge discovery in databases has reported many successes in domains such as fraud detection, targeted marketing etc., we found that the application of data mining techniques to health sector has been relatively few in comparison. We believe this is primarily due to two reasons. First, the data captured by health clinics are typically very noisy. Many of the patient records contain typographical errors, missing values, or wrong information such as street names or date of birth etc; and worse, many records are in fact duplicate records. Cleaning these data takes tremendous amount of effort and time. In addition, many of the data collected are not in the forms that are suitable for data mining. They need to be transformed to more meaningful attributes before mining can proceed. Second, the health doctors are usually too busy seeing patients each day. They cannot afford the time or the energy to sieve through the thousands of rules generated by some state-of-the-art mining techniques on the diabetic patient database.

Thus, it is important to present the discovered rules in some easy-to-understand fashion. In this paper, we will demonstrate how we address these concerns. To overcome the problem of noisy data, we have developed a semi-automatic data cleaning system. The system reconciles database format differences by allowing the doctors to specify the mapping between attributes in different format styles and the encoding schemes used. Once the format differences have been reconciled, the problem of identifying and removing duplicate records is addressed. To resolve the problem of too many rules generated by the state-of-the-art mining techniques, we apply a user-oriented approach that provides step-by-step exploration of the data in order to better understand the discovered patterns.

Clustering [2]-[8] is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters. Clustering can be considered the most important unsupervised learning technique so as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data. Clustering is ―the process of organizing objects into groups whose members are similar in some way.

A cluster is therefore a collection of objects which are ―similar‖ between them and are ―dissimilar‖ to the objects belonging to other clusters. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. Moreover, most of the data collected in many problems seem to have some inherent properties that lend themselves to natural groupings. Clustering algorithms are used extensively not only to organize and categorize data, but are also useful for data compression and model construction.

The main goal is to find the effects of Diabetes and evaluating the survival ratio in efficient manner. Accuracy, Sensitivity and Specificity are different metrics which will be evaluated in this survey.

2. Literature Survey

2.1. Survey on Diabetes

Hsu et al. [1] Real-life data mining applications are interesting because they often present a different set of problems for data miners. One such real-life application that we have done is on the diabetic patients databases. Valuable lessons are learnt from this application. In particular, we discover that the often neglected pre-processing and post-processing steps in knowledge discovery are the most critical elements in determining the success of a real-life data mining application. In this paper, we shall discuss how we carry out knowledge discovery on this diabetic patient database, the interesting issues that have surfaced, as well as the lessons we have learnt from this application. We will describe a semi-automatic means for cleaning the diabetic patient database, and present a step-by-step approach to help the health doctors explore their data and to understand the discovered rules better.

Rajesh, K., et al. [2] has worked on medical professionals need a reliable prediction methodology to diagnose Diabetes. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. The main goal of data mining is to discover new patterns for the users and to interpret the data patterns to provide meaningful and useful information for the users. Data mining is applied to find useful patterns to help in the important tasks of medical diagnosis and treatment. This project aims for mining the relationship in Diabetes data for efficient classification. The data mining methods and techniques will be explored to identify the suitable methods and techniques for efficient classification of Diabetes dataset and in mining useful patterns.

2.2. Survey on Clustering

Maulik et al. [3] has demonstrated that the hierarchical clustering builds a cluster hierarchy. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent.

Agglomerative (bottom up)

1. Start with 1 point (singleton).
2. Recursively adds two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

Divisive (top down)

1. Start with a big cluster.
2. Recursively divides into smaller clusters.
3. Stop when k number of clusters is achieved.

General steps of Hierarchical Clustering: Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering (defined by S.C. Johnson in 1967) is this:

- Start by assigning each item to a cluster, so that if we have N items, we now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
- Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now we have one cluster less.
- Compute distances (similarities) between the new cluster and each of the old clusters.
- Repeat steps 2 and 3 until all items are clustered into K number of clusters.

Chaudhari et al. [4] Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. This paper analyze the three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. Performance of the 3 techniques are presented and compared using a clustering tool WEKA.

Han et al. [5] in their book proved that the knowledge discovery process consists of an iterative sequence of steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining functionalities are characterization and discrimination, mining frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis. Three of the major data mining techniques are regression, classification and clustering. CLUSTERING is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters.

Lopez et al. [6] has proposed a classification via clustering approach to predict the final marks in a university course on the basis of forum data. The objective is twofold: to determine if student participation in the course forum can be a good predictor of the final marks for the course and to examine whether the proposed classification via clustering approach can obtain similar accuracy to traditional classification algorithms. Experiments were carried out using real data from first-year university students. Several clustering algorithms using the proposed approach were compared with traditional classification algorithms in predicting whether students pass or fail the course on the basis of their Moodle forum usage data. The results show that the Expectation-Maximisation (EM) clustering algorithm yields results similar to those of the best classification algorithms, especially when using only a group of selected attributes. Finally, the centroids of the EM clusters are described to show the relationship between the two clusters and the two classes of students.

Shou et al. [7] has studied that data mining is a technique to search potential valuable information from databases. Preventing personal data and high security data therefore pose a difficult task to IT experts. Shou et al. [7] has proposed a novel anti-datamining (ADM) database security scheme, which protect against data mining. The scheme makes use of hierarchical clustering where noise is added to change the cluster structure of data.

The proposed hierarchicalanti-clustering (HAC) scheme modifies the cluster structure of the original data. Experimented results show that data may be protected against during the HAC key can be used reverse the cluster structure to its original.

Rui et al. [8] proved that the clustering is a discovery process in data mining and can be used to group together the objects of a database into meaningful subclasses which serve as the foundation for other data analysis techniques. The authors focus on dealing with a set of spatial data. For the spatial data, the clustering problem becomes that of finding the densely populated regions of the space and thus grouping these regions into clusters such that the intracluster similarity is maximized and the intercluster similarity is minimized.

Lisboa et al. [9] has done research and study on Diabetes and proposed that the data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups such that the similarity within a group is larger than among groups. This paper reviews six types of clustering techniques- k-Means Clustering, Hierarchical Clustering, DBScan clustering, Density Based Clustering, Optics , EM Algorithm. These clustering techniques are implemented and analysed using a clustering tool WEKA. Performance of the 6 techniques are presented and compared.

3. Problem Formulation

The main purpose of this survey is to predict how likely the people with different age groups are being affected by diabetes based on their life style activities and to find out factors responsible for the individual to be diabetic. Hence it is motivating to implement quantized techniques in medical field to understand which age group of people are being affecting by diabetes.

The overall goal of the diabetes data mining process is to extract information from diabetes 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 preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

3.1. Research methodology

To attain the objective, step-by-step methodology is used in this dissertation. Subsequent are the different phases which are used to accomplish this work.

3.2. Orientation

This research work starts with the orientation in the area of Diabetes by consulting websites, reading news articles, participating in seminars and discussing with the experts. This research employs a structured method to obtain high quality information, called a Literature survey.

3.3. Literature survey

To explore the available knowledge on the area of data mining, Diabetes and nearest neighbor hierarchical clustering, literature survey will be conducted using a systematic approach. High quality papers are selected to explore the existing techniques.

3.4. Proposed algorithm implementation

Proposed algorithm, which will be designed and implemented in MATLAB. Different type of test will be conducted by taking different kind of Diabetes.

3.5. Performance analysis

In order to do performance analysis, comparisons will be made with different existing methods. Comparisons table and diagrams will be made based upon the outcomes of the experimental results. Different parameters will be considered and evaluate from available data set.

4. Objectives to be achieved

The objectives of this dissertation are:

(1) To briefly review the application of data mining techniques in Diabetes detection/diagnosis;

(2) To explore a novel analytic method with different feature selection methods;

(3) To compare the results obtained on different datasets and that reported by different authors in terms of detection performance and selected Diabetes types.

(4) Implementation of proposed algorithm in MATLAB to validate and verify proposed algorithm.
(5) Calculating different parameters like Accuracy, Sensitivity and Specificity.

5. Conclusion

The main purpose of this survey is to predict how likely the people with different age groups are being affected by diabetes based on their life style activities and to find out factors responsible for the individual to be diabetic. Hence it is motivating to implement quantized techniques in medical field to understand which age group of people are being affecting by diabetes.

The overall goal of the diabetes data mining process is to extract information from diabetes 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 preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

References

[1]. Hsu, Wynne, Mong Li Lee, Bing Liu, and Tok Wang Ling. "Exploration mining in diabetic patients databases: findings and conclusions." *InProceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 430-436. ACM, 2000.

[2]. Rajesh, K., and V. Sangeetha. "Application of Data Mining Methods and Techniques for Diabetes Diagnosis." *age 2, no. 3 (2012).*

[3]. U. Maulik, and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices", Pattern Analysis and Machine Intelligence*, IEEE Transactions, Vol. 24, 2002, pp. 1650-1654.*

[4]. Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining*", International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384*

[5]. J. Han, and M. Kamber, Data Mining: Concepts and Techniques, *Morgan Kaufmann, 2006.*

[6]. Lopez, M. I., J. M. Luna, C. Romero, S. Ventura, M. M. Molina, J. M. Luna, C. Romero et al. "Classification via clustering for predicting final marks based on student participation in forums}}." *In Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012}, vol. 42,* pp. 649-656. 2012.

[7]. Tung-Shou Chen; Jeanne Chen; Yuan-Hung Kao; Bai-JiunTu "A Novel Anti-Competitive Learning Neural Network Technique against Mining Knowledge from Databases", *Software Engineering, 2009. WCSE '09. WRI World Congress on, On page(s): 383 - 386 Volume: 4, 19-21 May 2009*

[8]. RuiXu; Wunsch, D., II "Survey of clustering algorithms*", Neural Networks, IEEE Transactions on, On page(s): 645 - 678 Volume: 16, Issue: 3, May 2005*

[9]. Lisboa, Paulo JG, Alfredo Vellido, Roberto Tagliaferri, Francesco Napolitano, Michele Ceccarelli, José D. Martín-Guerrero, and EliaBiganzoli. "Data Mining in Cancer Research [Application Notes]." *Computational Intelligence Magazine, IEEE 5, no. 1 (2010): 14-18*