Diabetic diagnosis through data mining

S. Sankaranarayanan[1], Dr. T. Pramananda Perumal[2]

[1]*Associate Professor of Computer Science, Government Arts College (Autonomous), Kumbakonam – 612 002, India & Research Scholar, R &D Centre, Bharathiar University, Coimbatore - 641 046, India*
[2]*Principal, Presidency College (Autonomous), Chennai - 600 005, India*
[1] profsankaranarayanan@yahoo.in, [2] pramanandaperumal@yahoo.com

Abstract

In the field of healthcare, patient inputs either go explicit or inherently implicit. Since the implicit are of buried patterns they need to be explored for proper diagnosis and identification. The elementary and aggregate forms of documented data are the key to open mines to mine hidden information. Diabetes mellitus disease (DMD) is a metabolic disorder affecting more people nowadays and hence the level of incidence increases exponential. Medical practitioners are now keen to use data mining techniques for such victims. Applying heuristics on patient records have proved to be worthwhile and valuable in predicting Diabetes. This paper parameterizes patient's pathological inputs for Diabetes mining. Important techniques such Rule mining and Decision trees are applied here for finding hidden knowledge.

Keywords: *DMD, HbA1c, Rule mining, Classification*

1. Introduction

Data mining of databases nowadays has become a dependable mechanism to predict decisions including diseases like heart attack, diabetic mellitus and other complicated ailments [1]. Statistical analysis of attribute oriented data and mining associative logic rules from a larger data are the classical approaches to data mining. This paper mainly focuses on Diabetic disease prediction using data classification. The outcomes of these are discrete and categorical and hence build models in the form of IF…THEN rules. Decision tree is yet another model in which a tree node denotes a test on parametric attributes and the leaves can model classes. Cluster analysis deals with data objects without a consultative (known) class label and thus contributes to unsupervised learning.

2. The Background

A simple diabetes disease prediction engine is devised here using attribute analysis of augmented patient data taken from susceptible and people prone to have diabetes disorder. Two different data sets are used here and identified compatible for diabetes data mining application. Data Classification method leads to some logical inferences in the form of IF…THEN rules and are enlisted as results of data analysis and mining of sample data as found in Table-1. Decision tree is also being built on another sample data set as found in Table-2. All these would contribute immensely useful in identifying Diabetes.

3. Material and Methods

3.1. Data Preprocessing

Data mining approaches require data to be preprocessed for removing noise, finding missing values and to identify incomplete and inconsistent data not to have crept in. Methods like normalization, eliminating data redundancy and clustering are immensely helpful. Also other methods such as cleaning, filtering and binning ensure for appropriate outcome by data mining. The method of hierarchical clustering using value hierarchy is used for data preprocessing here. Clustering algorithm partitions data into classes or groups called clusters. K-means algorithm is familiar in this context. It is nothing more than an iterative refinement method which initially identifies K clusters on N objects (K <= N) and places every object in any of K clusters and improves partitioning by moving objects from one group to other iteratively until some conditions are met.

3.2. Predictive data mining

Predictive data mining builds a data model from existing data to predict the behavior of new data set values [2]. Association Rule mining is the classical example for this. Data analysis of sample data set of Table-1 produces several inference rules (IF…THEN rules) to predict diabetic disease are shown here due to the process of mining frequent itemsets.

Table 1. Attributes for DMD prediction

Age (in yrs)
Sex (M / F)
Mass (in kg)
Height (in cm)
Blood Glucose Level (in mg / dL) Fasting (FBGL) Random(RBGL) Post-Prandial(PBGL)
Level of Obesity (in BMI), Body Mass Index

3.3. Mining Frequent Itemsets

Association Rule mining is a concept which portrays futuristic prediction of hidden patterns from a large transactional database and hence applied to medical data analysis [3]. An association rule inducted from an transaction T $\subseteq$ X , the m-itemset X = {i1, i2, i3,…,im} and D is the sample space of all transactions of the form T. Association Rule Mining is based on inference logic and inference calculus.

Association Rule Mining

A => B means that if (A) then (B) where A,B $\subseteq$ X and A $\cap$ B = Ø,

Here the transaction A => B holds D with support ' **s** ' v.i.z., the support (A => B) = the percentile AUB in D, i.e P(AUB)

Also the rule A => B holds D with confidence 'c' v.i.z., the confidence (A => B) = the percentile A that also contains B in D, i.e P (B|A)

Association rules derived thus far which satisfy both minimum support and minimum confidence levels are said to be strong rules. An itemset with k items is a k-itemset. For instance the set {Man, Diabetes} is a 2-itemset. The 'occurrence frequency' otherwise regarded to be the 'support count' of an itemset is the number of transactions that contains the given itemset. An itemset complies to minimum support when the occurrence frequency of the itemset is grater than or equal to the product of minimum support and total number of transactions in D. If an itemset satisfies minimum support count, then it is a frequent itemset. The set of frequent k-itemsets is commonly denoted by L_k

The algorithm of Association rule mining has two steps:

1. Finding all frequent itemsets

2. Generating strong association rules from the frequent itemsets

The overall performance of this algorithm lies in the first step i.e identifying possible frequent itemsets. Sample results of this approach are enlisted here below:

Example 1: If_then_rule induced in the diagnosis of DMD

IF Sex = MALE AND Age >= 30 AND

Mass >=(Height – 100)

AND (Fbgl >=100 OR Rbgl >=130 OR Pbgl >=155)

AND Bmi >=27

THEN Diagnosis = Incidence of DMD

Example 2: If_then_rule induced in the diagnosis of DMD

IF Sex = FEMALE AND Pregnant AND Age >= 20

AND (Fbgl >=120 OR Rbgl >=150 OR Pbgl >=175)

THEN

Diagnosis = Incidence of DMD in GESTATIONAL period

Example 3: If_then_rule induced in the diagnosis of DMD

IF Age = <45 AND Overweight = “no” AND Alcohol Intake = “never” THEN

Incidence level of DMD is low

Example 4: If_then_rule induced in the diagnosis of DMD

IF Age = >70 and Blood pressure = “High” AND

Smoking = “current” THEN

Incidence level of DMD is high

3.4. Decision Trees

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision trees are commonly used under classification for the purpose of decision making and knowledge representation [4]. Decision tree starts with a root node on which it is for users to take further traversal actions. From this node, users split each node recursively according to decision tree learning algorithm of choice.

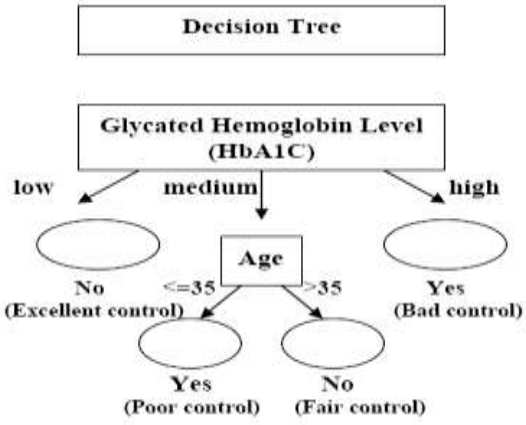
Decision tree generation consists of two phases, tree construction and tree pruning which identifies and removes noisy branches or outliers. A decision tree traversal always halts with a decision. Decision tree classification can be best used to mine Diabetes data [5]. The decision tree shown in Fig.1 is built from the very small training set as available in Table-2 with each row corresponding to a patient record instance.

Table 2. Data set used to build decision tree of Fig.1

Age	Gender	Glycated HG (**HbA1c**) level	Disease (Goal)
25	Male	Medium	yes
32	Male	High	yes
24	Female	Medium	yes
44	Female	High	yes
30	Female	Low	no
21	Male	Low	no
18	Female	Low	no
34	Male	Medium	no
55	Male	Medium	no

Training data: HbA1c = low when <= 5.5: medium when > 5.5 & < = 7: high when > 7

Figure 1. A decision tree built form data in Table 2



The data entity set contains three predictor attributes Age, Gender and Glycated Hemoglobin (HbA1C) level and one goal attribute, namely disease control value which indicates whether the corresponding patient DMD victim or not [6]. Decision tree technique can even classify an unknown class data. As an example, the data instance of the tuple (Age=27, Gender=male, HbA1C = medium, Goal =?) where "?" denotes the unknown value of the goal instance.

4. Conclusion

Medical related data is always huge and complex in nature as is derived from a variety of causes and rooting factors. In this paper, diabetes disease prediction system was developed using classification methods to predict the level of incidence of Diabetes disease accurately. The rules inducted using these approaches have the potential to devise an expert system for making better clinical decisions. In a thickly populated country with scarce resources such as India, public awareness can also be achieved through the effective knowledge dissemination and management.

5. References

1. Sellappan Palaniappan and Rafiah Awang [2008] Intelligent disease prediction system using data mining techniques, *International Journal of Computer Science and Network Security*, vol.8, (8), pp.343-350.
2. Jyoti Soni, Ujma Ansari, Dipesh Sharma and Sunita Soni [2011] Predictive data mining for medical diagnosis: an overview, *International Journal of Computer Science and Engineering*, vol. 3, pp.43-48.
3. K.Shekar, N.Deepika and D.Sujatha [2011] Association rule for classification of patients, *International Journal of Advanced Engineering Sciences and Technologies*, vol.11, (2), pp.253-257.
4. Z.H. Zhou, and Z.Q. Chen [2002] Hybrid Decision Tree, *Knowledge-Based Systems*, Vol. 15, pp. 515-528.
5. S. Sankaranarayanan and Pramananda T. Perumal [2014] A Predictive Approach for Diabetes Mellitus Disease through DataMiningTechnologies, *CPS-IEEE/iExplore*, 978-1-4799-2876 7/13$31.00©2013, IEEE, DOI10.1109/ WCCCT.2014.65, pp.231-233.
6. S. Sankaranarayanan and Pramananda T. perumal [2014] Diabetic prognosis through DataMining Methods and Techniques, *CPS-IEEE/iExplore*, Conference proceedings of International Conference of Intelligent Computing Applications: ICICA 2014, DOI March 2014

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

S. Sankaranarayanan and T. Pramananda Perumal [2014] Diabetic Diagnosis through Data Mining. *Indian Journal of Innovations and Developments,* Vol 3 (2), pp. 30-34.