ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
December 2012,
Vol. 5, No. (2):
Pgs. 215-221

Intelligent and Effective Diabetes Risk Prediction System Using Data Mining

KAWSAR AHMED[1], TASNUBA JESMIN[1], USHIN FATIMA[2], MD. MONIRUZZAMAN[2], ABDULLA-AL-EMRAN[2] and MD. ZAMILUR RAHMAN[1]

[1]Department of Information and Communication Technology,
Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902
[2]Department of Biotechnology and Genetic Engineering,
Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902

(Received: July 12, 2012; Accepted: August 04, 2012)

ABSTRACT

Diabetes is not only a disease but also responsible for occurring different kinds of diseases such as heart attack, kidney disease, blindness and renal failure. With respect to Bangladesh, Diabetes is a deadly, disabling and cost disease whose risk is increasing at alarming rate. The diagnosis of diabetes is a vital and tedious task. The detection of diabetes from some important risk factors is a multi-layered problem. Initially 400 diabetes and non-diabetes patients' data is collected from different diagnostic centre and data is pre-processed. After pre-processing data is clustered using K-means clustering algorithm for identifying relevant and non-relevant data to diabetes. Next significant frequent patterns are discovered using AprioriTid shown in Table 1 and Decision Tree algorithm shown in Table 2. Finally implement a system to predict diabetes which is easier, cost reducible and time saveable.

Key words: Data pre-processing, Data classification, AprioriTid algorithm, DT (Decision tree) algorithm, K-means clustering, Significant frequent pattern.

INTRODUCTION

Diabetes is one of deadly, metabolic and costly disease that increases blood sugar level. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases etc. If diabetes is uncontrolled then it increases blood glucose level more than 200mgl/dL which leads to micro and macro vascular disease complications[1]. The incidence of diabetes has soared worldwide in recent years and is expected to keep growing, with the greatest increase seen in metabolic forms of diabetes, notably type 2. This is blamed largely on the rise of obesity and the global spread of Western-style habits: physical inactivity along with a diet that is high in calories, processed carbohydrates and saturated fats and insufficient in fiber rich whole foods. However, other factors, such as environment may also be contributing, because cases of

autoimmune diabetes (type 1) are also becoming more common. The estimated number of people with diabetes has jumped from 30 million in 1985 to 150 million in 2000 and then to 246 million in 2007, according to the International Diabetes Federation. It expects this number to hit 380 million by 2025.According to World Health Organization there are more than one million people in this world who are suffering from diabetes. The prevalence of Type 2 Diabetes is increasing at an alarming rate in a developing country like Bangladesh in recent years[2]. Therefore The diagnosis of diabetes is a vital and tedious task. The detection of diabetes from some important risk factors for prevention of diabetes is a multilayered problem.

A widely recognized formal definition of data mining can be defined as "Data mining is the non- trivial extraction of implicit previously unknown and potentially useful information about data" [3]. Data mining has some fields to analysis of data such as classification, clustering, association rule etc. Now-a-days data mining has been used intensively and extensively by many organizations. In-healthcare, data mining is becoming increasingly popular[4]. Data mining provides the methodology and technology to analysis the useful information of data for decision making.

Data pre-processing is a tedious task of data mining. It mainly used for making analysis suitable and also making data suitable for clustering by deleting duplicate records and providing missing data according to past recorded data. The main benefits of data pre-processing minimizes memory.

Clustering is a process of dividing dataset into subgroups according to the unique feature. Clustering divided the dataset into relevant and non-relevant dataset to diabetes. AprioriTid[5] and Decision Tree algorithm[6] are mainly used to find out frequent patterns of dataset. Those algorithms are very simple and effective to find out frequent patterns. Frequent patterns are the sets of data that are frequently occurred into data warehouse. Significant frequent pattern is the set of data that are mostly responsible to diabetes. Using this significant pattern we implemented a prediction system for diabetes.

The main goal of our research is to develop a system that can be used by a person for testing his/her diabetes risk level.

Methodology

Data Collection

400 patients' data (200 diabetes patients and 200 non-diabetes patients) is obtained from different diagnostic centre. There are 200 male and 200 female patients whose age between 20 to 80 years old. From the previous studies 16 risk factors were considered for type 2 diabetes assessment in Bangladeshi population, which includes- age, gender, hereditary, previous health examination, use of anti-hypersensitive drugs, smoking, food habit, physical activity, BMI (Body Mass Index), waist circumference, mental trauma, uptake of red meat, hypertension, heart disease.

Data Pre-processing

Data pre-processing is a significant term of data mining. Making a suitable analysis and appropriate for clustering of collected data. This is the main concern of data pre-processing. Sometimes data warehouse is consisted with duplicate data and missing any values of data. Data pre-processing cleans the duplicates data and supplies the missing values according to the past recorded data. It also minimizes the memory and normalizes the values used to represent information in database.

Clustering of Collected Data

The process of partitioning and category of collected data into different subgroups where each groups have a unique feature is called clustering[7]. Clustering is another significant term of data mining. The clustering problem has been addressed in numerous contents besides being proven beneficial in many applications[8]. The goal of clustering is to classify objects or data into a number of categories or classes. The main benefits of clustering are that the data object is assigned to an unknown class and minimizes the memory. The K-means clustering[9] is a widely recognized clustering tool that is used for robotics, diseases and artificial intelligence application purposes[8]. Here k is a positive integer representing the number of clusters. The pre-processed data is clustered using the K-means clustering algorithm

with the value of k=2. This represents there is two clusters where one cluster contains relevant data to diabetes and another contains remaining data.

Discover Frequent Pattern

This is the most significant topics of data mining. It is considered as the major data mining problem that intends to find out the frequent items or patterns from the data warehouse[10]. There are different kinds of algorithms such as Apriori, AprioriTid, Decision Tree, FP-Tree whose aim is to mine interesting frequent patterns from databases like association rules, clusters, classifications and correlations etc.

After clustering, AprioiTid[5] and Decision Tree algorithms[6] is used to mine the frequent patterns. The AprioriTid and Decision Tree algorithms are the efficient algorithms of extracting the frequent patterns from databases.

1) L_1 = {large 1-itemsets};
2) C_1 = database D;
3) for (k = 2; Lk-1 ≠∅ ; k++) do begin
4) C_k = apriori-gen(L_{k-1});
5) $|C_k|$ = ∅;
6) forall entries t ∈ $|C_{k-1}|$ do begin
7) // determine candidate itemsets in C_k contained in the transaction with identifier t.TID
C_t = {c ∈ C_k | (c - c[k]) ∈ t.set-of-itemsets ^ (c - c[k-1]) ∈ t.set-of-itemsets};
8) forall candidates c ∈ Ct do
9) c.count++;
10) if (C_t≠∅;) then $|C_k|$ += < t.TID,C_t >;
11) end
12) L_k = {c ∈ C_k | c.count >= minsup}
13) end
14) Answer = ∪k Lk;

Input : II is a set of candidate attributes and S is a set of labeled instances

Output: A decision tree $\mathbf{T}$.

1. **If** ($\mathbf{S}$ is pure or empty) or ($\mathbf{\Pi}$ is empty) **Return** $\mathbf{T}$.
2. Compute $P_{\mathbf{S}}(c_i)$ on $\mathbf{S}$ for each class c_i.
3. **For** each attribute X in $\mathbf{\Pi}$, compute $IIG(\mathbf{S}, X)$ based on Equation 1 and 5.
4. Use the attribute X_{max} with the highest IIG for the root.
5. Partition $\mathbf{S}$ into disjoint subsets $\mathbf{S}_x$ using X_{max}.
6. **For** all values x of X_{max}
   - $\mathbf{T}_x$=$NT(\mathbf{\Pi} - X_{max}, \mathbf{S}_x)$
   - Add $\mathbf{T}_x$ as a child of X_{max}.
7. **Return** $\mathbf{T}$.

Pseudo code of AprioriTid & Decision Tree Algorithm

Significant Pattern Find-Out

After mining the frequent patterns using AprioriTid and Decision Tree algorithm, the weightage significant patterns are discovered by using the Equation (1) [8]

$$\mathbf{S}w(i)=\Sigma(Wi*Fi) \quad ...(1)$$

Where W_i is the weightage of each attribute and F_i represents number of frequency for each rule.

And significant Frequent Pattern is chosen by using the following Equation (2)

$$SFP=SW(n)\geq \varphi \text{ for all values of n} \quad ...(2)$$

Where SFP denotes significant frequent pattern and φ denotes significant weightage.

RESULTS

The experimental results are divided into two sections. One is significant frequent patterns discover and another is represents prediction tools to diabetes.

Result for Significant Frequent Pattern

Using data from data warehouse, the significant patterns are extracted for diabetes prediction. The collected data are pre-processed by deleting duplicate records and adding missing values. Then pre-processed data is clustered using K-means cluster algorithm with k=2. And finally significant frequent patterns are discovered using AprioriTid shown in Table 1 and Decision Tree algorithm shown in Table

Result for Prediction to Diabetes

Finally using the significant pattern the prediction tools to diabetes are implemented. Table 3 represents the frequent pattern parameters and their corresponding score and Figure-1 represents the risk level of diabetes which is implemented using Table 3.

Table 1: Significant Pattern and their corresponding Weightage value using AprioriTid algorithom

Significant Patterns	Weightage
Age-Hereditary-Sugar-Vegetables-Physical Activity-BMI-Waist-Red Meat	451.90
Age-Hereditary-Sugar-Vegetables -BMI-Waist-Red Meat	412.60
Age-Sugar-Vegetables-Physical Activity-BMI-Waist-Red Meat	391.30

Table 2: Significant Pattern and their corresponding Weightage value using Decision Tree algorithm

Significant Patterns	Weightage
Age-Hereditary-Sugar-Vegetables-Physical Activity-BMI-Waist-Red Meat	451.90
Age-Hereditary-Sugar-Vegetables-Physical Activity -BMI-Waist	412.30
Age-Sugar-Vegetables-Physical Activity-BMI-Waist-Red Meat	391.30

Table 3: Significant Pattern and their corresponding Weightage and Score

Parameters		Weightage	Score
Age		Age≤30	1
		30<Age≤50	3
		50<Age≤70	4
		Age>70	5
Relatives		No	1
		Grandparent, Uncle, Aunty	4
		Parents, Brother, Sister	6
Sugar		NO	1
		Yes	6
Vegetables		No	2
		Yes	1
Physical Activity		NO	3
		Yes	1
BMI		BMI≤24	1
		25≤BMI≤28	2
		BMI≤29	4
Red Meat		No	1
		Yes	3
	Men	Waist ≤88cm	1
Waist		89≤Waist≤98	4
		99≤Waist	5
	Women	Waist≤80cm	1
		81≤Waistd≤90	4
		91≤Waist	5

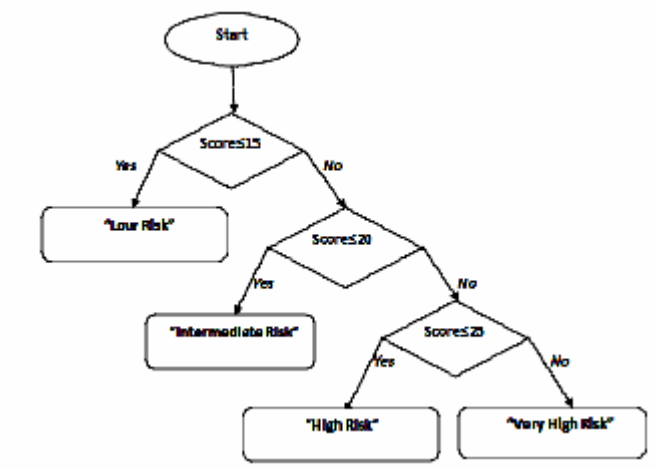

Fig. 1: Flow diagram of decision tree algorithm

Fig. 2: Diabetes prediction with low risk level

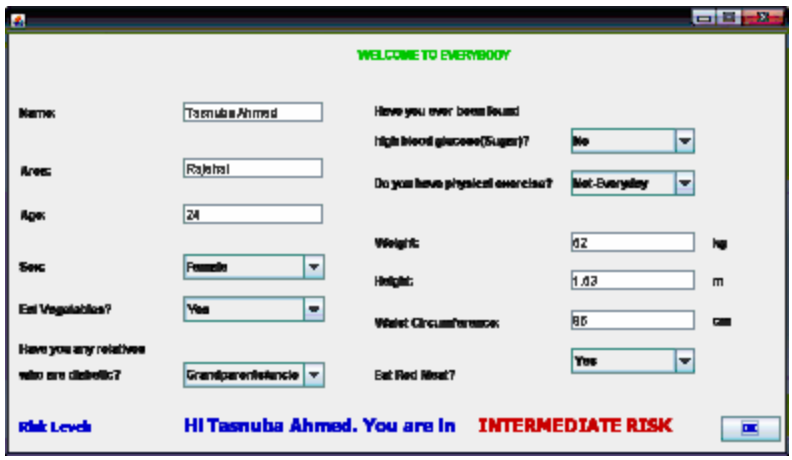

Fig. 3: Diabetes prediction with intermediate risk level

Fig. 4: Diabetes prediction with high risk level

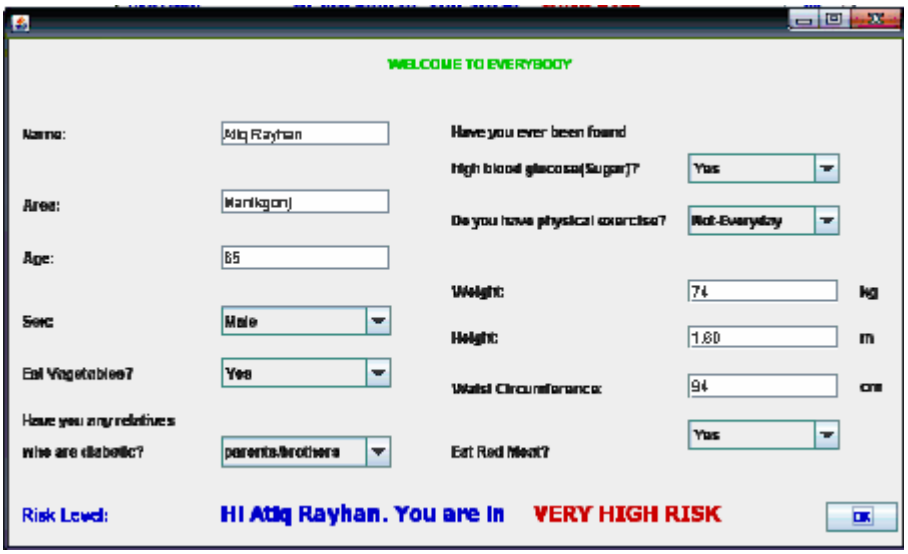

Fig. 5: Diabetes prediction with very high risk level

CONCLUTION

Millions of people in the Bangladesh and the world have diabetes. Most of them do not even know they have it. The ability to predict diabetes plays an important role in the diagnosis process. In this paper we have proposed an effective diabetes prediction system based on data mining. We have provided an efficient approach for the extraction of significant pattern from data warehouse for efficient prediction of diabetes. The proposed method is implemented using java. The proposed method can efficiently and successfully predict the diabetes.

REFERENCES

1. Manaswini Pradhan, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", *International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004)*, **2**(2): 303-311 (2011).
2. Unwin N, editors. IDF Diabetes Atlas. 4th ed. Brussels: *International Diabetes Federation* (2009).
3. Frawley, Knowledge Discovery in Databases: An Overview. The AAAI/MIT Press, Menlo Park, C.A (1996).
4. Hian Chye Koh, Data Mining Applications in Healthcare, *Journal of Healthcare Information Management.* **19**(2): .64-72.
5. Dr. Ilias Petrolias *"*Association rule tool An implementation of AprioriTID Algorithm*",* ID 2429851.
6. Jiang Su, " A Fast Decision Tree Learning Algorithm", Copyright ©c 2006, *American Association for Artificial Intelligence* (www.aaai.org).
7. Zakaria Nouir, "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering," *Journal of Communications*, **2**(6): 30-37 (2007).
8. Shantakumar B.Patil, **"**Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", *European Journal of Scientific Research* ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009 http://www.eurojournals.com/ejsr.htm
9. C. Ordonez, "Programming the K-Means Clustering Algorithm in SQL," *Proc. ACM Int'l Conf. Knowledge Discovery and Data Mining*, pp. 823-828 (2004).
10. Eric Li, "Optimization of Frequent Itemset Mining on Multiple-Core Processor", pp: 1275-1285 (2007).
11. American Diabetes Asscociation. Available: http://www.diabetes.org
12. T.Jayalakshmi, "A novel classification method for classification of diabetes mellitus using artificial neural networks". *International Conference on Data Storage and Data Engineering* (2010).
13. Muhammad Akmal Sapon , "Prediction of Diabetes by using Artificial Neural Network", *2011 International Conference on Circuits, System and Simulation* IPCSIT vol.7 (2011) © IACSIT Press, pp. 299-303 (2011).
14. Shantakumar B.Patil, **"**Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", *European Journal of Scientific Research* ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009 http://www.eurojournals.com/ejsr.htm
15. "Hospitalization for Heart Attack, Stroke, or Congestive Heart Failure among Persons with Diabetes", Special report: 2001 – 2003, New Mexico.
16. R. Agrawal, "Fast algorithms for mining association rules in large databases", *In Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, (1994).
17. Frank Lemke, "Medical data analysis using self-organizing data mining technologies," *Systems Analysis Modelling Simulation,* **43**(10): 1399 - 1408 (2003).
18. Huy Nguyen Anh Pham , "Prediction of Diabetes by Employing a New Data Mining Approach Which Balances Fitting and Generalization" , *R. Lee and H.-K. Kim (Eds.): Computer and Information Science*, SCI 131, pp. 11–26 (2008).
19. American Diabetes Association, http://www.diabetes.org/home.jsp (2007)
20. Pardha Repalli :"Prediction on Diabetes Using Data mining Approach".
21. Bishop, C. M.*, "*Pattern Recognition and Machine Learning, Springer*"*, 0-387-31073-8, New York (2006).