INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 1)
Available online at www.ijariit.com

Data Mining Techniques in Prediction of Risk Factors of Diabetes Mellitus

Omana .J
Prathyusha Engineering College, Tiruvallur, Tamil Nadu

Sujithra .S
sujithra.suji03@gmail.com
Prathyusha Engineering College, Tiruvallur, Tamil Nadu

Vishali .S
vishalisudharson@gmail.com
Prathyusha Engineering College, Tiruvallur, Tamil Nadu

Yuvashree .K
shreeyuvi@gmail.com
Prathyusha Engineering College, Tiruvallur, Tamil Nadu

ABSTRACT

Diabetes mellitus is a chronic disease, lifelong condition that affects the body's ability to use the energy found in food. The level of morbidity and mortality due to diabetes and its potential complications are enormous and pose significant healthcare burdens. It is a complex and time consuming task in detecting the risk of acquiring diabetes mellitus when a large amount of data is manually processed in the clinical environment. The objective of this paper is to simplify the process of analysing and detecting the risk of developing diabetes. Patients' details are gathered and stored in the form of Electronic medical record (EMR). Association rule mining and decision tree induction are applied to the records stored, in order to obtain the set of rules that are to be satisfied. C4.5 or Support vector machine is used to classify the data set accordingly and summarization techniques are used to summarize resultant possibility of acquiring diabetes.

Keywords: ***Data Mining, Mining Techniques, Association Rules, Support Vector Machine.***

1. INTRODUCTION

Diabetes is fast gaining the status of a potential epidemic in India with more than 62 million diabetic individuals currently diagnosed with the disease.[1,2] In 2000, India (31.7 million) topped the world with the highest number of people with diabetes mellitus followed by China (20.8 million) with the United States (17.7 million) in second and third place respectively. According to Wild et al. the prevalence of diabetes is predicted to double globally from 171 million in 2000 to 366 million in 2030 with a maximum increase in India. It is predicted that by 2030 diabetes mellitus may afflict up to 79.4 million individuals in India, while China (42.3 million) and the United States (30.3 million) will also see significant increases in those affected by the disease.[3,4] India currently faces an uncertain future in relation to the potential burden that diabetes may impose upon the country. Many influences affect the prevalence of disease throughout a country, and identification of those factors is necessary to facilitate change when facing health challenges.

Diabetes Mellitus is reaching potentially epidemic proportions in India. The level of morbidity and mortality due to diabetes and its potential complications are enormous and pose significant healthcare burdens on both family and society. In india, the steady migration of people from rural to urban areas, the economic boom, and a corresponding change in life style are all affecting the level of diabetes.

The propose extensions to incorporate the risk of diabetes into the process of finding an optimal summary. Association rule mining and decision tree induction are carried out in order to obtain the set of rules that are to be satisfied. C4.5 or Support vector machine and Bayesian classification are used to classify the data set accordingly and summarization techniques are used to summarize the obtained resultant data. The output can be visualized.

2. RELATED WORKS

2.1Extending association rule summarization techniques to assess the risk of diabetes mellitus.
Gyorgy J. Simon,¨ *Member, IEEE,* Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, M. Regina Castro and Peter W. Li

To apply association rule mining to electronic medical records (EMR) to discover sets of risk factors and their corresponding subpopulations that represent patients at particularly high risk of developing diabetes.

2.2 Utilization of data mining techniques for diagnosis of diabetes mellitus - a case study
Thirumal P.C. and Nagarajan N. Department of IT, Coimbatore Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

The principle of this study is to apply various data mining techniques which are noteworthy to prediction of diabetes mellitus and extract hidden patterns from the PIMA Indian diabetes dataset available at UCI Machine Learning Repository.

2.3 Predicting relative risk for diabetes mellitus using association summarization techniques
K. Sinduja1 & N. Saravanan. M. Tech Student and Associate professor, Department of Information Technology, K.S.R College of Engineering Tamil Nadu, India.

It aims to apply association rule mining to electronic medical records (EMR) to detect sets of risk factors and their corresponding subpopulations of patients.

2.4 Centers for Disease Control and Prevention.
National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011.

National Diabetes Fact Sheet uses both fasting glucose and haemoglobin A1c (A1c) levels to derive estimates for undiagnosed diabetes and prediabetes. These tests were chosen because they are most frequently used in clinical practice.

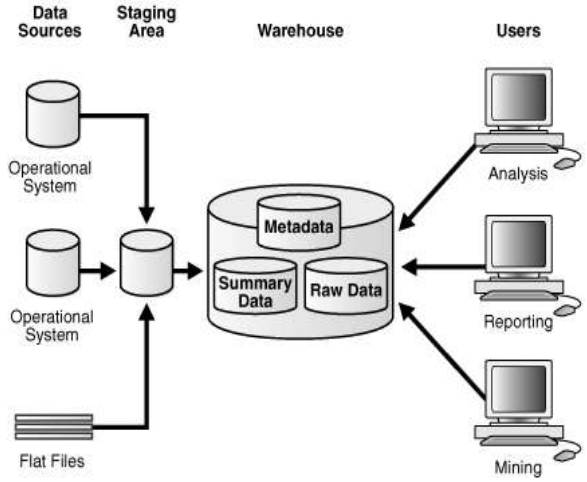
2.5 IDF Diabetes Atlas. The Economic Impacts of Diabetes.
Diabetes imposes a large economic burden on the national healthcare system. Healthcare expenditures on diabetes will account for 11.6% of the total healthcare expenditure in the world in 2010. About 95% of the countries covered in this report will spend 5% or more, and about 80% of the countries will spend between 5% and 13% of their total healthcare dollars on diabetes.

2.6 Miroslav Marinov, M.S., Abu Saleh Mohammad Mosa, M.S., Illhoi Yoo, Ph.D., And Suzanne Austin Boren, Ph.D., MHA
The objective of the study is to conduct a systematic review of the application of data mining techniques in the field of diabetes research. Using data mining to deal with the avalanche of clinical data collected from patients and generated from the research and management of diabetes is a valuable asset that can help researchers and clinicians provide better health care for the patients affected by this modern society disease. This is one more confirmation that data mining in biomedicine has a good future and will be used more and more in the area of diabetes in particular.

3. BACKGROUND

Data mining refers to extracting or ―mining‖ knowledge from large amounts of data.
Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, another popularly used term, Knowledge Discovery from Data, or KDD.

SYSTEM ARCHITECTURE



3.1 ASSOCIATION RULE

Frequent pattern mining searches for recurring relationships in a given data set. It introduces the enhanced frequent pattern mining for the discovery of interesting associations and correlations between itemsets in transactional and relational databases
*Implication:

X ➔ Y where X, Y $\subseteq$ I and X $\cap$Y = $\varnothing$;
*Support of AR (s)
X ➔ Y:

- ✓ Percentage of transactions that contain X $\cup$ Y
- ✓ The probability that a transaction contains X$\cup$Y.

*Confidence of AR (a)
X ➔ Y:

- ✓ Ratio of number of transactions that contain X $\cup$ Y to the number that contains X
- ✓ The conditional probability that a transaction having X also contains *Y*.

Once the frequent item sets from transactions in a database *D* have been found, it is straightforward to generate strong association rules from them (where *strong* association rules satisfy both minimum support and minimum confidence). This can be done using for confidence, which we show again here for completeness:

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

3.2 DECISION TREE INDUCTION

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (nonleaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or *terminal node*) holds a class label. The topmost node in a tree is the root node.

"How are decision trees used for classification?" Given a tuple, ***X***, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules.

3.3 C4.5 (GAIN RATIO)

Information gain measure is biased towards attributes with a large number of values. C4.5 (a successor of ID3) uses gain ratio to overcome the problem (normalization to information gain)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

The attribute with the maximum gain ratio is selected as the splitting attribute

3.4 SUPPORT VECTOR MACHINES

Support Vector Machines, a promising new method for the classification of both linear and nonlinear data. In a nutshell, a support vector machine is an algorithm that works as follows. It uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane with an appropriate nonlinear mapping to a sufficiently high dimension; data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using *support vectors* ("essential" training tuples) and *margins* (defined by the support vectors).

3.5 BAYES CLASSIFICATION

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
Baye's Theorem

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities:

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

Where X is data tuple and H is some hypothesis.
According to Bayes' Theorem,

P(H/X)= P(X/H)P(H) / P(X)

Bayesian Belief Network

Bayesian Belief Networks specify joint conditional probability distributions. They are also known as Belief Networks, Bayesian Networks, or Probabilistic Networks.

- A Belief Network allows class conditional independencies to be defined between subsets of variables.
- It provides a graphical model of causal relationship on which learning can be performed.
- We can use a trained Bayesian Network for classification.

There are two components that define a Bayesian Belief Network:

- Directed acyclic graph
- A set of conditional probability tables

4. METHODOLOGY

Initially, the system is to classify the system into two categories such that normal and abnormal ones. Normal are those that have no risk of developing diabetes whereas the abnormal ones are those has the probability of risk in acquiring diabetes.

4.1RISK ANALYSIS USING ASSOCIATION RULE

4.1.1 Data Pre Processing

1. Real world Data are generally

- ✓ Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- ✓ Noisy: containing errors or outliers
- ✓ Inconsistent: containing discrepancies in codes or names

2. Tasks in Data Preprocessing

- ✓ Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- ✓ Data integration: using multiple databases, data cubes, or files.
- ✓ Data transformation: normalization and aggregation.
- ✓ Data reduction: reducing the volume but producing the same or similar analytical results.
- ✓ Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

4.1.2 Data Cleaning

1. Fill in missing values (attribute or class value):
    - Ignore the tuple: usually done when the class label is missing.
    - Use the attribute mean (or majority nominal value to fill in the missing value.
    - Use the attribute mean (or majority nominal value) for all samples belonging to the same class.
    - Predict the missing value by using a learning algorithm: consider the attribute with the missing value as a dependent (class) variable and run a learning algorithm (usually Bayes or decision tree) to predict the missing value.

2. Identify outliers and smooth out noisy data:
    - **Binning**
        - Sort the attribute values and partition them into bins (see "Unsupervised discretization" below);
        - Then smooth by bin means, bin median, or bin boundaries.
    - **Clustering**: group values in clusters and then detect and remove outliers (automatic or manual)
    - **Regression**: smooth by fitting the data into regression functions.
3. Correct inconsistent data: use domain knowledge or expert decision.

In Association rule mining, Frequent pattern mining searches for recurring relationships in a given data set. It results in the enhanced frequent pattern mining for the discovery of interesting associations and correlations between itemsets in transactional and relational databases

4.2 RISK ANALYSIS USING BAYESIAN CLASSIFICATION

4.2.1 Bayesian network classifier

Let U = {x1,... , xn}, n ~ 1 be a set of variables. A Bayesian network B over a set of variables U is a network structure BS, which is a directed acyclic graph (DAG) over U and a set of probability tables BP = {p(u|pa(u))|u 2 U} where pa(u) is the set of parents of u in BS. A Bayesian network represents a probability distributions P(U) = Q u2U p(u|pa(u)).

Starting with a complete undirected graph, we try to find conditional independencies in the data. For each pair of nodes x, y, we consider sets Z starting with cardinality 0, then 1 up to a user defined maximum. Further-more, the set Z is a subset of nodes that are neighbors of both x and y. If an independency is identified, the edge between x and y is removed from the skeleton. The first step in directing arrows is to check for every configuration x - -z - - y where x and y not connected in the skeleton whether z is in the set Z of variables that justified removing the link between x and y (cached in the first step). If z is not in Z, we can assign direction x → z ←y. Finally, a set of graphical rules is applied to direct the remaining arrows.

5. RESULT

In this paper, the Bayes classification, Decision tree, Association rule and c4.5 algorithms are described in terms of diabetes risk prediction performance improvement. The efficiency of algorithms described in this paper is determined based on survey. From the survey, it was found that the support vector machine is capable of producing an accuracy of 73.34% before and 77.73% after Pre-processing. It was also observed that the accuracy level increases to 79.687% while processed with Boosting algorithm like Adaboost. Similarly, the J48 a decision tree constructing technique can reach an accuracy of 73.82% before and 86.46% after Pre-processing. The Naïve Bayes can achieve 78.1% of accuracy without the help of any boosting algorithms. Finally, the association techniques are used to identify some interesting rule present within the dataset.

6. CONCLUSION

The largest economic burden caused by diabetes is the monetary value associated with disability and loss of life as a result of the disease itself and its related complications, including heart, kidney, and eye and foot disease. Economists have used different methods to value disability and loss of life associated with diseases and the most appropriate method is still under debate. No matter what method is used, it is very likely the economic burden that is measured by the monetary value associated with this disability and loss of life would be far larger than the estimated economic burden using measures described above. Fortunately, the economic burden of diabetes can be reduced by implementing many inexpensive, easy-to-use interventions, and most of the interventions are cost-effective or cost-saving, even in the poorest countries. Tragically, these interventions are not widely used in poor and middle income countries. More resources should be invested to deliver these cost- effective interventions, in particular to those in the developing countries where the great majority of persons with diabetes live.

7. REFERENCES

[1] Pedro J. Caraballo, M. Regina Castro, Stephen S. Cha, Peter W. Li, and Gyorgy J. Simon. Use of association rule mining to assess diabetes risk in patients with impaired fasting glucose. In AMIA Annual Symposium, 2011.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In VLDB Conference, 1994.

[3] Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. In Knowledge Discovery and Data Mining, 1999.

[4] Varun Chandola and Vipin Kumar. Summarization – compressing data into an informative representation. Knowledge and Information Systems, 2006.

[5] Gary S Collins, Susan Mallett, Omar Omar, and LyMee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. BMC Medicine, 2011.

[6] Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England Journal of Medicine, 346(6), 2002.

[7] Gang Fang, Majda Haznadar, Wen Wang, Haoyu Yu, Michael Steinbach, Timothy R Church, William S Oetting, Brian Van Ness, and Vipin Kumar. High-order snp combinations associated with complex diseases: efficient discovery, statistical power, and functional interactions. PLoS One, 7(4):e33531, 2012.

[8] Mohammad Al Hasan. Summarization in pattern mining. In Encyclopaedia of Data Warehousing and Mining, (2nd Ed). Information Science Reference, 2008.

[9] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In American Association for Artificial Intelligence (AAAI), 1997. Paper ID: SUB152157 1121 International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438 Volume 4 Issue 3, March 2015 www.ijsr.net Licensed under Creative Commons Attribution CC BY

[10]Terry M. Therneau and Patricia M. Grambsch. Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health. Springer, 2010.

[11]Ruoming Jin, Muad Abu-Ata, Yang Xiang, and Ning Ruan. Effective and efficient itemset pattern summarization: Regression based approach. In ACM International Conference on Knowledge Discovery and Data Mining (KDD), 2008.

[12]Aysel Ozgur, Pang-Ning Tan, and Vipin Kumar. RBA: An integrated framework for regression based on association rules. In SIAM International.

[13]Bing Liu, Wynne Hsu, and Yiming Ma. Integrating classification and association rule mining. In ACM International Conference on Knowledge.

[14]Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In SIAM International Conference on Data Mining (SDM), 2003.

[15]Peter W. Wilson, James B. Meigs, Lisa Sullivan, Caroline S. Fox, David M. Nathan, and Ralph B. D"Agostino. Prediction of incident diabetes mellitus in middle-aged adults–the Framingham offspring study. Archives of Internal Medicine, 167, 2007.

