Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Survey on Predictive Analysis of Diabetes in Young and Old Patients

Komal Agicha, Priyanka Bhatia, Neha Badlani, Ashutosh Menghrajani, Abha Tewari
Computer Engineering, VESIT, Maharashtra,
India

Abstract— This project concentrates upon predictive analysis of diabetic patients on the basis of age and gender using various data mining techniques like classification algorithms. Also this tool recognizes possibility of diabetes for particular person. R studio will be used as it is a free and open source integrated development environment (IDE) for R, a programming language used for statistical computing and graphics. We conclude that analysis will be provided to the physician and also prediction of the possibility of diabetes provided by our tool would help physician to accordingly provide treatment plans accordingly.

Keywords— Data Mining; Classification Technique; Bayesian network; Decision Induction Algorithm; R Studio

I. INTRODUCTION

The term Data Mining is one that is used frequently in the research world, but it is often misunderstood by many people. Sometimes people misuse the term to mean any kind of extraction of data or data processing. However, data mining is so much more than simple data analysis. According to Doug Alexander at the University of Texas, data mining is, "the computer-assisted process of digging through and analysing enormous sets of data and then extracting the meaning of the data.Today most of the people are affected by diabetes and some of them even don't know that they are having such chronic disease. Moreover there should be analytics done to recognize the count of people affected by diabetes on analysing data records. As data mining is a computer assisted process of digging through data record, it would be easy for a physician to compare new patient record with previous and provide treatment plans. Therefore it is important to design a tool which automates the manual work on mathematical models used for analysis. Diabetes is a disease characterized by abnormal metabolism, most notably hyperglycaemia, and an associated heightened risk for relatively specific long-term complications affecting the eyes, kidney, and nervous system.Today most of the people are affected by diabetes and some of them even don't know that they are having such chronic disease. So through this project people will become aware about the risk factors causing diabetes and the treatments/prevention to be taken.
Data analysis technique used to predict the diabetes for people of different age groups by analysing their symptoms. The purpose of this analysis is to allow the physicians to provide treatment to people according to their age and gender and to know the approximate number of people of different age groups affected by diabetes.

II. RELATED WORK

1. Effective hyper sensitive data using data mining in Saudi Arabia

In this investigation, the data sets of NCD were employed, The data set was analysed on the real time report from a particular hospital in Saudi Arabia in 2005.The tool used was Oracle Data Miner(ODM) for predicting treatment plans. Treatment plans varied from age to age and also the bifurcation was provided considering various symptoms of diabetes. Reports were generated on analysis and also treatment plans were proposed by this tool.

2. A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies

In this article we studied about their survey on forecast of Diabetes Mellitus in humans. Patient records were checked and current analysis of human prone to diabetes for a particular range of years was provided to physician. Therefore physician would propose treatment plans accordingly. Also health programs were proposed in this survey.

III. METHODOLOGIES

A. Data Mining and Architecture:

- Knowledge base: Knowledge Base is a simple guide to evaluate the resulting pattern after data mining. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).
- Data mining engine: It is an essential component of data mining architecture where the relevant data when passed to data warehouse server is further given to data mining engine for classification.

© 2015, IJARCSSE All Rights Reserved

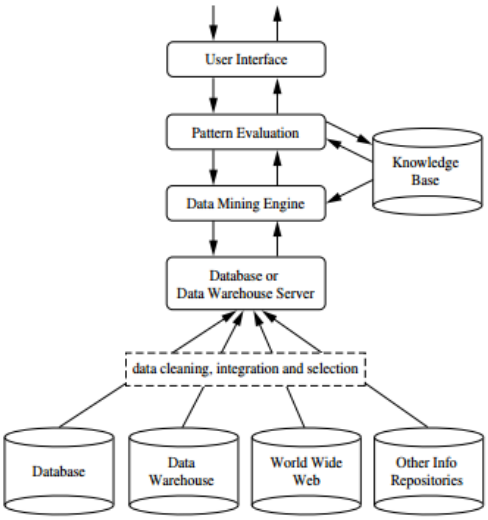

Figure-1

- Pattern evaluation module: This component is mainly used for evaluating efficient patterns for a particular data mining tool. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.
- User interface: This module involves communication between authorized users and the data mining tool, thus giving appropriate privileges to the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

B. Techniques:

1. Classification Technique:

Data classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analysing or "learning from" a training set made up of database tuples and their associated class labels.

A tuple, X, is represented by an n-dimensional attribute vector, X = (x1, x2,..., xn), depicting n measurements made on the tuple from n database attributes, respectively, A1, A2,..., An Each tuple, X, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute. The class label attribute is discrete-valued and unordered. It is categorical in that each value serves as a category or class. The individual tuples making up the training set are referred to as training tuples and are selected from the database under analysis. In the context of classification, data tuples can be referred to as samples, examples, instances, data points, or objects. Because the class label of each training tuple is provided, this step is also known as supervised learning (i.e., the learning of the classifier is "supervised" in that it is told).

C. Algorithms to be used:

1. Classification by Decision Tree Induction:

Decision tree induction is the learning of decision trees from class-labeled tuples. A decision tree is a flowchart-like tree structure, where each internal node is a test on each attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The top node is the root node. Some decision tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), whereas others can produce non binary trees. "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.

Algorithm:

1. Generate decision tree.
2. Generate a decision tree from the training tuples of data partition D.
3. Input: Data partition, D, which is a set of training tuples and their associated class labels;
4. attribute list, the set of candidate attributes;
5. Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.
6. Output: A decision tree. Method:

© 2015, IJARCSSE All Rights Reserved

7. create a node N;
8. if tuples in D are all of the same class, C then
9. return N as a leaf node labeled with the class C; (4) if attribute list is empty then
10. return N as a leaf node labeled with the majority class in D; // majority voting
11. apply Attribute selection method(D, attribute list) to find the "best" splitting criterion;
12. label node N with splitting criterion;
13. if splitting attribute is discrete-valued and multiway splits allowed then // not restricted to binary trees
14. attribute list ← attribute list − splitting attribute; // remove splitting attribute
15. for each outcome j of splitting criterion // partition the tuples and grow subtrees for each partition
16. let Dj be the set of data tuples in D satisfying outcome j; // a partition
17. if Dj is empty then
18. attach a leaf labeled with the majority class in D to node N;
19. else attach the node returned by Generate decision tree(Dj , attribute list) to node N; end for
20. return N;

2. Classification By Bayesian Network:

Bayesian classification is based on Bayes' theorem. Let ***X*** be a data tuple. In Bayesian terms, ***X*** is considered "evidence." As usual, it is described by measurements made on a set of *n* attributes. Let *H* be some hypothesis, such as that the data tuple ***X*** belongs to a specified class *C*. For classification problems, we want to determine $P(H \mid \boldsymbol{X})$, the probability that the hypothesis *H* holds given the "evidence" or observed data tuple ***X***. In other words, we are looking for the probability that tuple ***X*** belongs to class *C*, given that we know the attribute description of ***X***.

$P(H \mid \boldsymbol{X})$ is the posterior probability, or *a posteriori probability*, of *H* conditioned on ***X***. Similarly, $P(\boldsymbol{X}j \mid H)$ is the posterior probability of ***X*** conditioned on *H*.

$P(H)$, $P(\boldsymbol{X}j \mid H)$, and $P(\boldsymbol{X})$ may be estimated from the given data, as we shall see below. Bayes' theorem is useful in that it provides

a way of calculating the posterior probability, $P(H \mid \boldsymbol{X})$, from $P(H)$, $P(\boldsymbol{X}j \mid H)$, and $P(\boldsymbol{X})$.

Bayes' theorem is

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let *D* be a training set of tuples and their associated class labels. As usual, each tuple is represented by an *n*-dimensional attribute vector,

$\boldsymbol{X} = (x1, x2, : : : , xn)$, depicting *n* measurements made on the tuple from *n* attributes, respectively, *A*1, *A*2, : : : , *An*.

2. Suppose that there are *m* classes, *C*1, *C*2,: : : , *Cm*. Given a tuple, ***X***, the classifier will predict that ***X*** belongs to the class having the highest posterior probability, conditioned on ***X***. That is, the naïve Bayesian classifier predicts that tuple ***X*** belongs to the class *Ci* if and only if

$P(Ci \mid \boldsymbol{X}) > P(Cj \mid \boldsymbol{X})$ for $1 \leq j \leq m; j \neq i;$

Thus we maximize $P(Ci \mid \boldsymbol{X})$. The class *Ci* for which $P(Ci \mid \boldsymbol{X})$ is maximized is called the *Maximum posteriori hypothesis*. By Bayes' theorem,

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

3. As $P(\boldsymbol{X})$ is constant for all classes, only $P(\boldsymbol{X} \mid Ci)P(Ci)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C1) = P(C2) = \ldots.. = P(Cm)$, and we would therefore maximize $P(\boldsymbol{X} \mid Ci)$. Otherwise, we maximize $P(\boldsymbol{X} \mid Ci)P(Ci)$. Note that the class prior probabilities may be estimated by $P(Ci) = \mid Ci,D \mid / \mid Dj \mid$,where $\mid Ci,D$ is the number of training tuples of class *Ci* in *D*.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(\boldsymbol{X} \mid Ci)$. In order to reduce computation in evaluating $P(\boldsymbol{X} \mid Ci)$, the naive assumption of class conditional independence is made. This presumes that he values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$P(\boldsymbol{X} \mid Ci) = \prod_{k=1}^{n} P(xk \mid Ci) = P(x1 \mid Ci) * P(x2 \mid Ci) * \ldots\ldots\ldots P(xn \mid Ci).$$

We can easily estimate the probabilities $P(x1 \mid Ci)$, $P(x2 \mid Ci)$, ……., $P(xn \mid Ci)$ from the training tuples. Recall that here *xk* refers to the value of attribute *Ak* for tuple ***X***. For each attribute, we look at whether the attribute is categorical or continuous-valued. For instance, to compute $P(\boldsymbol{X} \mid Ci)$, we consider the following:

(a) If *Ak*is categorical, then $P(xk \mid Ci)$ is the number of tuples of class *Ci* in *D* having the value *xk* for *Ak*, divided by $\mid Ci,D \mid$, the number of tuples of class *Ci* in *D*.

(b) If *Ak* is continuous-valued, then we need to do a bit more work, but the calculation is pretty straightforward. A continuous-valued attribute is typically assumed to have a Gaussian distribution with a mean μ and standard deviation s, defined by

© 2015, IJARCSSE All Rights Reserved

$$g(x, \mu, \sigma) = 1/\sqrt{2\pi}\sigma e^{-((x-\mu)(x-\mu))/2\sigma}$$

so that $P(xk \mid Ci) = g(xk, \mu Ci, \sigma Ci)$.

D. Diagrammatic Representation:

1.1 Flow chart diagram:

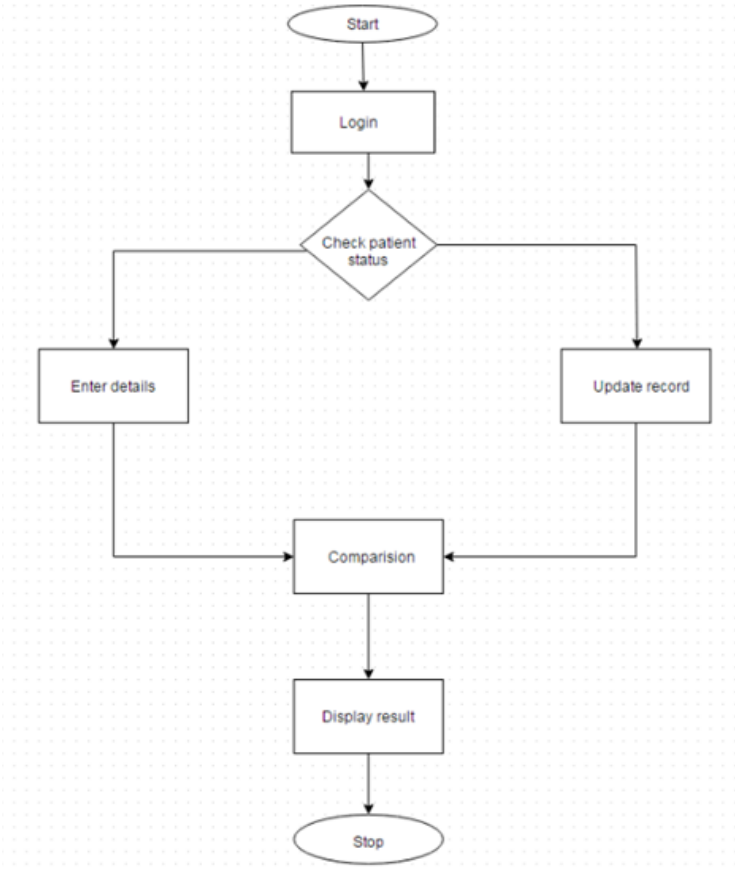

Figure-2

Flow charts are diagrammatic representation which help in representing the working of various algorithms or techniques used, functions of different modules present in the system in the form of blocks. They show the flow of how modules are interrelated to each other in the form of arrows. Thus raising a solution model of the system. In our system the physician has to first login, then he will check patient's status and if the patient is new, he will accordingly enter details. If the patient is old, his new detail will be updated. The details entered would be compared with the current analysis and results will be displayed whether the patient has diabetes or not.

1.2 Use case Diagrams:

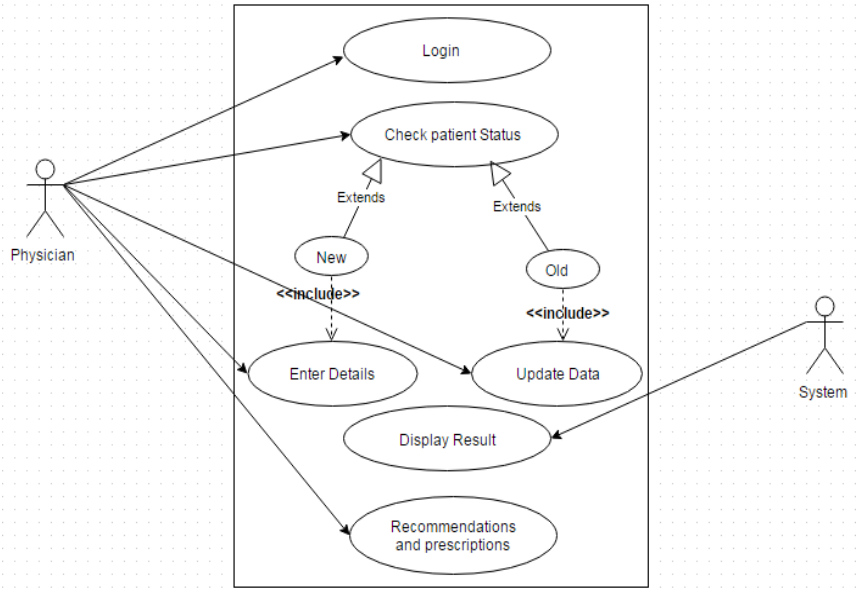

Figure-3

© 2015, IJARCSSE All Rights Reserved

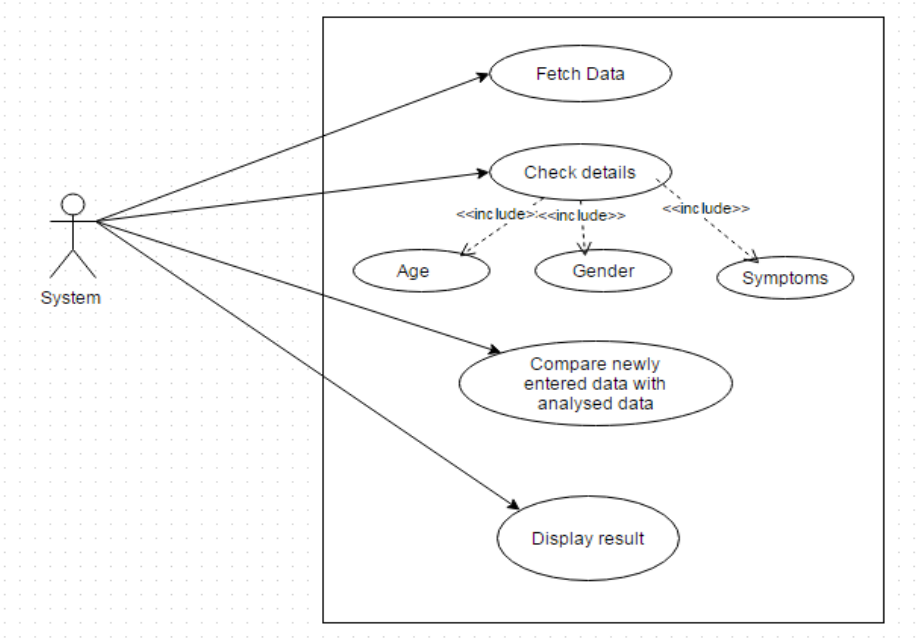

Figure-4

Use Case Diagrams helps in showing the dynamic behaviour of the system. Dynamic behaviour means how system works when it is in running state. They are used to gather requirements and also get the outside view of a system. As we can see in figure 3, there are various use cases like login, patient status, enter details, update data, display results and recommendations and perceptions. Here the physician is an actor which interacts with all the modules. Figure 4 shows the working of our data mining tool.

E. Tool to be used:

R Studio:

R Studio is a free and open source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.R Studio is available in two editions: R Studio Desktop, where the program is run locally as a regular desktop application; and R Studio Server, which allows accessing R Studio using a browser while it is running on a remote Linux server.

IV. CONCLUSIONS

The prevalence of diabetes is increasing among patients. Public health awareness of simple measures such as low sugar diet, exercise, and avoiding obesity should be promoted by health care providers. In this study, predictions on the effectiveness of different treatment methods for different age groups would be elucidated. Diet control, weight reduction, exercise and smoking cessation are mutually beneficial to each other for the treatment of diabetes. Thus on the basis of analytics done by our tool, probability of occurrence of diabetes and treatment would be provided.

ACKNOWLEDGMENT

The heading of the Acknowledgment section and the References section must not be numbered.Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

REFERENCES

[1] Effective hyper sensitive data using data mining in Saudi Arabia http://www.ncbi.nlm.nih.gov/pubmed/20978930

[2] Zhaohui Wu College of Computer Science, Zhejiang University,China, From Big Data to DataScience: A Multi-disciplinary Perspective

[3] Almazyad, A.S., Ahamad, M.G., Siddiqui, M.K., Almazyad, A.S., 2010. Effective hypertensive treatment using data mining in Saudi Arabia. Journal of Clinical Monitoring and Computing 24 (6), 391–401. http://dx.doi.org/10.1007/s10877-010-9260-2, Springers.

[4] V. H. Bhat, P. G. Rao, and P. D. Shenoy, "An Efficient Prediction Model for Diabetic Database Using SoftComputing Techniques," Architecture, Springer-Verlag Berlin Heidelberg, pp. 328-335, 2009..

[5] Nishchol Mishra, Dr.Sanjay Silakari, "Predictive Analytics: A Survey, Trends, Applications, Oppurtunities & Challenges", International Journal of Computer Science and Information Technologies, vol. 3(3), 4434- 4438 4434, 2012.

© 2015, IJARCSSE All Rights Reserved

[6] Promises and Challenges of Big Data Computing in Health Sciences http://www.journals.elsevier.com/big-data-research

[7] Andre W. Kushniruk, "Predictive Analytics and Forecasting in Health Care: Integrating Analytics with Electronic Health Records", SAS Institute Inc, 2008.

[8] Wullianallur Raghupathi, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential",Health Information Science and Systems, vol. 2(3) pp. 2-10, 2014.

[9] Health Monitoring System by Prognotive Computing using Big Data Analytics http://www.sciencedirect.com

[10] Journal of Diabetes & Metabolism , ISSN:2155-6156

© 2015, IJARCSSE All Rights Reserved