EXCELLENT PUBLISHERS

International Journal of Current Research and Academic Review

ISSN: 2347-3215 Volume 2 Number 10 (October-2014) pp. 91-98

www.ijcrar.com

Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes

Srideivanai Nagarajan[1*], R.M.Chandrasekaran[2], and P.Ramasubramanian[3]

[1]P.G. Dept.of Information Technology, Kumararani Meena Muthiah College, Chennai, India
[2]Department of Computer Science & Engineering, Annamalai University, Chidambaram, India
[3]Ranipet Engineering College, Ranipet, India
Corresponding author

KEYWORDS

Data Mining, Gestational Diabetes, Classification, Prediction

A B S T R A C T

During pregnancy when a woman has high blood sugar (glucose) level, who does not have diabetes before pregnancy is said to be gestational diabetes. According to the recently announced diabetes criteria, it is found that around18% of pregnant women have gestational diabetes. Data Mining is a field of computer science which is used to discover new patterns for large data sets. Classification is an important task in data mining. In various areas of medicine, data mining helped in improving the results along with other methodologies. The aim of this paper is to improve the diagnosis of gestational diabetes with the application of data mining techniques. Also, this paper analyses the performance of supervised learning algorithms viz ID3, Naïve Bayes, C4.5 and Random Tree. Experimental results have been proved that Random Tree serves to be the best one with highest accuracy and least error rate.

Introduction

Gestational Diabetes

When pregnant women who never had diabetes before but when she has high blood sugar level during pregnancy, is said to be gestational diabetes. According to the recently announced diabetes criteria it is found that around 18% of pregnant women have diabetes. At the time of pregnancy, the levels of some hormones made in the placenta (the organ which connects the baby by the umbilical cord to the uterus) will get increased. This will help to move nutrients from mother to the growing baby. Placenta produces other hormones which help in preventing the mother from developing low blood sugar value. They will oppose the role of insulin. During subsequent weeks of pregnancy**,** these hormones will generate high blood sugar levels. In order to reduce high blood sugar levels, the body generates more insulin to furnish glucose into the cells which can be used for energy.

During pregnancy the mother's pancreas produces more insulin (almost thrice the normal value) in order to defeat the result of the pregnancy hormones in blood sugar levels. If the pancreas fails to produce sufficient insulin in order to overcome the consequence of the increased hormones during pregnancy, there will be an increase in the blood sugar levels. This will result in gestational diabetes.

Gestational diabetes may affect the developing fetus throughout the pregnancy. During earlier stage of pregnancy, a mother's diabetes can cause birth defects to the fetus and it may result in an increased rate of miscarriage. Several birth defects that occur affect major organs like brain and heart.

From the fourth to ninth month of pregnancy, the diabetes of a mother can lead to excess-nutrition and over growth to the baby. If the size of the baby is large it increases risks at the time of delivery.
Due to over-nutrition of the fetal and because of hyperinsulinemia the blood sugar level of the baby may hit a very low value after birth, because the fetal won't be able to receive the high blood sugar from the mother as before.

GDM (Gestational Diabetes Mellitus) affects at least 7 percent [10] and possibly as many as 18 percent of pregnancies in the United States. As reported in [11], in India GDM affects nearly 17% of pregnant women. However, with proper treatment, a gestational diabetic mother can give birth to a healthy baby despite of having diabetes [12].

Classification Algorithm

Classification is a supervised machine learning technique which assigns labels or classes to different objects or groups [13]. Classification is a two step process: First step is model construction, which is used to analyze the training data set of a database. The second step is model usage; here the constructed model is used for classification. The accuracy of the classification is estimated according to the percentage of test samples or test data set that are correctly classified.

Significance of the study

This paper focuses on how data mining techniques are applied to predict the risk for gestational diabetes in the data set collected from pregnant women. The study of related works is presented in section II. The methods and materials which include description about dataset, system design supervised learning algorithms are discussed in section III. The section IV describes the experimental results, Section V discusses the limitations and finally section VI gives the conclusion and future work of this paper.

Literature Survey

Stephen E. Brossette [1] proposed the Data Mining Surveillance System which tracks many association rules with three datasets. This system automatically identifies new, unexpected and potentially interesting patterns in hospital infection control and public health surveillance data.

Jie Gao and Jorg Denzinger [2] proposed a data mining approach called CoLe, for early diabetes detection. CoLe is a multi-agent system framework with multiple miner agents and a combination agent. The main goal of CoLe is to get hybrid knowledge that can describe given data from multiple aspects.

B.M. patil [3] used Apriori association algorithms to classify type-2 diabetes. The author generated 4 association rules for the dataset with class value = "yes" for diabetes and also generated 10 association rules for the class value = "No". The author also used preprocessing techniques to improve the quality of the dataset.

Huda Yasin [4] devised a method for investigating factors which have higher prevalence for the risk of hepatitis C virus. The author compared the proposed method with nearly 20 classification techniques which includes Naïve bayes, GRNN, CART etc., and proved the proposed system is having the highest accuracy rate of 89.6%.

A comparative study on Breast Cancer prediction was done by J.Padmavathi [5]. The Author compared the result of Radial Basis Function and Multilayer Perceptron neural network Methods with Logistic Regression method. The author concluded that Radial Basis Function, neural network method predicts the disease with 97% accuracy.

Duarte Ferreira [6] used various classification algorithms like j48, simple CART, SMO and simple logistics, Bayes net, Naïve bayes to diagnose the neonatal jaundice and proved that simple logistics is the best model for the dataset.

Daveedu raju Adidela [7] used Fuzzy Id3 method to predict Diabetes. The author uses the system for predicting the disease from huge sample data set as it initially clusters the data and applies the classification on clusters. The author proposes a hybrid classification system which builds from Em algorithm for clustering and fuzzy ID3 algorithm to obtain decision tree for individual clusters.

Sonu Kumari [8] applied Neural Network system for diagnosing diabetes mellitus and proved that the results are 92.8% accurate.

S.Shyed Shajahaan [9] applied various data mining algorithms like ID3, C4.5, CART, Random Tree and Naïve Bayes techniques to model breast cancer data and compared the results. The author concluded that Random tree gives an accuracy of 100% and error rate 0f 0.000.

In this paper, a comparative study is done with various classification algorithms like ID3, Naïve Bayes, C4.5 and Random Tree which used to predict the risk for gestational diabetes.

Methodology

Data Mining Algorithms used

Classification is an important task in data mining. There are several classification algorithms available to mine the data and these algorithms are used in several disciplines. The classification techniques also play a vital role in analyzing the data and to predict information [14]. Some of the classification algorithms used to predict gestational diabetes - ID3, Naïve Bayes, C4.5 and Random Tree algorithm. They are being used depending upon the specificity of the problem; these techniques have their own advantages and drawbacks.

Discussion

This study consists of two stages, data pre-processing and application of supervised learning algorithms viz ID3, Naïve Bayes, C4.5 and Random Tree to data set. In preprocessing stage, we have applied equal interval binning with approximate values based on medical expert advice to the data set. In this paper WEKA data mining tool is used for modeling gestational diabetes data.

This WEKA tool is an open source data mining software mainly used for research and academic purposes.

Dataset

The Dataset used in this work is clinical data set collected from the St. Isabella Hospital, Mylapore, Chennai and from National Institute of Diabetes, Digestive and Kidney Diseases and contains records of about 600 patients. In particular, all patients here are pregnant and above 21 years.

Table.1 The attributes used in the experimentation

Attribute	Description	Type
Preg	Number of times pregnant	Numeric
Plas	Plasma glucose concentration - 2 hours in an oral glucose tolerance test	Numeric
Pres	Diastolic blood pressure (mm Hg)	Numeric
Mass	Body Mass Index (weight in kg/(height in m)^2)	Numeric
Pedi	Diabetes pedigree function (Family history details)	Numeric
Age	Age (years)	Nominal

Data Preprocessing

In real world, data is not always complete and in the case of the medical data, it is always true. To remove the number of inconsistencies which are associated with data, we use Data preprocessing. Many data preprocessing techniques are given in [9] [3]. During this study, we removed instances which had zero values for the attributes – Pregnant, Plasma Glucose, Diastolic BP and BMI. In [8] the authors have proved that list wise deletion is an efficient technique instead of replacing the values with techniques like mean, mode, random imputation, two regression imputations, and a Bayesian model-based procedure [8]. For this study for data preprocessing supervised attribute filtering technique is used. Among various filters available discretize filter is used to derive good intervals of data. After pre-processing only 540 instances remain out of 600.

System Design

This section explains the steps involved in building the classification model. The frame work is shown in fig1.

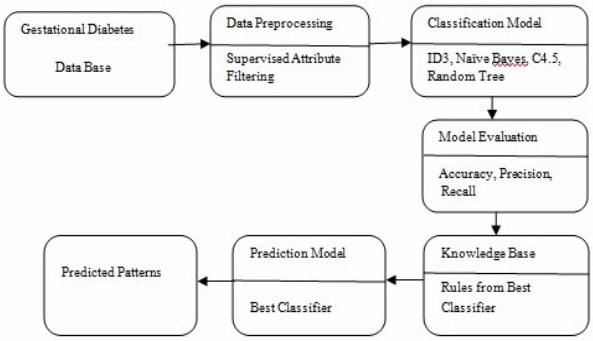

Figure.1 – System Design

Accuracy Measures

In order to perform data modeling, several classification algorithms are applied in medical datasets, which are implemented in WEKA. For this paper we have chosen: ID3, Naïve Bayes (a Naïve Bayes classifier using estimator classes), C4.5 and Random tree algorithms. The tests were executed by means of internal cross validation 10-folds. By using the internal cross-validation it will be easy to find out how the excellence of a learning algorithm will be affected in the separate sets of data. From the average

performance on the test data set, the function of the classifier will be derived. [4].

Accuracy of each method represents how far the set of tuples are being classified. Recall and precision are the accuracy measures used for this study.

TP - Positive tuples classified by the basic classifiers

TN - Negative tuples classified by the basic classifiers

FP - Positive tuples which are being incorrectly classified

FN - Negative tuples which are being incorrectly classified

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Results and Discussion

The Proposed system is executed with all the classifier algorithms discussed in section III and the results are visualized using the data mining tool Tanagra. The error rate and accuracy of the algorithms are evaluated. The error rates of ID3, Naïve Bayes, C4.5, Random Tree are shown in figure 2, figure 3, figure 4 and figure 5 respectively.

Classifier Performances

Error rate	0.2099
Values prediction	**Confusion matrix**

Value	Recall	1-Precision
yes	1.0000	0.2099
no	0.0000	1.0000

	yes	no	Sum
yes	128	0	128
no	34	0	34
Sum	162	0	162

Figure 2 – Error Rates of ID3

Error rate	0.0123
Values prediction	**Confusion matrix**

Value	Recall	1-Precision
yes	1.0000	0.0154
no	0.9412	0.0000

	yes	no	Sum
yes	128	0	128
no	2	32	34
Sum	130	32	162

Figure.3 Error Rate of Naïve Bayes

Error rate	0.0000
Values prediction	**Confusion matrix**

Value	Recall	1-Precision
yes	1.0000	0.0000
no	1.0000	0.0000

	yes	no	Sum
yes	128	0	128
no	0	34	34
Sum	128	34	162

Figure.4 Error Rate of C4.5

Error rate			0.0000			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		yes	no	Sum
yes	1.0000	0.0000	yes	128	0	128
no	1.0000	0.0000	no	0	34	34
			Sum	128	34	162

Figure.5 Error Rate of Random Tree

The comparison of error rates of the classifiers are shown in figure 6

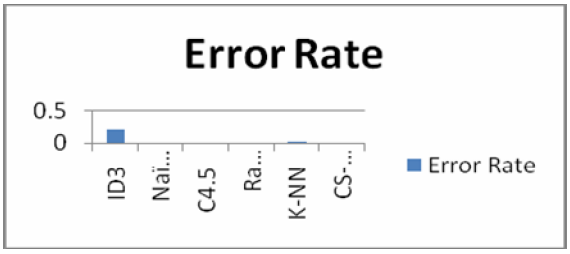

Figure.6 Basic Classifier Error Rate

From Figure 6 It was observed that C4.5, Random Tree and CS-CRT classifiers have least error rate as 0.000 compared to the other classifiers in predicting the risk of gestational diabetes. The classifiers with the error rate and accuracy values are shown in Table 2.

Classifier	Error Rate	Accuracy Value
ID3	0.0209	0.938
Naïve Bayes	0.0123	0.934
C4.5	0.000	0.903
Random Tree	**0.000**	**0.938**

Table.2 Classifiers Error rate and Accuracy values

From the table 2 it is clear that Random Tree algorithm is having less error rate and more accuracy value.

The accuracy values of various classifiers are shown in figure 7

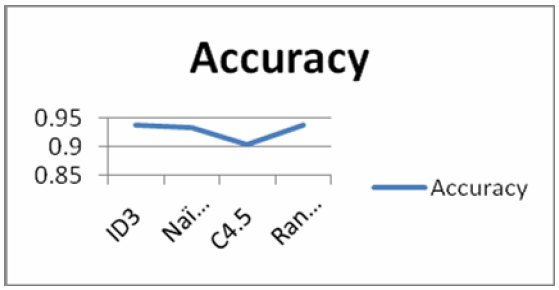

Figure.7 Basic Classifier Accuracy Value

The classifiers with corresponding precision and recall values listed in Table 3.

Classifier	Precision		Recall	
	Yes	No	Yes	No
ID3	0.953	1	1	0.824
Naïve Bayes	0.953	0.824	0.953	0.824
C4.5	0.939	0.839	0.961	0.8
Random Tree	0.955	5	0.977	0.903

Table.3 Accuracy Measures Precision and Recall

The gestational diabetes data for about 540 tuples with 7 attributes are analyzed for their accuracy and error rate with various classification algorithms. Fig 8 shows the accuracy measures Precision and Recall for various algorithms.

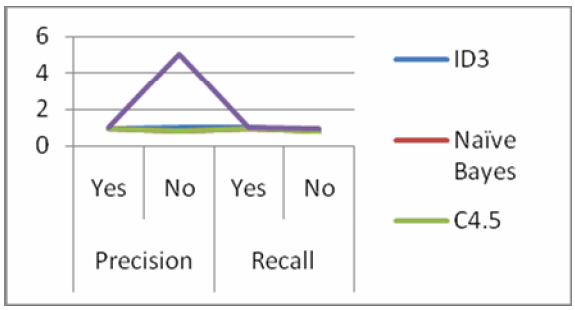

Figure.8 Precision and recall values of the classifiers

From the above analysis it is found that random tree algorithm is the best one when compared to other classifiers for classifying gestational diabetes dataset which contains 540 tuples.

J48 Decision Tree

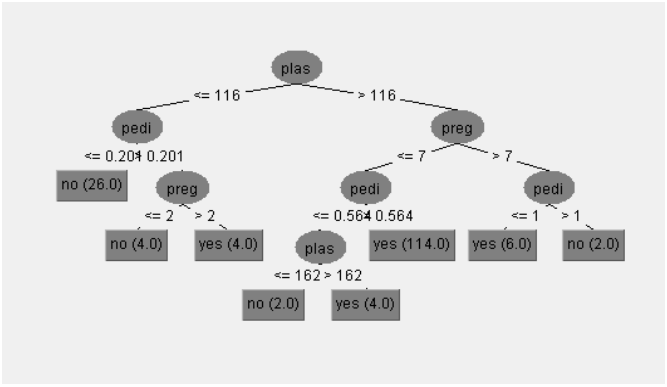

Figure 9 J48 Decision Tree

From figure 9 decision tree, it is revealed that among the seven attributes used for the study, the attributes plasma, pedigree value and number of times being pregnant plays a vital role in classifying gestational diabetes. According to the tree, for any pregnant woman with plasma glucose concentration value less than 116 mg/dl with pedigree value less than 0.2, the chances of getting diabetes are less. If plasma glucose concentration value is less than 116 mg/dl and number of previous pregnancies is greater than 2, then the risk of getting diabetes is very high. In the same way even though plasma glucose concentration value is between 116 mg/dl and 162 mg/dl but number of previous pregnancies is less than 7 with pedigree value less than 0.5, the chances of getting diabetes during gestational period are less. Finally, the figure 9 reveals that pregnant women with plasma glucose concentration value greater than 116 mg/dl, number of previous pregnancy greater than 7 with pedigree value 1 the chances are more for becoming diabetic.

Conclusion and Future work

Preventing gestational diabetes is still one of the most important problems that gynecologists and diabetologists face nowadays. It is very important to prevent or control gestational diabetes because Gestational Diabetes Mellitus (GDM) may go away after pregnancy, but women who have GDM seven times more are likely to develop type2 diabetes than women who do not have GDM in pregnancy. The children of gestational diabetic mother may also have a greater risk of obesity and type2 diabetes [10].

Through this study it is found that the data mining techniques are important and it leads to valid approaches for predicting the risk of gestational diabetes. So it is our recommendation to use new techniques like data mining for decision making in medical fields, which improves the diagnosis of diseases like gestational diabetes. Further this study also throws light on exploration and utilization of new technologies such as data mining to support medical decision, which improves in diagnosing the risk for gestational diabetes.

Acknowledgement

Our sincere and heartfelt thanks to Dr. Jalaja Ramesh., M.B.B.S.,D.C.P.,D.N.B (Gen Med)., C. Diab (Steno Centre Denmark),

PGHSC Diab, Senior Consultant Diabetologist, St. Isabel Hospital, Mylapore, Chennai – 600004 for providing facility for data collection.

References

[1]. Stephen E. Brossette, Alan P. Sprague, Ph.D., J. Michael Hardin Ph.D., Ken.B. Waites, M.D., Warren T. Jones Ph.D, Stephen A. Moster, Ph.D., "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance", Journal for the American Medical Informatics Association, Volume 5, Pg- 4, 1998.

[2]. Jie Gao and Jorg Denzinger, Robert C.James, "CoLe: A Cooperative data Mining Approach and its Application to Early Diabetes Detection", Proceedings of the fifth International Conference on Data Mining (ICDM'05), 2005.

[3]. Patil.B.M. Joshi.R.C, Durga Toshniwal, "Association rule for classification of type-2 diabetic patients", IEEE - Second International Conference on Machine Learning and Computing, DOI 10.1109/ICMLC, Pg-67, 2010

[4]. Huda Yasin, Tahseen A. Jilani, Madiha Danish, " Hepatitis-C Classification using Data Mining Techniques", International Journal of Computer Applications, Volume 24- Pg-3, 2011.

[5]. J.Padmavathi, "A Comparitive Study on Breast Cancer Prediction using RBF and MLP", International Journal of Scientific & Engineering Research, Volume 2, Issue 1, January 2011, ISSN – 2229-5518.

[6]. Duarte Ferreira, Abilio Oliveira, Alberto Freitas, "Applying data mining techniques to improve diagnosus in neonatal jaundice", BMC Medical Informatics and Decision Making, 1472-6947/12/143, 2012.

[7]. Daveedu raju Adidela, Lavanya Devi.G, Jaya Suma.G, Appa Rao Allam, "Application of Fuzzy ID3 to Predict Diabetes", International Journal of Advanced Computer and Mathematical Sciences, Vol 3, Issue 4, Pg-541-545, 2012, ISSN 2230-9624.

[8]. Sonu Kumari, Archana Singh, " A Data mining Approach for the diagnosis of diabetes mellitus", IEEE Journal Pg - 12, 2012, ISBN Number 978-1-4673-4603-0.

[9]. Syed Shajahaan.S, Shanthi.S, ManoChitra.M, "Application of Data Mining techniques to Model Breast Cancer Data", International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9901:2008 Certified Journal, Volume 3, Issue 11, 2013.

[10]. Available in the website http://www.cdc.gov/- in the chapter "National Center for Chronic Disease Prevention and Health Promotion" under the heading "Gestational Diabetes". This website is developed and maintained by Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, 2011.

[11]. Available in the web site - www.thehealthsite.com under the heading Diseases and conditions with subheading women's health.

[12]. Available in the web site http://diabetes.webmd.com/guide/gestational_diabetes under the heading Pregnancy and Gestational Diabetes.

[13]. Data Mining Concepts and Techniques – Jiawei Han and Micheline Kamber, Second edition, ELSEVIER Publisher, pg- 285-289.

[14]. Data Mining – A Tutorial Based Primer by Richard Roiger and Michael Geatz, Fourth Edition, Pearson Publisher, Pg- 100- 102