ISSN 2278-3083

International Journal of Science and Applied Information Technology (IJSAIT), Vol. 3 , No.4, Pages :82 - 86 (2014)
Special Issue of ICECT 2014 - Held during September 01, 2014 in The Golkonda Hotel, Hyderabad, India

An Empirical Study about the role of Classification Algorithms for Diabetic Data Analysis

Dr.V.V. JayaramaKrishniah[1]Dr.K.RamChand[2] Dr.D.V.Chandra Shekar[3]

1 Department of CSE, ASN Womens Engineering CollegeTenali, AP,jkvemul@gmail.coma
2 Department of CSE, ASN Womens Engineering College, Tenali AP, ramkolasani@yahoo.com
3 Department of CSc,TJPS College, Guntur, AP,chand.info@gmail.com

Abstract: Knowledge management and data mining approaches have been widely used in many important applications in scientific, business and Bio Medial domains in recent years.This paper focuses on Bio Medical Domain Analysis, which addresses the bio medical and medical related data, which uses the inherited methodologies used in data mining, artificial intelligence. The aim of this work developed in this paper was to address some of the difficulties of data analysis, in particular, and to come up with a methodology that would makeup ambiguity, anomalies and limited knowledge representation and inconsistency of information generally available with the processed data, more over Biological systems are inherently stochastic and uncertain.

INTRODUCTION

The amount of digital data has been exploding during the past decade, while the number of scientists, engineers and analysts available to analyse the data has been static. To bridge this gap requires the solution of fundamentally new research problems like how was the massive data with high dimensional dataset can be organized, a data set with different kinds of predictions and situations. The straight forward solution for these kinds of questions is Data Mining. Data Mining can be defined as the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Alternatively, it can be defined as the process of data selection and exploration and building models using vast data stores to uncover previously unknown patterns.

In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Several factors have motivated the use of data mining applications in healthcare. In the age of managed care where the cost- effectiveness and necessity of medical procedures are closely scrutinized for patient data analysis, it is not surprising that data mining has begun to play an increasingly important role as data repositories swell with valuable information. Health care data is massive. It includes patient centric data, resource management data and transformed data. Health care organizations must have ability to analyse data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care.

The main aim of Data Mining is Knowledge Discovery in Databases (KDD). KDD is a process that aims at finding valid, useful, novel and understandable patterns in data .Methods of data analysis and automatic processing are treated as knowledge discovery. In many cases it is necessary to classify data in some way or find regularities in the data. That is why the notion of similarity is becoming more and more important in the context of intelligent data processing systems. It is frequently required to ascertain how the data are interrelated, how various data differ or agree with each other, and what the measure of their comparison is.

RELATED STUDY

Diabetes is an illness which occurs as a result of problems with the production and supply of insulin in the body. People with diabetes have high level of glucose or “high blood sugar”. This leads to serious long-term complications such as eye disease, kidney disease, nerve disease, disease of the circulatory system, and amputation this is not the result of an accident. Diabetes also imposes a large economic impact 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 Type-2 diabetes mellitus (T2DM) is the most common type of diabetes and accounts for 90-95% of all diabetes patients and most common in people older than 45 who are overweight. However, as a consequence of increased obesity among the young, it is becoming more common in children and young adults. In T2DM, the pancreas may produce adequate amounts of insulin to metabolize glucose (sugar), but the body is unable to utilize it efficiently. Over time, insulin production decreases and blood glucose levels rise. T2DM patients do not require insulin treatment to remain alive, although up to 20% are treated with insulin to control blood glucose levels. Diabetes has no obvious clinical symptoms and not been easy to know, so that many diabetes patient unable to obtain the right diagnosis and the treatment. Therefore, it is important to take the early detection, prevent and treat diabetes disease, especially for T2DM. Recent studies by the National Institute of Diabetes and Digestive and Kidney Diseases (DCCT) in India shown that effective control of blood sugar level is beneficial in preventing and delaying the progression of complications of diabetes. Adequate treatment of diabetes is also important, as well as lifestyle factor such as smoking and maintaining healthy bodyweight. According to this context, data mining and machine learning could be used as an alternative way in discovering knowledge from the patient medical records and classification task has shown remarkable success in the area of employing computer aided diagnostic systems (CAD) as a “second opinion” to improve diagnostic decisions.

Data Preparation

The training dataset used for data mining classification was the Pregnancy Diabetes Database extracted from UCI Machine Learning. The dataset contains 768 record samples, each having 8 attributes. We used this dataset for our classification exercise, as the data is complete.

The following Table I illustrated the description about the Dataset used.

S.No	Attribute	Type
1	Number of times pregnant	Continuous
2	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Continuous
3	Diastolic blood pressure (mm Hg)	Continuous
4	Triceps skin fold thickness (mm)	Continuous
5	2-Hour serum insulin (mu U/ml)	Continuous
6	Body mass index (kg/m)^2)	Continuous
7	BMI type	Discrete
8	Diabetes pedigree function	Continuous
9	Age (years)	Continuous
10	Class Variable(0,1)	Discrete

Methodology :Classification

Classification is a classic data mining technique based on machine learning. Classification trees are methodologies to classify data into discrete ones using the tree structured algorithms. Basically classification is used toclassify each item in a set of data into one of predefined set of classes or groups**.** For example, one tree structured classifier uses blood pressure, age to classify whether the patient as risk or not.A Rule-based classificationextracts a set of rules that show relationships between attributes of the data set and the class label. The main purpose of these classifications is to expose the structural information contained in the data. Analyst can easily understood the propagation and relationships between the data items as response variable or target variable and explanatory or predictor variable.

In predicative classification, we use a representation called confusion matrix, which is a table with two rows and two columns, represents the number of false positives, false negatives, true positives and true negatives. This allows more detailed analysis than mere propagation of correct guess or accuracy.

EXPERIMENTAL ANALYSIS

This section provides the performance comparison of the different classification algorithms like “C4.5”, “ID3”, “C-RT”,”MLP”,”K-NN” and “Naïve Bayes”. For data analysis, we used a data mining tool called Tanagra. The training data set consists of 768 instances with 9 different attributes. The instances in the dataset are representing the results of different typesof testing to predict the accuracy of heart disease. The performance of the classifiers isevaluated and their results are analysed.

For data analysis, we defined a new attribute called “BMI Type” based on Training datasets attribute “BMI”. Basically the Body Mass Index is an continuous attribute, we categorise the BMI into six Discrete Values like “Under Weight”, “Normal”, “Overweight”, “Obese Class I”, “Obese Class II” and “Obese Class III”. The following algorithm defines, how the six BMI Type was were created.

```
if (bmi< 18.5)
bmitype = "Underweight"
if ((bmi>=18.50) && (bmi< 25))
bmitype = "Normal"
if ((bmi>=25) && (bmi< 30))
bmitype = "Overweight"
if (bmi>= 30 &&bmi<35)
bmitype = "Obese Class I"
if (bmi>= 35 &&bmi<40)
```

bmitype = "Obese Class II"
if (bmi>= 40)
bmi type =”Obese Class III

For making predictions we classify the four different responses as “Diabetic and Good Control”, “High Stage of Diabetic”, Very High Diabetic” and “Normal”.

We evaluated the performance of the classifier with different initial attributes like “Plasma Glucose Concentration”, “Diabetes Pedigree Function”, ”BMI” and “Diastolic Blood Pressure”.

The following figures shows the confusion matrix for all classification methods

Fig:1. Confusion Matrix for Training Dataset with “Diastolic Blood Pressure” for C4.5

	High Stage of Diabetic	**No Diabetic**	**Very High Diabetic**	**Diabetic and Good Control**	**Sum**
High Stage of Diabetic	25	147	3	0	175
No Diabetic	15	422	2	0	439
Very High Diabetic	11	105	6	0	122
Diabetic and Good Control	2	29	1	0	32
Sum	53	703	12	0	768

Fig:2. Confusion Matrix for Training Dataset with “Diastolic Blood Pressure” for MLP

	High Stage of Diabetic	**No Diabetic**	**Very High Diabetic**	**Diabetic and Good Control**	**Sum**
High Stage of Diabetic	47	128	0	0	175
No Diabetic	52	387	0	0	439
Very High Diabetic	35	87	0	0	122
Diabetic and Good Control	1	31	0	0	32
Sum	135	633	0	0	768

Fig:3. Confusion Matrix for Training Dataset with “Diastolic Blood Pressure” for K-NN

	High Stage of Diabetic	**No Diabetic**	**Very High Diabetic**	**Diabetic and Good Control**	**Sum**
High Stage of Diabetic	42	121	7	5	175
No Diabetic	52	330	34	23	439
Very High Diabetic	21	83	15	3	122
Diabetic and Good Control	4	21	1	6	32
Sum	119	555	57	37	768

Fig:4. Confusion Matrix for Training Dataset with “Diastolic Blood Pressure” for ID3

	High Stage of Diabetic	**No Diabetic**	**Very High Diabetic**	**Diabetic and Good Control**	**Sum**
High Stage of Diabetic	62	113	0	0	175
No Diabetic	62	377	0	0	439
Very High Diabetic	38	84	0	0	122
Diabetic and Good Control	3	29	0	0	32
Sum	165	603	0	0	768

Fig:5. Confusion Matrix for Training Dataset with "Diastolic Blood Pressure" for C-RT

	High Stage of Diabetic	No Diabetic	Very High Diabetic	Diabetic and Good Control	Sum
High Stage of Diabetic	0	175	0	0	175
No Diabetic	0	439	0	0	439
Very High Diabetic	0	122	0	0	122
Diabetic and Good Control	0	32	0	0	32
Sum	0	768	0	0	768

Fig:6. Confusion Matrix for Training Dataset with "Diastolic Blood Pressure" for Naïve Bayes

	High Stage of Diabetic	No Diabetic	Very High Diabetic	Diabetic and Good Control	Sum
High Stage of Diabetic	0	175	0	0	175
No Diabetic	0	439	0	0	439
Very High Diabetic	0	122	0	0	122
Diabetic and Good Control	0	32	0	0	32
Sum	0	768	0	0	768

Table 2: Comparison of Error rates for Classification Algorithms

S.No	Algorithm Name	Error Rate				Average Error rate
		Plasma glucose concentration	Diabetes pedigree function	BMI	Diastolic blood pressure	
1	C 4.5	0.0000	0.4076	0.3919	0.4102	0.3024
2	ID3	0.0334	0.4268	0.4282	0.4284	0.3292
3	C-RT	0.0108	0.4282	0.4284	0.4284	0.3239
4	MLP	0.0410	0.4232	0.4297	0.4349	0.3322
5	K-NN	0.0000	0.3607	0.3906	0.4883	0.3099
6	NAÏVE BAYES	0.0286	0.4271	0.4284	0.4284	0.3281

From the above table, we notice the error rate was change with respect to the initial attribute consider for the classification.

In a classification task, the precision for a class is the *number of true positive* i.e. the *number of items correctly labeled as belonging to the positive classdivided by the total number of elements labeled as belonging to the positive class*i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class. Recall in this context is defined as the *number of true positivesdivided by the total number of elements that actually belong to the positive class* i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been.

Table3 : Precession and Recall Values for the Classification Algorithms (A. High Stage of Diabetic, B. No Diabetic, C. Very High Diabetic, D. Diabetic and Control)

Algorithm Name	Predictor	1-Precession				Recall			
		Plasma glucose concentration	Diabetes pedigree function	BMI	Diastolic blood pressure	Plasma glucose concentration	Diabetes pedigree function	BMI	Diastolic blood pressure
"C4.5"	**A**	0.0000	0.3333	0.4595	0.5283	1.0000	0.0571	0.2286	0.1429
	B	0.0000	0.4106	0.3765	0.3997	1.0000	0.9909	0.9317	0.9613
	C	0.0000	0.3333	0.5263	0.5000	1.0000	0.0820	0.1475	0.0492
	D	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
"ID3"	**A**	0.0000	1.0000	1.0000	0.6242	1.0000	0.0000	0.0000	0.3543
	B	0.0000	0.4284	0.4284	0.3748	1.0000	1.0000	1.0000	0.8588
	C	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
	D	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
"C-RT"	**A**	0.0000	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
	B	0.0000	0.4284	0.4284	0.4284	1.0000	1.0000	1.0000	1.0000
	C	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	D	0.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
"MLP	**A**	0.0000	1.0000	0.7273	0.6519	1.0000	0.0000	0.0171	0.2686
	B	0.0679	0.4246	0.4254	0.3886	1.0000	1.0000	0.9909	0.8815
	C	0.0000	0.0000	1.0000	1.0000	1.0000	0.0328	0.0000	0.0000
	D	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000
K-NN	**A**	0.000	0.4960	0.5182	0.6471	1.0000	0.3600	0.3771	0.2400
	B	0.000	0.3162	0.3321	0.4054	1.0000	0.8770	0.8337	0.7517
	C	0.000	0.4507	0.5570	0.7368	1.0000	0.3197	0.2869	0.1230
	D	0.000	0.5556	0.7500	0.8378	1.0000	0.1250	0.0313	0.1875
"Naïve Bayes".	**A**	0.0541	1.0000	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000
	B	0.0266	0.4256	0.6263	0.4284	1.0000	0.9932	0.9932	1.0000
	C	0.0000	0.5556	0.6250	1.0000	0.9180	0.0328	0.0246	0.0000
	D	0.0000	1.0000	1.0000	1.0000	0.6250	0.0000	0.0000	0.0000

From the above table 3, we notice, Precession and Recall values were varies with respect to algorithm and initial attribute for the classification, among C4.5 and C-RT shows the better results than other classification models.

CONCLUSION

In this paper, we discussed the various classification techniques decision making. From the results obtained through the different classification methods, We noticed, some of algorithms struggle with ambiguity even though gives the better results in the case of C-RT and MLP ,because the initial or starting attribute plays crucial role in the construction of the decision tree and Prune percentage is also high in the case of those algorithms.

REFERENCES

[1] http://diabetes.co.in.

[2] Han, J., Kamber, M.: Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers (2000).

[3] Margaret H. Dunham,-"Data Mining Techniques and Algorithms", Prentice Hall Publishers.

[4] SantiWulanPurnami, S.P. Rahayu and AbdullahEmbong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.

[5] PardhaRepalli, "Prediction on Diabetes Using Data mining Approach".

[6] Joseph L. Breault., "Data Mining Diabetic Databases:Are Rough Sets a Useful Addition? " .

[7] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method ", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[8] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

[9] UCI Machine Learning Repository- Center for Machine Learning and Intelligent System, http://archive.ics.uci.edu.