Middle-East Journal of Scientific Research
24 (Recent Innovations in Engineering, Technology, Management & Applications): 207-211, 2016
ISSN 1990-9233; © IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.RIETMA134

An Effective Prediction of Diabetics Using ID3 Classification Algorithm

[1]S. Sathya and [2]A. Rajesh

[1]Department of Computer Science& Engineering, C. Abdul Hakeem College of Engineering & Technology, Tamilnadu, Research Scholar, St.Peter's University, Chennai, India
[2]Professor & Head in the Computer Science & Engineering,
C. Abdul Hakeem College of Engineering & Technology, Tamilnadu, India

Abstract: This research focus on predicting diabetic using ID3 classification algorithm. ID3 algorithm is well suited for robust missing value. ID3 gives better performance when compared with many other classification algorithms. The UCI Machine learning diabetic dataset has been used for implementation in ID3 algorithm. The dataset contains around 50 attributes and 10Z, 1767 instances. Among this, around 100 instances with 50 attributes have been used for implementation. This implementation has been done with three different modes of runs with 10 fold cross validation of training and test data. Run 1: The actual diabetic dataset collected from UCI Machine learning database has been given as input to ID3 algorithm and measures were recorded. Later, the same dataset has been preprocessed to remove missing values, inconsistent, noisy and erroneous data. The data cleaning and implementation has been done in two ways. Run 2: The dataset has been cleaned by unsupervised learning method and then fed as input to the ID3 algorithm and the measures were noted. Run 3: The dataset has been cleaned by supervised learning method and given as input to the ID3 algorithm and the measures were noted. The comparative analysis has been made between the different modes of run and the accuracies and other measures were tabulated and charted for analysis. It is found that the data cleaning with supervised learning method gives better accuracy of 63% while unsupervised learning yields 56% and actual data set without preprocessing gives the poor result of 48% accuracy.

Key words: ID3 classification · Diabetic · Data preprocessing · Weka · Machine learning · Prediction

INTRODUCTION

Diabetes has affected over 422 Million adults which causes around 1.5 million worldwide with a majority of them being women. According to the WHO report, by 2025 this number is expected to rise to over 380 million [1]. Diabetic is one of the most dangerous disease leading huge volumes of death rate every year particularly in developed countries. It is necessary to introduce systems to help doctors to predict diabetes earlier. Developing a tool to be embedded in the hospitals management system to help the healthcare professionals in diagnosing patients is important [2]. Diabetes is a lifelong chronic condition that affects the human body by reducing the insulin which carries glucose into the blood cells. This increases the sugar level in the body leading to different complications like stroke, heart disease, blindness, kidney failure and death.

Diabetic patients generally have the following symptoms and they are increased thirst, frequent urination, Weight loss, increased hunger, Slow-healing infections, Blurred vision, Nausea and Vomiting [3]. HUMAN body needs energy for activation. The carbohydrates are broken down to glucose, which is the important energy source for human body cells. Insulin is needed to transport the glucose into body cells. There are two types of diabetes such as type 1 and type 2. The insulin deficiency is the outcome of diabetes [4]. Diabetes is a disease that can produce terrible complications of blindness, kidney failure, amputation and premature cardiovascular death [5]. The leading complication of diabetes is shown in Fig 1. Preventing the disease of diabetes is an ongoing area of interest to the healthcare community. Increased awareness and treatment of diabetes should begin with prevention [6]. Diabetes mellitus is a chronic disease and is caused when the body

Corresponding Author: S. Sathya, Department of Computer Science& Engineering, C. Abdul Hakeem College of Engineering & Technology, Tamilnadu, Research Scholar, St.Peter's University, Chennai, India.

loses its ability to turn food into energy [7]. Diabetes is one of the high prevalence diseases worldwide with increased number of complications, with retinopathy as one of the most common one [8]. One of the useful applications in the field of medicine is the incurable chronic disease diabetes [9]. In this research, the ID3 algorithm is implemented with UCI Machine learning diabetic dataset using WEKA, a popular data mining tool.

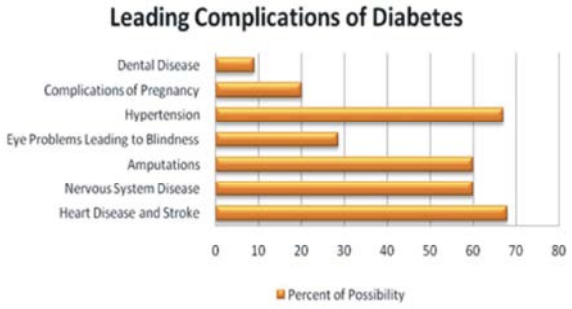

Fig. 1: Leading complications of Diabetes

Data Mining: The usage of data mining techniques in disease prediction is to reduce the test and increase the accuracy of rate of detection [10]. Data Mining is the process of extracting hidden knowledge from large volumes of raw data [11]. It is the process of selecting, exploring and modeling large amount of data and has become an increasingly pervasive activity in all areas of research in medical science. It is the core step, which results in the discovery of useful hidden patterns from massive databases [12].

ID3 Algorithm: J. Ross Quinlan originally developed ID3 at the University of Sydney. He first presented ID3 in 1975. ID3 is based on the Machine learning strategy. The function of ID3 classification algorithm is used in this paper. The ID3 algorithm constructs a decision tree from a well-known training data set of examples. The resulting tree is used to classify test datasets. The example has several attributes and belongs to a class. The leaf nodes of the decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch to another decision tree being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node. . If new records from test data were given as input, it can predict the class label. The generated tree contains Root Node, Branch Node and Leaf or class Node. The root node represents test on an Attribute, branch node represents Outcome of the test and leaf node represents the decision made after testing all attributes called class labels. The path generated from root node to leaf node represents the classification rules. Decision tree can handle both categorical and numerical data [13].

Weka: Waikato Environment for Knowledge Analysis, called shortly WEKA, is a set of state-of-the art data mining algorithms and tools to in-depth analyses. Weka is a state-of-the-art facility for developing machine learning techniques and their application to real-world data mining problems [14]. The author of this environment is University of Waikato in New Zealand. The programming language of WEKA is Java and its distribution is based on GNU General Public License. These complex algorithms may be applied to data set in the aim of detailed analyses and evaluation of data mining examination. There are three main ways of WEKA use. First is analyzing data mining methods ?outputs to learn more about the data; next is generation of model for prediction of new instances and finally the last but most important for this master?s thesis feature, comparison of data mining methods in order to choose the best one as a predictor e.g. in Medical Decision Support System. The new machine learning schemes can also be developed with this package [15].

Implementation

Run 1: ID3 Algorithm without Data Cleaning:

Table 1: ID3 Algorithm Without Data Cleaning (Run 1)

Measures	Values	Accuracy
Correctly Classified	48	48
Incorrectly Classified Instances	52	52
Kappa statistic	-0.0252	
Mean absolute error	0.3839	
Root mean squared error	0.4482	
Relative absolute error	100.8857%	
Root relative squared error	103.0091%	
Coverage of cases (0.95 level)	99%	
Mean rel. region size (0.95 level)	99.3333%	
Total Number of Instances	100	

The actual dataset without cleaning has been fed as input to the ID3 algorithm. It is found that 48 instances have been correctly classified and 52 instances have been incorrectly classified. Run 1 gives the kappa statistic of -0.0252, Mean absolute error of 0.3839, Root mean squared error of 0.4482 , Relative absolute error of 100.8857%, Root relative squared error of 103.0091%, coverage of cases at 0.95 level of 99% and Mean relational region size at 0.95 level of 99.333%. The measures were tabulated in the Table 1: ID3 algorithm without data cleaning for analysis.

Table 2: Weighted Average of ID3 algorithm without data cleaning

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.071	0.103	0.333	0.071	0.118	0.463	NO
	0.9	0.92	0.495	0.9	0.638	0.482	>30
	0	0	0	0	0	0.395	<30
Weighted Average	0.48	0.503	0.387	0.48	0.369	0.467	

Table 3: Confusion Matrix of ID3 algorithm without data cleaning

A	B	C	Classified as
3	39	0	NO
5	45	0	>30
1	7	0	<30

Table 4: Measures of ID3 algorithm with unsupervised data cleaning. (Run 2)

Measures	Values	Accuracy
Correctly Classified	56	56
Incorrectly Classified Instances	44	44
Kappa statistic	0.0723	
Mean absolute error	0.3553	
Root mean squared error	0.4372	
Relative absolute error	92.2987%	
Root relative squared error	99.841%	
Coverage of cases (0.95 level)	96%	
Mean rel. region size (0.95 level)	90%	
Total Number of Instances	100	

It is found that the weighted average measures of ID3 algorithm without data cleaning as TP Rate of 0.46, FP Rate of 0.503, Precision of 0.387, Recall of 0.48, F-Measure of 0.369 and ROC of 0.467. The measures were tabulated in Table 2 for analysis.

The values obtained in confusion matrix indicate the True positive rate, True negative rate, false positive rate and false negative rate of the whole instances and the respective values were tabulated in Table 3: Confusion Matrix for analysis.

Run 2: ID3 Algorithm With Unsupervised Data Cleaning: The actual dataset without cleaning has been fed as input to the ID3 algorithm. It is found that 56 instances have been correctly classified and 44 instances has been incorrectly classified. Run 2 gives the kappa statistic of 0.0723, Mean absolute error of 0.3553, Root mean squared error of 0.4372, Relative absolute error of 92.2987%, Root relative squared error of 99.841%, coverage of cases at 0.95 level of 96% and Mean relational region size at 0.95 level of 90%. The measures were tabulated in the Table 4 analysis.

It is found that the weighted average measures of ID3 algorithm with unsupervised data cleaning as TP Rate of 0.56, FP Rate of 0.496, Precision of 0.651, Recall of 0.56, F-Measure of 0.439 and ROC of 0.552. The measures were tabulated in Table 5 for analysis.

Table 5: Weighted Average of ID3 algorithm with unsupervised data cleaning. (Run 2)

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.086	0	1	0.086	0.158	0.559	NO
	0.981	0.913	0.558	0.981	0.711	0.548	>30
	0	0.022	0	0	0	0.547	<30
Weighted Average	0.56	0.496	0.651	0.56	0.439	0.552	

Table 6: Confusion Matrix of ID3 algorithm with unsupervised data cleaning. (Run 2)

A	B	C	Classified as
3	31	1	NO
0	53	1	>30
0	11	0	<30

Table 7: Measures of ID3 algorithm with supervised data cleaning. (Run 3)

Measures	Values	Accuracy
Correctly Classified	63	63
Incorrectly Classified Instances	37	37
Kappa statistic	0.2732	
Mean absolute error	0.2655	
Root mean squared error	0.4041	
Relative absolute error	73.6537%	
Root relative squared error	95.4829%	
Coverage of cases (0.95 level)	90%	
Mean rel. region size (0.95 level)	55%	
Total Number of Instances	100	

The values obtained in confusion matrix indicates the True positive rate, True negative rate, false positive rate and false negative rate of the whole instances and the respective values were tabulated in Table 6 confusion Matrix for analysis.

ID3 With Supervised Data Cleaning: The actual dataset without cleaning has been fed as input to the ID3 algorithm. It is found that 63 instances have been correctly classified and 37 instances has been incorrectly classified. Run 3 gives the kappa statistic of 0.2732, Mean absolute error of 0.2655, Root mean squared error of 0.4041, Relative absolute error of 73.6537%, Root relative squared error of 95.4829%, coverage of cases at 0.95 level of 90% and Mean relational region size at 0.95 level of 55%. The measures were tabulated in the Table 7 analysis.

It is found that the weighted average measures of ID3 algorithm with supervised data cleaning as TP Rate of 0.63, FP Rate of 0.36, Precision of 0.632, Recall of 0.594, F-Measure of 0.594 and ROC of 0. 744. The measures were tabulated in Table8 for analysis.

Table 8: Weighted Average of ID3 algorithm with supervised data cleaning. (Run 3).

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.378	0.127	0.708	0.378	0.493	0.726	NO
	0.902	0.592	0.613	0.902	0.73	0.769	>30
	0	0.01	0	0	0	0.638	<30
Weighted Average	0.63	0.36	0.632	0.594	0.594	0.744	

Table 9: Confusion Matrix of ID3 algorithm with supervised data cleaning. (Run 3)

A	B	C	Classified as
17	27	1	NO
5	46	0	>30
2	2	0	<30

The value obtained in confusion matrix indicates the True positive rate, True negative rate, false positive rate and false negative rate of the whole instances and the respective values were tabulated in Table 9: confusion Matrix for analysis.

RESULTS AND DISCUSSION

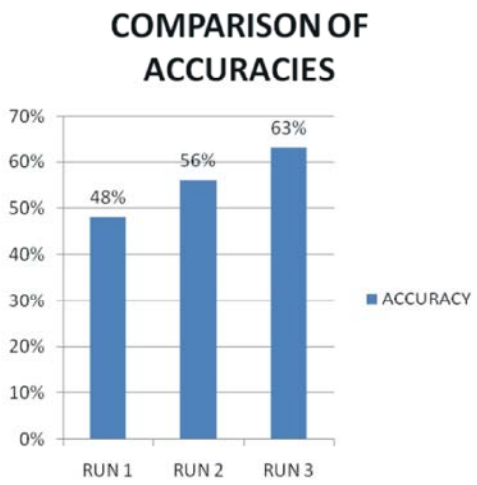

Fig. 2: Comparison of accuracies

CONCLUSION

The diabetic dataset from UCI Machine learning has been implemented with ID3 algorithm in three modes as mentioned above as Run 1, Run 2 and Run 3. In Run-1, the actual UCI Machine learning dataset and ID3 algorithm is implemented and the instances classified correctly are 48 and incorrectly classified instances are 52. In Run-2, under unsupervised learning of data cleaning and ID3 algorithm, 56 instances were classified correctly and 44 instances were classified incorrectly. In Run-3, under supervised learning of data cleaning and ID3 classification algorithm, 63 instances were classified correctly and 37 instances were classified incorrectly. It is agreed that Run-3 gives an improved accuracy rate of 63% and better than the other two runs. It is found that the ID3 suffers due to overfitting, that is the tree behaves well for datasets similar to training set and not well for real time datasets. This issue can be addressed by further researches in ID3 algorithm.

REFERENCES

1. Iyer, A., S. Jeyalatha and R. Sumbaly, 2015. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.
2. Evirgen, H. and M. Çerkezi, 2014. Prediction and Diagnosis of Diabetic Retinopathy using Data Mining Technique. The Online Journal of Science and Technology, 4(3).
3. Vijayan, V. and A. Ravikumar, 2014. Study of data mining algorithms for prediction and diagnosis of diabetes mellitus. International Journal of Computer Applications, 95(17).
4. Durairaj, M. and G. Kalaiselvi, 2015. Prediction Of Diabetes Using Soft Computing Techniques-A Survey. International Journal Of Scientific & Technology Research, 4(03): 2277-8616.
5. http://www.psu.edu.sa/megdam/sdma/Downloads/Posters.
6. Sa-ngasoongsong, A. and J. Chongwatpol, 2012. An analysis of diabetes risk factors using data mining approach. Oklahoma state university, USA.
7. Jothikumar, R., R.V. Sivabalan and A.S. Kumarasen, 2015. Data Cleaning Using Weka For Effective Data Mining In Health Care Industries, International Journal of Applied Engineering Research, 10(30).
8. Sathya, S. and A. Rajesh, 2015. Performance Analysis On Diabetes Prediction With Different Classification Algorithms Using Weka, International Research Journal In Advanced Engineering And Technology, 1(4): 178-190.
9. Sanakal, R. and S.T. Jayakumari, 2014. Prognosis of Diabetes Using Data mining Approach-Fuzzy C Means Clustering and Support Vector Machine. International Journal of Computer Trends and Technology, 11(2): 94-8.

10. Thirumal, P.C. and N. Nagarajan, 2015. Utilization of Data mining Techniques for Diagnosis of Diabetes Mellitus- A Case Study. ARPN Journal of Engineering and Applied Sciences, 10(1): 8-13.
11. Radha, P. and B. Srinivasan, 2014. Predicting Diabetes by cosequencing the various Data Mining Classification Techniques. IJISET-International Journal of Innovative Science, Engineering & Technology, 1(6).
12. Aljumah, A.A., M.G. Ahamad and M.K. Siddiqui, 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2): 127-136.
13. Sathya, S., A. Rajesh and R. Manivannan, 2014. Prediction of diabetes using Decision Trees, International Journal of Applied Engineering Research, 9(24): 27165-27178, ISSN 0973-4562.
14. Jothikumar, R. and R.V. Sivabalan, 2015. Performance Analysis on Accuracies of Heart Disease Prediction System Using Weka by Classification Techniques. AJBAS, 9(7): 741-749.
15. Jothikumar, R., R.V. Sivabalan and E. Sivarajan, 2015. Accuracies of j48 weka classifier with different supervised weka filters for predicting heart diseases, ARPN Journal of Engineering and Applied Sciences, 10(17): 7788-7793, ISSN 1819-6608.